IT UNIVERSITY OF COPENHAGEN

A Comparative Study of Feature Robustness and Sex Bias in Logistic Regression and CNN for Skin Cancer Detection

Submitted by
Nikolette Pedersen, *nizp@itu.dk*
Regitze Sydendal, *resy@itu.dk*
Andreas Wulff, *lawu@itu.dk*

Supervisor
Veronika Cheplygina, *vech@itu.dk*

6th Semester, Spring 2023

Course Code: BIBAPRO1PE

IT University of Copenhagen
May 14, 2024

Robustness and sex differences in skin cancer detection: A case study using CNN and logistic regression

IT University of Copenhagen, Denmark
Nikolette Pedersen, Regitze Sydendal, Andreas Wulff
{nizp, resy, lawu}@itu.dk

Supervisor: **Veronika Cheplygina**
vech@itu.dk

Abstract

The detection of skin cancer using convolutional neural networks (CNNs) has reached high results, yet challenges regarding the reproducibility of results and biases in model training is still an acknowledged problem. This study extends prior work of Petersen et al. 2022 of studying robustness in two distinct models: logistic regression (LR) and CNN, which they did in regards to Alzheimer's disease. We explore the bias in sex in skin cancer detection, using the PAD-UFES-20 dataset composed of 2298 smartphone images of skin lesions. Our two models consist of: an LR model trained on manually selected features from medical guidelines, ABCDE and the 7-point checklist, and a pre-trained ResNet-50 model that will process raw images of the lesions. We evaluate these models in alignment with Petersen et al. 2022: across multiple training datasets with varied sex-composition to determine their robustness. Our results showed that both the LR and the CNN were robust to the dataset shift, but the results also revealed that the CNN had a significant higher accuracy (ACC) and area under the receiver operating characteristics (AUROC) for males than for females. We wish for these findings to contribute to the growing field of investigating potential bias in popular medical machine learning methods. The data and relevant scripts to reproduce our results can be found in our Github[1].

1 Introduction

The classification of skin lesions using machine learning methods aims to enhance the accuracy (ACC) and accessibility of diagnosing skin lesions, such as melanoma and other skin conditions. Over the last years ACC scores have been reached that excel the ACC of dermatologists (Pham et al., 2021). However, the reporting of very high ACC scores over the recent years of detecting skin cancer, lacks investigation in bias across different skin types and sex. Groh et al., 2021 annotated the Fitzpatrick 17k dataset, which consists of clinical images sourced from two dermatology datasets with Fitzpatrick skin type labels. The Fitzpatrick scale classifies skin colours into skin types that range from 1-6 with 1 being the lightest skin type and 6 being the darkest. Based on these labels they find that their deep learning neural network was more accurate on skin types similar to those it was trained on. Sies et al., 2022 investigated the performance difference in sex related to skin cancer detection, and they found that there was no difference in performance. However, they highlight that we cannot rule out bias yet for all AI-based dermoscopic skin cancer classification systems. They also stated that research of sex-related differences of using convolutional neural network (CNN) performance is something that is yet to be addressed in the field of skin lesions.

Therefore, we have investigated the sex-related performance and robustness in regards to two models for skin cancer detection, a logistic regression (LR) with handcrafted extracted features and a CNN trained on raw images of skin lesions.

Our experimental setup will follow Petersen et al., 2022 who conducted a study regarding Alzheimer's disease and the feature robustness of certain models. They found that while LR, which was trained on manually selected features, is robust to different dataset compositions, CNN generally improved its performance for both male and female subjects when including more female subjects in the training dataset. This was done by training on several imbalanced sex-composition datasets and then evaluating the models' performance on a sex-balanced dataset.

The data that will be used in this study is the PAD-UFES-20 dataset, which is composed of smartphone images of skin lesions, and not dermoscopic images. Pacheco and Krohling, 2020 devel-

[1] GitHub

oped an application in collaboration with the Dermatological Assistance Program (PAD) at the Federal University of Espírito Santo (UFES), hence the name. Therefore, further research into this topic does not only lead to valuable findings, but also increases the accessibility to healthcare, particularly for those who have limited access to certified dermatologists, which is why they developed this application.

The LR, which relies on manually selected features, uses methods produced by former first-year data science students which extracts the key visual characteristics from two different medical guidelines: the ABCDE method and the 7-point checklist, more details in Table 1. The second model used is a pre-trained PyTorch ResNet-50 (resnet50), afterwards trained on the raw images of the lesions. The evaluation of the models will be done in compliance with the methodology described by Petersen et al., 2022.

