Project 2: Medical Imaging

Alexander Thoren, Josefine Nyeng, Miranda Speyer-Larsen, Pedro Prazeres

June 02, 2023

GitHub Repository: `https://github.com/FIYEP-2023/Skin-cancer`

Abstract

This study investigates the reliability of medically-defined, computer-based measurements in predicting the malignancy of skin lesions, specifically asymmetry, colour, and border irregularity. Additionally, computational features extracted from skin lesions to enhance the accuracy of computer-based skin cancer detection were explored. The research is based on a dataset of 1200 skin lesions, each accompanied by a corresponding diagnosis. The data is prepared through manual image segmentation, resulting in isolated skin lesions and their associated masks. These are used to computationally extract features. Classifiers are trained using cross-validation, and optimised through trial and error. Based on an evaluation of various metrics from the cross-trained models, Logistic Regression is selected as the final model. When applied to the testing data, it accurately identifies 62.7% of cancerous lesions. Furthermore, 12 general features are selected through feature selection and used to train a model that results in a KNN classifier that accurately identifies 67.2% lesions, an improvement of 4.5% over the previous model.

Introduction

Early detection of skin cancer is crucial as it vastly increases the chances of survival. In fact, when detected early, most skin cancers have a survival rate of at least 95% (Moffitt Cancer Center, n.d). Dermatologists diagnose skin cancer using some variation of the ABCD rating system. This system rates the skin lesion based on 4 different features: asymmetry, border irregularity, colour, and diameter (Betul Tas, 2020). We want to investigate how well this system can be translated to a computer-based classifier in order to predict cancerous lesions. Thus we want to answer the following research question: “How reliably can computer-based measurements of asymmetry, colour, and border irregularity be used to predict cancerous skin lesions?” For our open question, we would like to further investigate if the computer-based diagnosis of skin lesions can be improved by extracting general features from the skin lesions, which are otherwise not considered by dermatologists as they cannot be assessed manually. Thus our secondary research question will be "How can computational features extracted from skin lesions improve the computer detection of skin cancer?"

Literature review

Computer based diagnosis

Due to the high survival rate that comes with early diagnosis, technology is constantly being developed in the hopes of aiding dermatologists in the diagnosis of skin cancer. Several apps have been developed to help people assess their skin lesions from home. For example, the app SkinVision requires a photograph of the skin lesion, and assesses the lesion as either high, low, or medium risk (The Medical Futurist, 2019). However, relying solely on such apps can be dangerous in case it outputs the wrong diagnosis. A 2022 study discovered that the apps evaluated had an accuracy of 59% on average (Memorial Sloan Kettering Cancer Center, 2022).

Data Exploration

The given dataset consisted of 1200 images of skin lesions, accompanied by a metadata csv file that included the diagnosis for each image as well as other variables such as age, gender, and history of skin cancer. However, since this project revolves around image classification, we will focus on the images and not on the additional variables provided. The cancerous diagnoses were Basal Cell Carcinoma (BCC), Squamous Cell Carcinoma (SCC), and Melanoma (MEL) and the skin diseases were Actinic Keratosis (ACK), Nevus (NEV) and Seborrheic Keratosis (SEK). The distribution of these diagnoses was quite varied as can be seen in Figure 1 below.

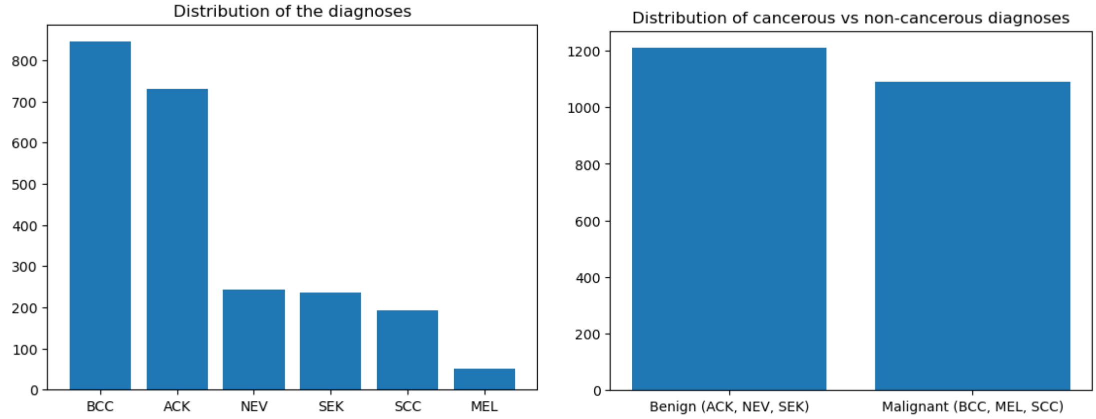

Figure 1: Figure showing the distribution of diagnoses and cancerous vs non-cancerous skin lesions