2 Related work

In the growing field of utilising medical imaging as a diagnostic aid, deep learning has become an important tool in enhancing and speeding up the diagnostic processes. This can be especially beneficial for low- and middle-income countries, where access to certified dermatologists might not be directly available. Recent research has showed the capability of these machine learning models including CNNs, to achieve diagnostic ACC levels comparable to healthcare professionals in various medical domains (Liu et al., 2019).

Nonetheless, despite the achieved diagnostic ACC levels, the field is not without its challenges, especially concerning bias in skin cancer detection methods. (Guo et al., 2022) reported that out of 1408 records, only 136 studies met their inclusion criteria when it came to reporting diverse skin types, which are already underrepresented in machine learning research for skin cancer detection. Sies et al., 2022 described that although the CNNs approval rates for clinical use are increasing, sex-related differences of CNN diagnostic performance is yet to be assessed.

But still, the issue of bias in AI models is increasingly acknowledged and it is getting more popular to explore ways to mitigate biases (Li et al., 2021). Petersen et al., 2022 explores the robustness of classification models in the context of MRI-based Alzheimer's disease detection. They found that LR models trained on manually selected features showed robust performance, while the performance of CNNs improved with more female representation in the training data for both females and males, showing that the CNN was less robust when meeting a skewed training dataset.

This bachelor thesis aims to extend these themes and build upon this research. Motivated by this, we investigated the robustness of classification models in the context of detecting skin cancer in lesions. In 2020 A. proposed to use a CNN as a skin cancer detection system. He used the HAM10000 dermoscopy image database (Tschandl, 2023), and employed 3400 images consisting of multiple diagnoses. He chose to group the different skin lesions into binary classes, malignant and benign. The CNN received only raw images of the lesions, no segmentation or preprocessing was applied. Still, a classification ACC of 84% was achieved. This setup of using the raw images and binary classification will be implemented in our study. It will also allow us to follow the experimental setup from Petersen et al., 2022.

Looking outside the medical deep learning community, the medical field has developed several guidelines to identify cancer, and more specifically melanoma. The two medical guidelines which will be the base for the manually selected features is the ABCDE method and the 7-point checklist method, which are both defined in Table 1. These two guidelines are developed to look for certain characteristics in regards to detecting melanoma, but since the skin cancers present in the dataset carry some of the same characteristics, such as irregular texture and color change, these methods were used regardless (Cancer Council Victoria).

3 Data

3.1 Original skin lesion data

3.1.1 PAD-UFES-20

The PAD-UFES-20 dataset consists of 2298 images and corresponding metadata. Each row in the `metadata.csv` represents the metadata of an image of a skin lesion with multiple features. There are 1641 unique skin lesions, and 1373 patients in the dataset. There are 6 different diagnoses, as can be seen in Table 1. Grouping them as such, will result in a binary classification task, skin disease/skin cancer. Any missing

ABCDE method	**Dermatological guideline for detecting melanoma.** A: Asymmetry, B: Border, C: Colour, D: Diameter, E: Evolution (Corewell Health).
7-point checklist	**A relatively new dermatological guideline for diagnosing melanoma by simplifying the standard pattern analysis.** 1. Atypical pigment network, 2. Blue-whitish veil, 3. Atypical vascular pattern, 4. Irregular streaks, 5. Irregular pigmentation, 6. Irregular dots/globules, 7. Regression structures (DermLite).
Skin diseases	**ACK**: Actinic Keratosis, **NEV**: Nevus, **SEK**: Seborrheic Keratosis (Pacheco and Krohling, 2020).
Skin cancers	**BCC**: Basal Cell Carcinoma, **SCC**: Squamous Cell Carcinoma, **MEL**: Melanoma (Pacheco and Krohling, 2020).

Table 1: Background terms related to dermatological conditions and guidelines

data entry in the feature `gender` was removed from the dataset. Moving forward, we chose to define the column gender as sex, to keep it binary. Council Of Europe has defined sex as: *"the different biological and physiological characteristics of males and females, such as reproductive organs, chromosomes, hormones, etc."*. We also removed any duplicates of `lesion_ids` and kept the first occurrence. The final dataframe had 1179 data entries. Figure 1 shows an imbalance in the dataset, with a majority of people with cancerous skin lesions. This will cause challenges further on, which will be accounted for accordingly (Hvilshøj, 2022).

The dataset is also imbalanced in its distribution for skin colour, classified using the Fitzpatrick scale. The distribution is skewed, with a majority towards lighter skin types. The lack of darker skin types in the dataset will be brought up in the discussion. The graph showing the distribution can be found in Appendix A.

3.1.2 Dataset implications

When exploring the data we found some inconsistencies, such as the `metadata.csv` containing `lesion_ids` belonging to multiple patients. We found 45 instances of the aforementioned case. The number of unique skin lesions increase from 1641 to 1686. We also found issues with instances of patients having different `lesion_ids`, but the images of the lesions were the same. Our supervisor contacted the authors Pacheco and Krohling and got confirmation that these instances were indeed mistakes in the dataset. We changed all the wrong `lesion_ids` to a unique value and progressed with the modified data saved under the name `fixed_metadata.csv`. To find more details about the mistakes we found, refer to Appendix B.

3.2 Crowd sourced segmentations and features

3.2.1 Feature extraction method

We got access to 14 former first-year data science students' GitHub repositories in order to get their code for their handcrafted methods. The resources were gathered through the course *First Year Project* at the IT-University of Copenhagen, managed by Veronika Cheplygina, which is also to whom we will be referencing, when referring to the students' code (Cheplygina, 2023).

To ensure anonymity, we anonymised all the students and assigned each group with a unique number that we chose. Afterwards, we evaluated each group's project, and we did so using these factors: the project's references, the degree of described details and the feasibility of implementing and understanding the code.

We chose three different projects: two projects that used the ABC(DE) method[2] and one project using the 7-point checklist, Table 1. We chose two ABC(DE) methods due to the majority of the student groups preferring this method. Choosing two approaches to compare and pick features from, increased the likelihood of us finding high-quality implementations. We chose the 7-point checklist to gain more insight on different characteristics that the ABCDE method does not account for. The groups we ended up using code from, were Group

[2]The ABCDE method, in relation to the students' code, will be refereed to as ABC(DE), due to the students not building methods for measuring the diameter or the evolution in the lesions.

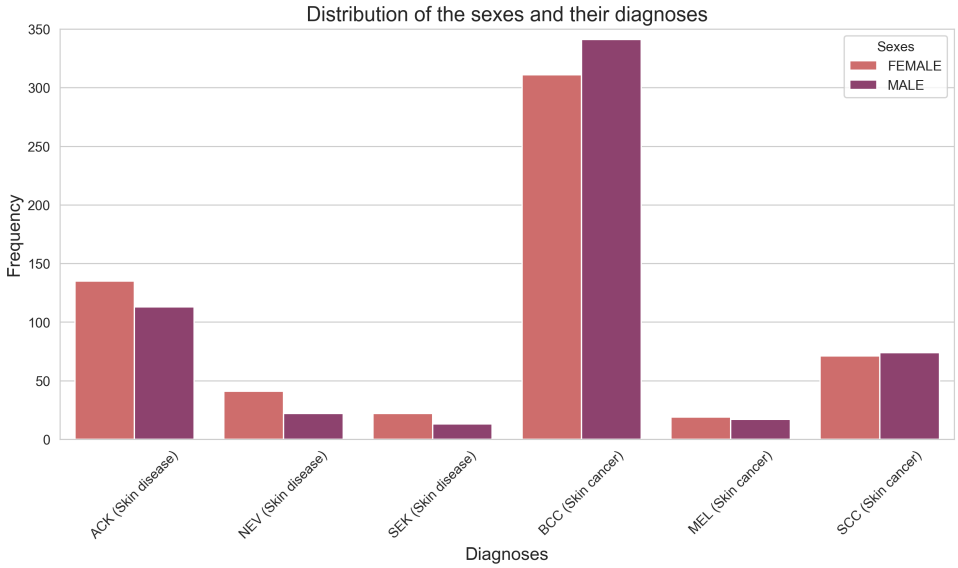

Figure 1: Final dataframe distribution of the sexes and their diagnoses

4 and 9, who both used the ABC(DE) method and Group 5 who used the 7-point checklist. After we chose the suitable student code, we made a new script combining all their methods, in order to be able to run them all at the same time.

3.2.2 Segmentations

Segmentations were required to utilise the data science students' handcrafted feature extraction methods. The segmentations of the lesions were provided from current first-year data science students, and to obtain all the student-segmentations, we had to manually access all the groups' GitHub repositories. All but 200 segmentations were provided and the missing ones were done by us. All of this was made using Label Studio (Label Studio). The segmentations are all available in our GitHub.