As we can see Melanoma is vastly underrepresented compared to, for example, Basal Cell Carcinoma. However, if we split the diagnoses into cancerous and non-cancerous we can also see in Figure 1 that the dataset has a somewhat even distribution of the two classes.

Data preparation

The images were segmented in order to create binary masks of the skin lesion. This was done manually using a photo editing software such as Photoshop to highlight the lesion area and encode it as 1, and encode the surrounding area as 0. Around 150 images were segmented by each group member, resulting in a total of 623 segmented images. A segmented lesion can be seen in Figure 2.

Additionally, to aid in the computer feature creation, the images were all resized to 1024

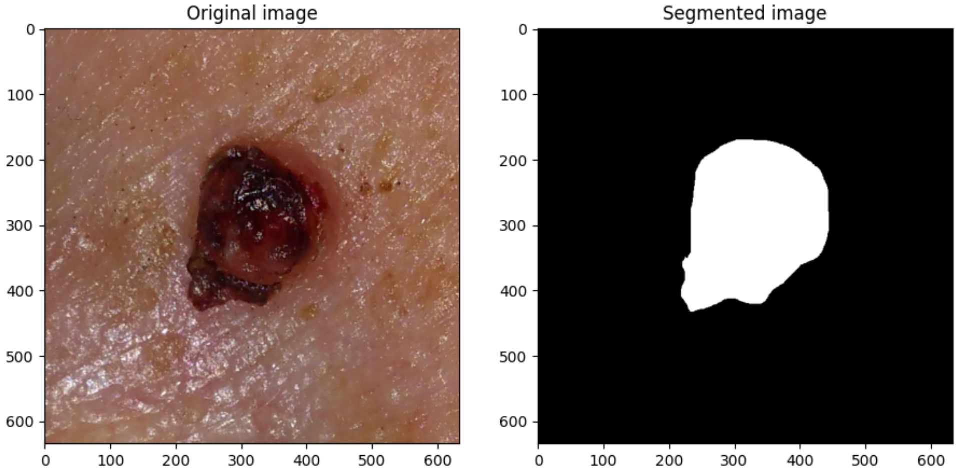

Figure 2: Original lesion image and lesion segmentation

by 1024 pixels. This was to ensure the comparability of the feature measurements since differently sized images of the same lesion could result in different outcomes.

Measuring ABCD features

The diameter feature in the ABCD rating system was excluded from our feature space due to the diversity of angles, distance and quality of the images, invalidating any size measurements based on the pictures themselves. We therefore focused only on asymmetry, colour, and border irregularity.

Asymmetry

The binary masks obtained from the segmentation process were used to assess the asymmetry of each lesion. The masks were folded on 4 different axes: horizontally, vertically, and along each diagonal. The non-overlapping areas of the folded masks were taken to represent the asymmetric parts of the lesion. In practice, this was done by splitting the mask along the axes, reflecting one of the sides, and then subtracting the two sides from each other. The process for the vertical axis is shown in Figure 3 below.