4 Methodology

4.1 Upsampling data

We are dealing with a sparse and imbalanced dataset. By splitting our data like Petersen et al., 2022, our training-validation sets would have a constant skin cancer/skin disease fraction of $293/122$, or 29.4%. Because the datasets were skewed to such a degree, we anticipated that the models would simply default to predict skin cancer. To deter this, we upsample the skin disease data by augmenting all of the non-cancerous skin lesions once. The techniques we used to augment the data was flipping the image, randomly sharpen and adding Gaussian blur (PyTorch). We chose these techniques, as they augment the images without introducing excessive distortion or changing the colour. This was deemed especially important for future works regarding the Fitzpatrick scale. Using this, our training-validation datasets achieved a constant fraction of $293/243$, still in the favour of cancerous data. This translates to roughly 45.4% of non-cancerous lesions. The feature extraction was run for the same augmented images, as the CNN was trained on.

We also implemented and tried the synthetic minority oversampling technique (SMOTE) for the LR, which creates additional synthetic data by using a K-nearest neighbour algorithm (Imbalanced Learn). The reason for not using SMOTE was due to the difficulty of comparing the LR and the CNN, when using two different upsampling methods. We therefore decided on using augmented images to upsample our data for both models.

4.2 Dataset splits

In order to be able to evaluate the performance of the models, we followed the setup of Petersen et al., 2022, which allows us to end up with 125 training-validation sets with 5 test sets for the LR. The CNN had 50 training-validation datasets with 2 designated test sets. This was chosen due to the long training time for the CNN. It took approximately 20 hours to train for 50 training-validation sets. Each training-validation dataset would then contain these sex-ratios in regards to the amount

of females, $0\%, 25\%, 50\%, 75\%$ and 100%.

The following paragraph will be a quick walk-through of how the splitting works to utilise our data the most. Firstly, the data is split into 4 categories being `cancer_female`, `non_cancer_female`, `cancer_male` and `non_cancer_male`. 26 data subjects of each category will be assigned into each of the 5 test sets, so each test set consists of 104 data subjects. For each test set, there will be 5 training-validation datasets, and they will consist of all the remaining data in addition to the test dataset that does not correspond to their own test dataset. The 5 training-validation datasets consist of the 5 ratios mentioned earlier for females, so $0\% - 100\%$, with increments of 25%. We apply 5 folds to each training-validation set, which gives us 25 combinations of the training-validation datasets for all test sets. In total, it sums up to 125 training-validation datasets. When evaluating the model the test dataset will also be split into female and males, to see the difference in performance for every ratio separately. Figure 2 visualises how the test data and the training-validation data for `test_set_1` is being made for the first ratio 0%. The LR is then parsed these data splits with the extracted features. The CNN which requires images, matches the `img_id` from the metadata to the corresponding image.

4.2.1 Own implementation

To create the splits of the data, we used the code of Petersen et al., 2022, but due to our data being different from theirs, we had to do a lot of modifications to their code, to split our data in a similar way and account for various issues regarding the way we upsample data.

The first modification we did, was making sure there was no data leak. Since we have instances where patients have multiple skin lesions, we ensured that patients of such cases in the test set, were not included in the corresponding training set. Edge cases occurred where patients had both a cancerous and a non-cancerous lesion. Due to the way the data is split, this would bypasss our earlier modification, resulting in a possible data leak, which we have also rectified. Another thing we had to take into account is when we added the augmented images to our data splitting. When we augmented our data, we made sure that no augmented data would be in the test sets, as we try to keep the test set as realistic an approximation of reality as possible (Barreto, 2022).

Lastly, we made sure that the original image and the augmented image were either in different training-validation set ratios, or if they happened to be in the same ratios that they were assigned to the same fold, in order not to have them in corresponding training-validation sets. This was to make sure that during cross-validation, the model would not validate itself on data it had seen before.

4.3 Handcrafted methods and feature selection

The groups each provided different ways to measure the same features. The first two groups used different approaches to calculate the ABC(DE) methods, and the third group provided ways to measure the 7-point checklist. Not all of their features were used. We will shortly elaborate on some of the chosen features and some which were not.

4.3.1 Group 4

Multicolour rate Group 4 developed a method that focuses on analysing the multicolour rate of a skin lesion. They calculated the multicolour rate as the Euclidean distance between the two most common colours in the segmented part of the skin lesion.

RGB channels and average HSV Another aspect of their method involved the use of RGB values to create channels representing averages of each color value. To measure the average hue, saturation, and value, they converted the image from RGB-space to HSV-space.