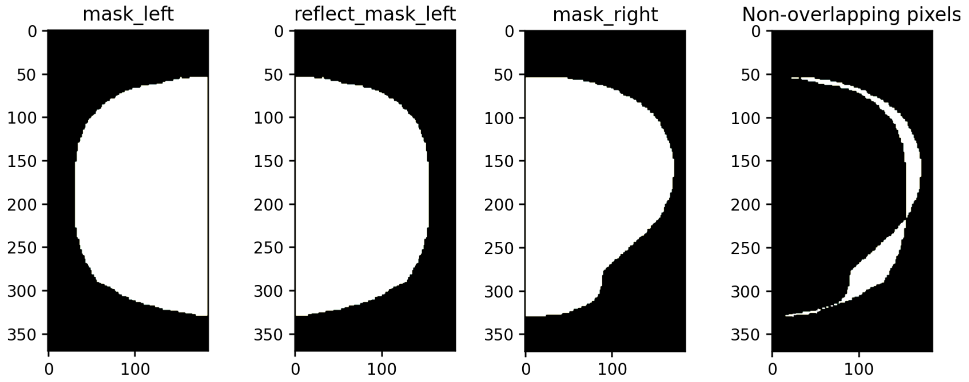

Figure 3: The asymmetry along the vertical axis

The amount of non-overlapping pixels was determined for each axis on which the mask was folded and then divided by the lesion's total area, in order to standardise the measurement. Then an average over the 4 axes was taken as the final measure of asymmetry. This produced an asymmetry score for each skin lesion between 0 and 1, where values closer to 1 represent a higher degree of asymmetry. This can be seen in Figure 4 below.

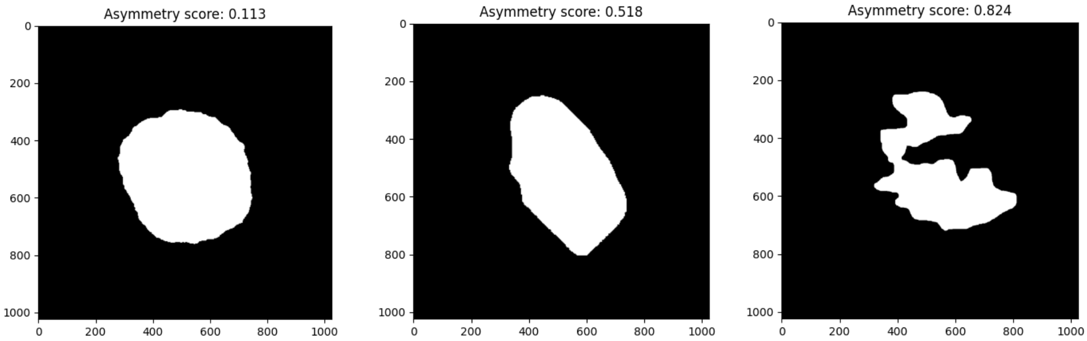

Figure 4: Comparing asymmetry scores of 3 lesions
Scores in order from left to right: 0.113, 0.518, 0.824.

Colour

The masks were also used during the extraction of colour, in order to isolate the lesion itself. The general idea was to have defined RGB colour-values for each of the 6 colours determined by ABCD (white, light-brown, dark-brown, blue-grey, red and black). Then a measure of how prevalent each of these colours was in the image was taken as the final colour feature. In practice, this was done by dividing each lesion into 100 different segments defined by colour similarity, and finding the average colour in each segment. These were then assigned one of the predefined RGB-colours by comparing the segment's average colour to the predefined colours and taking the closest. Finding which colour was closest was calculated using the Manhattan distance. The result was 6 features that represent how much of each colour is present in the skin lesion. In figure 5, we can see a lesion split into segments, where each segment has been coloured both its average colour, and the assigned colour it is closest to.

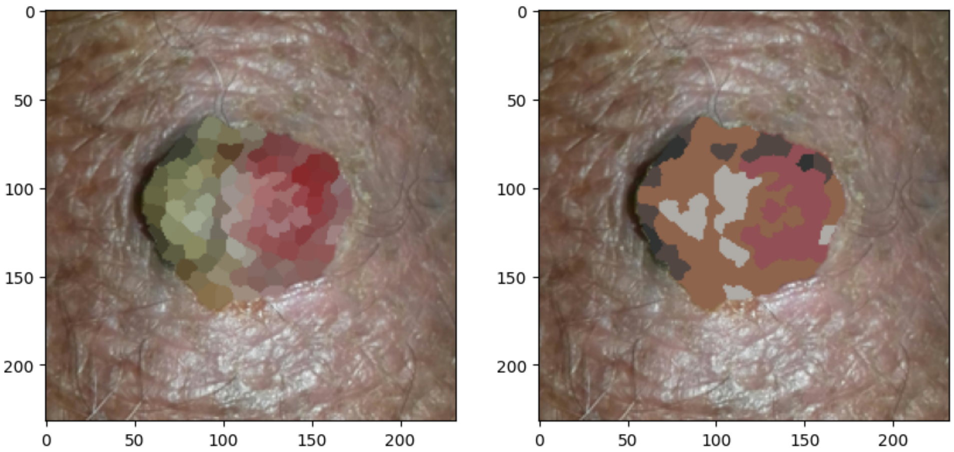

Figure 5: Getting colours from skin lesions by creating 100 segments based on colour similarity. Left: Segments coloured by their average colour. Right: Segments coloured by assigned predefined colours based on closest distance.

Compactness

Finally, compactness was calculated by also using the binary masks. Compactness is given by the isoperimetric ratio, or how close the lesion's shape is to a circle. The lesion's perimeter was extracted by eroding the shape, and the total area was determined by summing the pixels within the mask. These values were then plugged into formula 1, where C is compactness, p is the perimeter, and a is the area.

$$C = \frac{p^2}{4\pi a} \tag{1}$$

The returned value is lower the more circular a shape is, therefore measuring the border irregularity of a lesion. Two examples can be seen in figure 6.

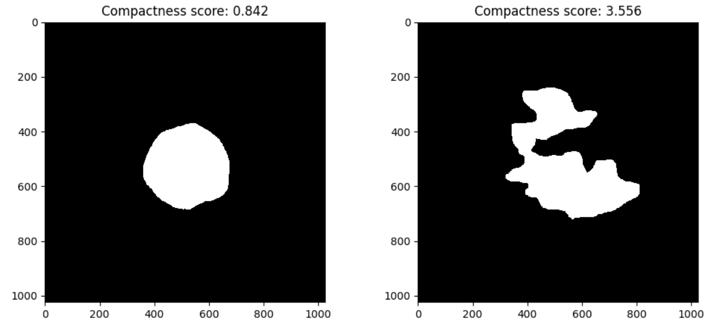

Figure 6: Comparing compactness scores of two lesions
Left: 0.842, Right: 3.556.

General features

For the open question, we used a method to get general features from the images that provided us with several hundred additional features. These were combined with the ABCD features to create a new feature space. We extracted the general features by applying 72 different filters to the lesion images, such as edge and texture detection. To ensure we did not use any data from outside the lesion area, the original lesion was first masked using the masks obtained from segmentation before filtering. Some of the resulting filtered images are shown in Figure 7.

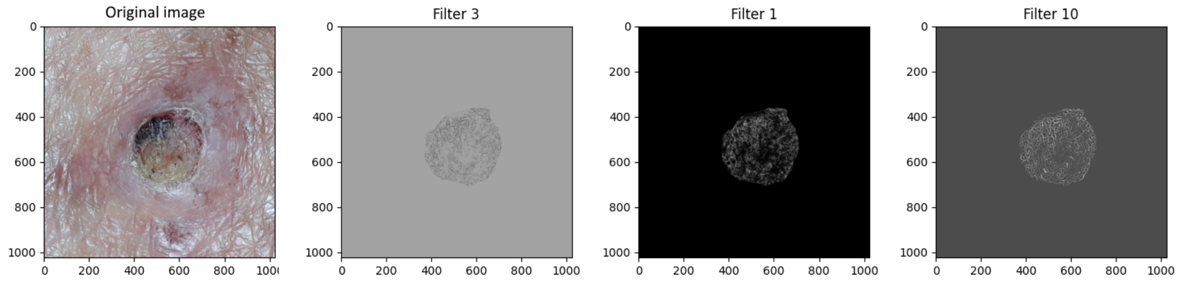

Figure 7: Example outputs from image filters

The pixels in each of these filtered images have varying and measurable levels of intensity. This intensity can be visualised in a histogram, where each bin collects the pixels in some intensity range (Figure 8).

Each histogram includes 9 bins, and each of these bins can represent a feature. Before that, however, the bins needed to be standardised across all images. This was done by finding a representative image and using the bin widths and positions from said image on all images. After doing this, each image only differed in the number of pixels in each bin,