Compactness They also measured the compactness of the skin lesions to analyse the border. Compactness is defined by the formula:

$$c = \frac{l^2}{4\pi A}$$

where l is the length of the border and A is the area of the mask, previously referred to as the student-segmentations. The closer the compactness value is to 1, the more circular the lesion is. Values greater than 1 indicate more irregular borders. The mask images are binary arrays, where the masks are represented by 1's and the area outside of the mask is represented by 0's. They performed calculations by summing together the masks' array and the outline of the masks' array, and subtracting the mask of the lesion from the outline of the mask to obtain the result.

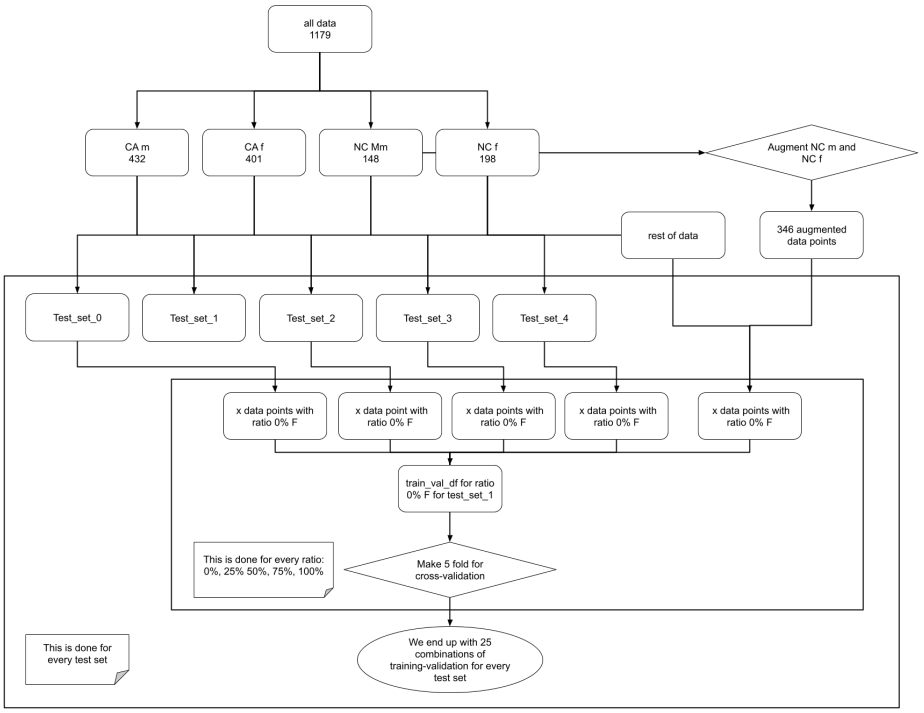

Figure 2: Flowchart simulating the data split

4.3.2 Group 9

Asymmetry Group 9 evaluated the lesions based on the best possible asymmetry, the worst and the mean asymmetry. They used the masks of the lesions and removed excess pixels around the mask, so their array was as small as possible. Then, they folded the array around its center, and only the areas where both axes of the lesion are asymmetrical were summed. They have done this both horizontally and vertically and the sum of those values are then divided by the total area and multiplied by 2, to give an asymmetry score between 0 and 1. The image is then rotated by 22.5°, and they continue these steps and keep rotating the image until they have reached 90°, and finally, they get the mean asymmetry.

Relative color features They also used relative colours as features, referred to as F1, F2, F3, F10, F11 and F12. They used these features as they can compensate for possible variations in the image that are caused by illumination. They followed the formulas of Celebi et al., 2008.

Dominating HSV Another way they looked at colours in the lesion was by extracting the dominating hue, saturation and value. They did this by converting the image into a HSV-space, flattening the image and then applying a K-means model to cluster the colours of the lesion.

Variance of HSV The variance of hue, saturation and value were also extracted. They did this by converting the image into a HSV-space again, extracting the HSV-mean and afterwards calculating the variance of HSV.

4.3.3 Group 5

Atypical pigment network For the first feature, atypical pigment network, Group 5 uses OpenCV's image processing abilities, so the function extracts and enhances the relevant features.

Blue whitish veil The second feature, blue whitish veil, finds the blue veil of a skin lesion. To do this it loops through each pixel of the image, and looks for pixels where the RGB-value satisfies the following predicate: $B > 60, R - 46 < G < R + 15$, (Madooei et al., 2015).

Atypical vascular pattern The third feature is the atypical vascular pattern, which was calculated by measuring the amount of red pixels in an image.