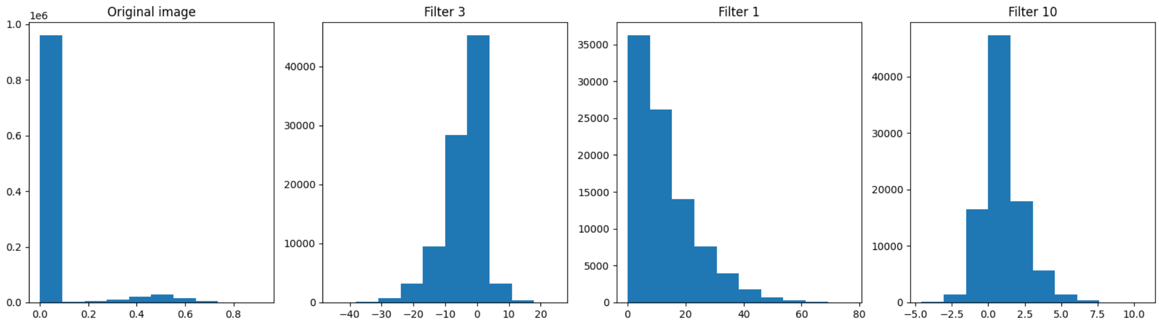

Figure 8: Histograms of filtered image pixel intensities

and this value was then used as a feature. Since there are 72 filters creating histograms, and 9 bins for each histogram, this resulted in the creation of 648 features.

Most of these features are most likely not very valuable, but this is a technique that allows one to create a large amount of features with relative ease and speed. So while not all features are valuable for predictions, some may have good predictive power that could aid our classifier in diagnosing cancerous lesions.

Feature selection

For the open question, we generated 648 additional features for a total of 656 features, in order to investigate if any of them would improve the predicting abilities of our classifier. Since our segmented data only contains 623 samples, a method of reducing our feature count must be employed. We decided to employ a feature selection technique called SelectKbest, which involves evaluating every feature by some scoring function and selecting the k top-scoring features. The scoring function used was the ANOVA F-value, whose score represents how much of the output variance is explained by the features. The number chosen for k was found through hyperparameter optimisation (See "Optimisation").

Classification

To build our classifier we began by splitting the data into 80% training data, and 20% test data. We trained our classifier on the training data using cross-validation and then eventually tested and assessed it based on the test data. We decided to try out two common methods of classification: K-Nearest Neighbours (KNN) and Logistic Regression.

Cross-validation

It is important to validate the classifier so optimisations can be made to improve its performance. There exist several techniques for classifier validation, and we decided to employ cross-validation as it helps prevent overfitting, is scalable, and has low predictive variance. Cross-validation works by splitting the data into k folds. Here it is important to preserve the data distribution across each fold to prevent the introduction of bias, which we ensured by using a stratified split that takes the labels of the data into account. Once the data has been split into k folds, in our case 5, a model is then trained on $k-1$ folds (the training data) and evaluated on the final fold (the validation data). This is one split. This is repeated for $k-1$ more splits, each time training a model and choosing a different fold as the validation fold. This technique is visualised in Figure 9.

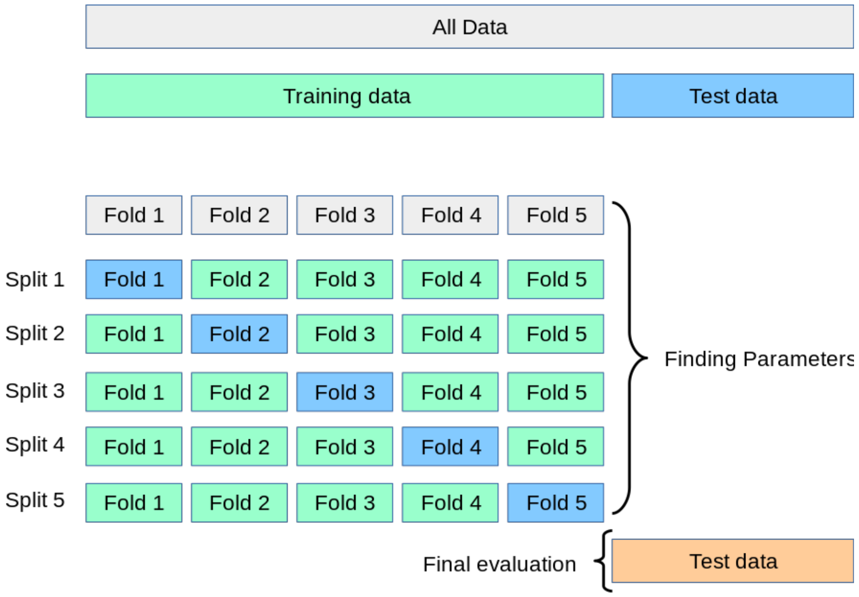
All Data
Training data
Test data
Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Split 1 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Split 2 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Split 3 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Split 4 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Split 5 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Finding Parameters
Final evaluation
Test data

Figure 9: Cross-validation visualised

To validate the models performance, we trained a model on each split, and averaged the results. This gives a good estimation for how the model will perform on new data, and thus helps guide the decisions surrounding the model, specifically hyperparameter optimisation.