Irregular streaks The fourth feature is irregular streaks, which they extract by taking the grayscale image, and then applying a threshold that detects

the contours, so that they can compute the border of the lesion.

Irregular pigmentation The fifth feature is irregular pigmentation, wherein the image was grayscaled again and applied the Otsu thresholding (Gopalakrishnan, 2023) to create a binary mask. They identified the regions with circularity below 0.6 as suggestive of irregular pigmentation.

Irregular dots/globules The sixth feature is irregular dots/globules. It is found by measuring the presence of globules, or dots, in a lesion by implementing blob detection technique from Scikit-Image.

Measurement of regression structures The seventh feature is measurement of regression structures. Group 5 transforms the image colour format to HSV and define a lower and an upper bound. They create masks and count the pixels within these bounds (Cheplygina, 2023).

4.3.4 Feature selection

For the LR, we chose to not use all the features, as we did not want to overfit our model(RITHP, 2023). Before we chose the final features, we had to exclude all the features that were highly correlated with each other, as they are deemed redundant. We used the Pearson correlation coefficient to quantify the correlation. We therefore defined a threshold of 0.80, so features with a correlation above the absolute value 0.80 would be excluded in our feature selection. However, we wanted to remove the features with the least correlation with the target feature, `is_cancerous`, which are our gold labels. In order to determine which features to exclude we started by sorting the features based on their interconnectivity. The ones which had a correlation above the threshold with the majority of other features were chosen. The one with the lowest target correlation amongst those was then dropped. This meant that some features, which previously had correlation with the now dropped feature, no longer had that interconnectivity and became viable. This was repeated until no features with correlation above our threshold remained. This approach meant that only one feature with high correlation with only the target feature was removed, namely pigment network coverage, as this had correlation with the more valuable blue whitish veil. After removing the redundant features, we chose the 10 features with the highest correlation with our gold labels. Those were: `F2`, `F11`, `sat_var`, `blue_veil_pixels`, `avg_green_channel`, `mean_asymmetry`, `F1`, `average_hue`, `compactness_x` and `dom_hue`.

5 Experiment

5.1 Experimental setup: Logistic regression

For our experimental setup, we chose to use scikit-learn's `LogisticRegression`. We made a pipeline that would first standardise the features, and then we fed these standardised features to the LR. As we deemed LR to be a simple, yet powerful model, we also chose to use `GridSearchCV` to search over an exhaustive list of the best parameters for our model, to help improve the performance. `GridSearchCV` has its own cross-validation splitting strategy, but our objective was to use the aforementioned splits, so we had to use scikit-learn's `PredefinedSplit` to preserve our own folds (scikit-learn). As we are training an LR model on 125 training-validation sets, we chose to use MLflow to log all the models, their corresponding best combination of parameters and the metrics: ACC, precision, recall, F1-score and area under the receiver operating characteristics (AUROC) (Narkhede, 2021). Our metrics were chosen based on Pacheco and Krohling's paper. They used some of these metrics for their CNN, which we do as well, to make it easy to compare the two models. Although our dataset is no longer imbalanced, due to our upsampling with augmented images, we still decided to keep these metrics and then discussed the results we found interesting.

5.2 Experimental setup: CNN

We decided to use the specific CNN-model ResNet-50. We chose this CNN because Pacheco and Krohling, 2020 used a ResNet-50, and it was the best performing model out of all the CNN-models, in terms of ACC. The same metrics were also logged on MLflow as done with the LR. We used a ResNet-50 from the PyTorch library (resnet50). First step is loading the ResNet-50 model with pre-trained weights. The weights we chose were the default ResNet-50 weights supplied by PyTorch, as these were recommended for image classification. The next step was to modify the last layer, which was originally designed for a different classification task, in order to adapt

Metrics	LR	CNN
Mean ACC, f	0.682 ± 0.059	0.650 ± 0.057
Mean ACC, m	0.684 ± 0.080	0.687 ± 0.053
Mean AUROC, f	0.745 ± 0.052	0.716 ± 0.075
Mean AUROC, m	0.727 ± 0.081	0.758 ± 0.059
Mean Precision	0.666 ± 0.052	0.655 ± 0.065
Mean Recall	0.743 ± 0.069	0.763 ± 0.162
Mean F1-score	0.701 ± 0.047	0.689 ± 0.081

Table 2: The mean of the metrics and their standard deviation for the LR and the CNN

the pre-trained ResNet-50 model to binary classification. The third step was to preprocess the dataset. This was done by resizing the images to 224x224 pixels, which is the input size expected by the ResNet-50 architecture. The images are then converted to tensors which can be processed by ResNet-50 and normalised according to standard and mean values obtained from Imagenet, which is standard practice. The fourth step is creating `Datasets` and `Dataloaders`. These are implemented by obtaining metadata from our data split and collecting the images corresponding to the `img_ids` in the metadata. This data is then processed through the third step.