Optimisation

Before doing evaluation on the classifiers, there are some variables called hyperparameters that need to be accounted for. The two main hyperparameters are the k best features chosen during feature selection, and the K neighbours used for the KNN classifier. Both of these parameters can have a large impact on the performance of the classifier, so it is important to optimise them to ensure the best possible performance.

The parameters were optimised using trial and error, by comparing the performance of classifiers with different hyperparameter values. To optimise the k features, we measured the F1-score of the classifier for every value of k between 0 and 50. This allowed us to plot a graph of feature count versus score, allowing us to choose the value of k that gave the most improvement. For the K neighbours, we utilised the same method, but used error rate instead of F1-score, as it is more representative of the impact caused by changes in the number of neighbours, K. This meant that we had to choose K such that the error rate was minimal. The process was repeated several times, switching between optimisation of neighbours and features. Using one to optimise the other and vice-versa until they both reached stable optimal values. The final graphs can be seen in Figure 10. This resulted in the following optimised values: 12 top features and 11 neighbours.

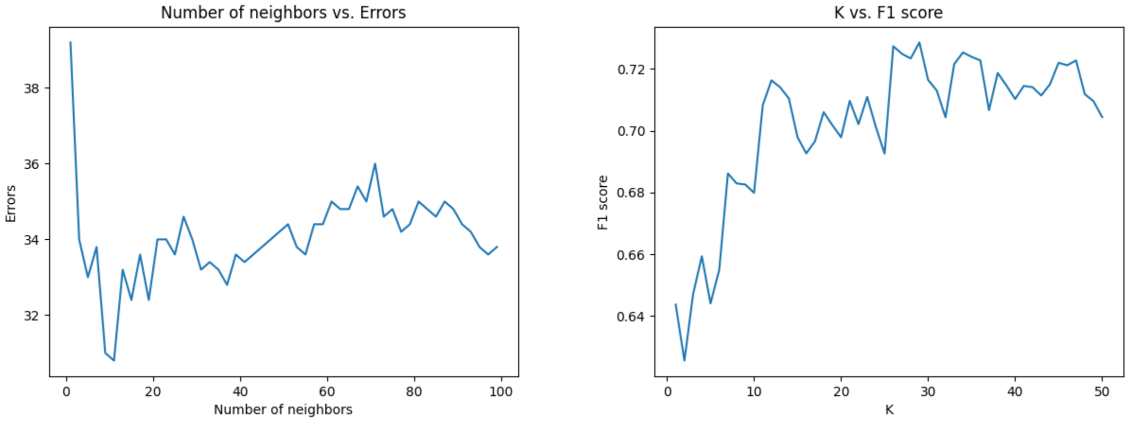

Figure 10: Final optimisation graphs

Classifier selection

The evaluation metrics and confusion matrices for both the KNN and the Logistic Regression classifiers can be seen in Table 1. This is the result of averaging the evaluation metrics of all cross-trained models.

Classifier	**Accuracy**	**Precision**	**Recall**	**F1 score**
KNN	0.647	0.674	0.668	0.671
Logistic regression	0.675	0.698	0.698	0.697

KNN confusion matrix

		True diagnosis	
		Positive	Negative
Prediction	Positive	35.8	17.4
	Negative	17.8	28.6

Logistic Regression confusion matrix

		True diagnosis	
		Positive	Negative
Prediction	Positive	37.4	16.2
	Negative	16.2	29.8

Table 1: Cross-validation results of KNN and Logistic Regression classifiers.

Accuracy and precision are not optimal for this use-case, as they both reward correct classifications with no regard for incorrect ones. In our case, we want to focus on minimising the number of false negatives, since it is preferable to detect cancer when there is none, rather than labeling a cancerous lesion as safe. Therefore, recall and the F1 score were used to assess the classifiers since they both penalise false negatives. All these metrics can also be summarised in a confusion matrix, which displays the number of true positives, false positives, true negatives, and false negatives.

Logistic regression was chosen as the final model since it has a better recall and F1-score. This is also confirmed by looking at the confusion matrices, as all values are improved in the logistic regression model when compared to the KNN model.

Results and discussion

ABCD features

After selecting the logistic regression algorithm as our classifier, we ran it on the testing data. The final evaluation metrics of this classifier are shown in Table 2.

The classifier performs slightly worse on the testing data, but that is to be expected. This may indicate that our classifier is slightly overfitting to our training data. The model has

Accuracy	**Precision**	**Recall**	**F1 score**
0.624	0.656	0.627	0.641

Confusion matrix

		True diagnosis	
		Positive	Negative
Prediction	Positive	42	22
	Negative	25	36

Table 2: Evaluation on testing data for Logistic Regression model.

a recall of 0.627, which means that it correctly identifies 62.7% of the cancerous lesions. Ideally, this would be closer to 100% however, it is not unacceptably low.

Recall is slightly lower than precision, indicating more false negatives than false positives, and this is confirmed by looking at the confusion matrix. This is a bit disappointing as we were aiming for a high recall score to minimise the number of false negatives.

A reason for this could be the insufficient amount of positive samples of each type of cancer in the training data. Melanoma and Squamous Cell Carcinoma were particularly underrepresented compared to Basal Cell Carcinoma, which dominated the dataset. Therefore, the high rate of false negatives could be caused by the misclassification of melanoma lesions. However, confirming this is outside the scope of our investigation.

General features

For our open question we repeated this process for the combined feature space, including the general features and the ABCD features. Feature selection was used to reduce the number of features to 12, all of which were general features. The highest-ranked ABCD feature was the "amount of red colour" and was only positioned as the 16th best.

Both a KNN and Logistic Regression model were trained on these 12 features, and the results indicated that the KNN classifier was superior, a reversal of the results for the classifier using only the ABCD features. This suggests that some characteristic of the general features either caused the KNN classifier to improve and/or caused the Logistic Regression classifier to worsen. It is possible that the general features manifest non-linear relationships with the outcome, and since logistic regression only supports linear solutions this could be why the KNN model performs better.

The evaluation metrics of the general-feature KNN model are shown in Table 3

Accuracy	**Precision**	**Recall**	**F1 score**
0.640	0.662	0.672	0.667

Confusion matrix

		True diagnosis	
		Positive	Negative
Prediction	Positive	45	23
	Negative	22	35

Table 3: Evaluation on testing data for general-feature KNN model.

As we can see, these are slightly better than the ones for the ABCD classifier, albeit with slightly more false positives and fewer true negatives. The classifier has a recall value of 0.672, or 67.2%, which is a 4.5% increase compared to the ABCD classifier. This means we have slightly fewer false negatives which is highly desirable since this means fewer cancerous lesions go undetected. We also see an improvement in all the other evaluation metrics. This could indicate that the general features have good predicting powers that help improve the classification of cancerous lesions. However, because the difference is so small it could also be due to sheer randomness and not due to an improvement in the model.

Limitations and improvements

Although we are moderately happy with the performance of our model, there are still many ways our classifier could be improved. An obvious improvement would be to have a larger and better quality sample to train on. For example, images with more clearly defined lesions, less blur, and taken from the same distance would improve the quality of our feature space. It would also make measurements such as diameter comparable, allowing us to include it in our feature space. Furthermore, we could have experimented with other classifiers that could have been better suited for our purpose.

Our feature creation process also had some limitations that could be improved upon. For example, our colour feature works by comparing the colours found in the lesion to 6 predefined colours. While these colours were based on the colour information given by the ABCD rating system, the exact RGB values chosen to represent each colour were determined based on our own interpretations, and could therefore be biased. Moreover, the asymmetry feature could also be improved by folding the mask over several axes, instead of just the 4 we used. This could reduce the warping effect of the apparent lesion shape that results from pictures being taken at an angle.

Lastly, since we normalised all the features, they all had the same weight in our classifier. However, this may not line up with the professional assessment of dermatologists, as certain features or combinations of features might be more indicative of cancer than others. Therefore, further research into the weight of each feature would help improve the model.