The fifth step is defining the loss function and optimizer. The loss function used is a BCEWithLogitsLoss, which is typically used when having a binary classification task. It is an inner sigmoid function with an outer cross entropy loss function. The optimizer chosen is the Adaptive Moment Estimation (ADAM) which is a commonly used optimizer from the PyTorch library with a learning rate of 0.001. Finally, the model is trained for 10 epochs, and the training data is fed to the model in batches of 32, with shuffling enabled to ensure diverse gradient updates.

6 Results

Table 2 shows the mean of the different metrics for the two models we logged in MLflow.

Figure 3 and 4 illustrate the LR and CNN respectively. They are trained on the sex ratios, but then tested separately on female and male. They are then evaluated by their ACC and their AUROC. The x-axis shows the sex ratios, and the y-axis are the scores for the models' metrics. To identify the significance of the models performance, we fit a regression line on the data and performed a regression t-test to see if the slope could determine any kind of incline or decline in performance. Here, we are dealing with multiple comparisons, female/male, CNN/LR, and ACC/AUROC, which means we end with a total of 8 tests. We are therefore using a Bonferroni correction to avoid Type I errors. Our correction factor is 8 and the corrected threshold is therefore $\alpha = 0.006$. m is the slope of the linear regression and the null-hypothesis is $m = 0$, μ is the average ACC and AUROC (± standard deviation) across all sex ratios. Looking at the p-values across all tests done with both models, nothing is under the threshold and therefore, not statistically significant. The statistical tests and their results can be found in Table 3.

For females, Figure 3 shows that the lowest standard deviation for ACC was achieved at ratio 0.5. At ratio 1.0 some of the lowest and highest ACC scores were achieved. For the AUROC scores, the lowest standard deviation were achieved with a ratio of 0.0 and it increases for every ratio and reaches the highest at ratio 1.0 for females. The same pattern for AUROC male is not observed, as the scatter seems fairly consistent across all ratios. The lowest standard deviation for males in regards to ACC was also achieved at 0.5.

In Figure 4, we see for all plots that the smallest standard deviation is at ratio 0.5. The highest ACC scores for females are reached in ratios above < 0.5, whereas with males, the highest ACC scores are across all ratios. The CNN's mean ACC for males was 0.687, and the mean ACC for females was 0.650. We chose to perform a Mann-Whitney U test, a non-parametric test, to compare our two groups, females' and males' mean ACC and AUROC scores (SciPy). We found that it is statistically significant that the mean ACC and AUROC for males are higher than it is for females, as can be seen in Table 3. It had a p_{value} of 0.0006 and a p_{value} of 0.007 respectively, which is well below our threshold, which for this test is 0.025, due to our comparison of the two sexes for the tests.

7 Discussion

7.1 Summary of results

The results showed us that for both LR and CNN, the null hypothesis of $m = 0$ could not be rejected, as none of our tests are statistically significant. However, for the CNN, we found that the mean of the ACC and AUROC scores are statistically sig-