Conclusion and further research

In conclusion, our objective for this project was to investigate how well computerised measures of asymmetry, colour, and border irregularity could be used to predict skin cancer. Our results showed that our classifier has an F1 score of 0.641. This is not a bad result for a skin lesion classifier, especially with the limited resources this project has had. We saw that computer-aided detection of skin lesions is possible to some extent. The performance indicates that it is useful as an early skin cancer warning system, but it cannot be used as a definitive diagnosis tool.

For the open question, we saw that the classifier trained on general, computer-generated features, had better performance than the model trained on features based on the ABCD system. We conclude that this can indicate that the computer may be able to find correlations between skin cancer and some of these general features that would not be possible for humans. As the performance beats that of the ABCD model, this can be a valuable tool, especially if it is developed further. A downside to this classifier, however, is the obfuscated nature of the features, making it impossible to understand or explain how or why the classifier makes certain decisions.

The topic of lesion classification based on general, machine-generated features would be an interesting topic to further investigate, as there are several algorithms that allow features to be extracted from images. Analysing these features and optimising them to detect skin cancer could be a captivating exploration.

However, it is important to acknowledge that computer-generated diagnoses cannot, as of today, replace the expertise of medical professionals. If one becomes suspicious of a skin abnormality, it is highly recommended to seek professional evaluation in order to ensure an accurate diagnosis.

References

Skin Cancer Survival Rate. (n.d.). Moffitt Cancer Center.
https://moffitt.org/cancers/skin-cancer-nonmelanoma/survival-rate
Retrieved June 1., 2023 from
http://web.archive.org/web/20230601141244/https://moffitt.org/cancers/skin-cancer-nonmelanoma/survival-rate

Tas, B. (2020). Faster Evaluation of ABCD Rule and Total Dermoscopic Score for Nevomelanocytic Lesions: Dermoscopic Score Scale. Journal of Skin and Stem Cell, 6(2).
https://doi.org/10.5812/jssc.102138
Retrieved June 1., 2023 from
http://web.archive.org/web/20230601141554/https://brieflands.com/articles/jssc-102138.html

SkinVision Review: Dermatology App Under The Microscope. (2022, May 18). The Medical Futurist.
https://medicalfuturist.com/aspiring-dermatology-app-under-the-microscope-the-skinvision-review/
Retrieved June 1., 2023 from
http://web.archive.org/web/20230601141759/https://medicalfuturist.com/aspiring-dermatology-app-under-the-microscope-the-skinvision-review/

A Warning Against Using Phone Apps To Detect Skin Cancer. (2022, December 8). Memorial Sloan Kettering Cancer Center.
https://www.mskcc.org/news/warning-against-using-phone-apps-detect-skin
Retrieved June 1., 2023 from
http://web.archive.org/web/20230601141003/https://www.mskcc.org/news/warning-against-using-phone-apps-detect-skin

Appendix

Annotation guide

The annotation guide was based on the ABCD rating system, which is scored as shown in Table 4 below (Betul Tas, 2020).

ABCD, Features	Value	Weight Factor	Subscores
Asymmetry		1.3	0 - 2.6
Symmetric	0		
One-axis asymmetry	1		
Two-axis asymmetry	2		
Border irregularity		0.1	0.5 - 3
0 to 8-border	0 - 8		
Colour (one point for each)		0.5	0.5 - 3
White, light brown, dark brown, blue-grey, red, black	1 - 6		

Table 4

Manual annotation results

In Table 5 below we have the annotation results from our manual annotation process. We annotated 10 images in total, where each image was annotated by two annotators separately. We visually assessed the asymmetry and colour for each image and gave them scores in line with the ABCD rating system.

Images	Annotator 1		Annotator 2		Annotator 3		Annotator 4	
	Asy.	Col.	Asy.	Col.	Asy.	Col.	Asy.	Col.
PAT_115_1138_870.png	1	1	2	2				
PAT_902_1718_852.png					1	2	0	2
PAT_115_177_575.png	2	3					1	2
PAT_135_202_593.png			1	1	0	1		
PAT_155_240_211.png	1	3	0	2				
PAT_1298_1050_519.png					0	1	0	1
PAT_1439_1522_211.png	1	2					1	2
PAT_1415_1437_743.png			2	1	2	1		
PAT_1481_1674_735.png	2	1					2	2
PAT_1414_1433_570.png			2	2	2	4		

Table 5