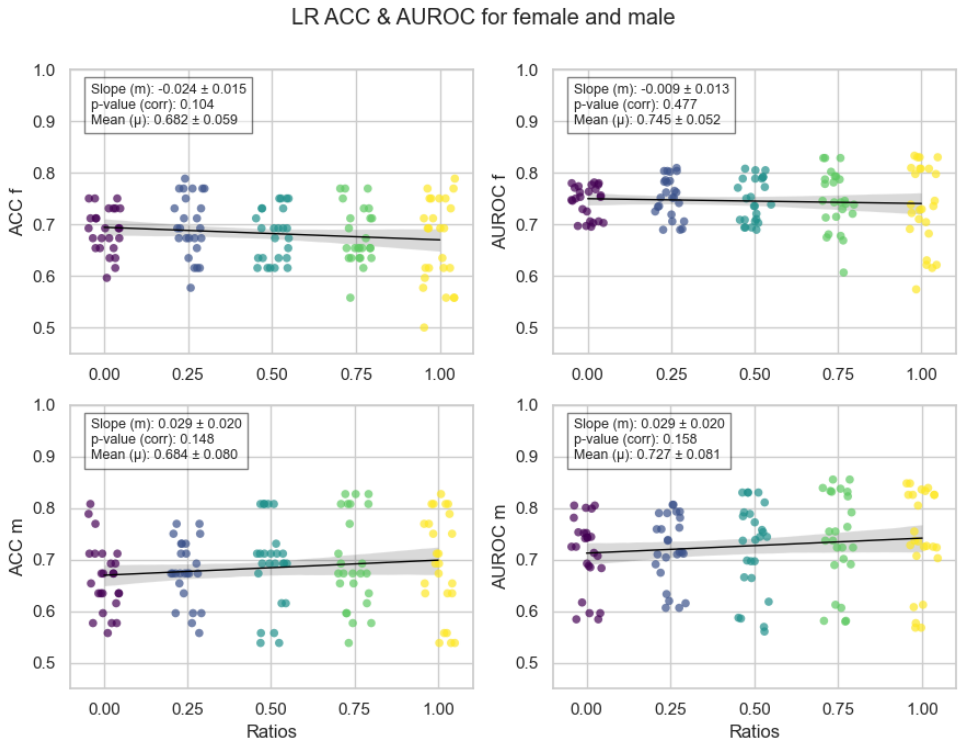

Figure 3: ACC and AUROC, male and female, for LR

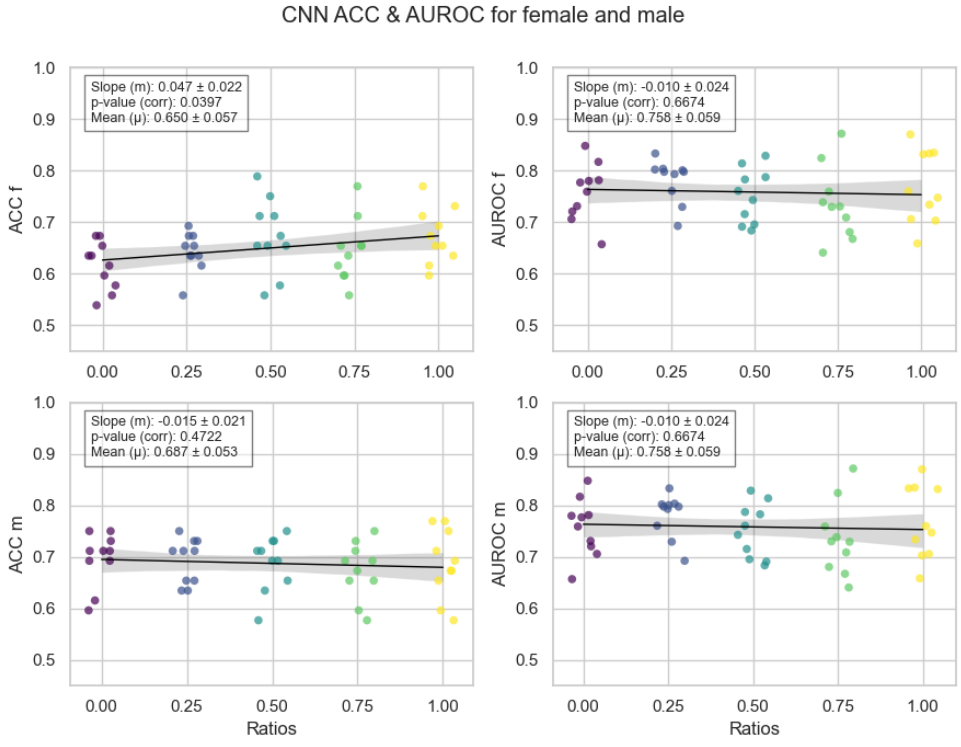

Figure 4: ACC and AUROC, male and female, for CNN

Statistical test	Null Hypothesis (H0)	*p*-value
Threshold: $\frac{0.05}{8} = 0.006$ H_0: $m = 0$, μ is the average ACC and AUROC (± standard deviation)		
LR ACC, f		0.104
LR AUROC, f		0.477
LR ACC, m		0.148
LR AUROC, m		0.158
CNN ACC, f		0.0397
CNN AUROC, f		0.3374
CNN ACC, m		0.4722
CNN AUROC, m		0.6674
Threshold: $\frac{0.05}{2} = 0.025$ H_0 for mean metrics: $\mu_{\text{f}} = \mu_{\text{m}}$		
Mean LR ACC (f & m)		0.904
Mean LR AUROC (f & m)		0.219
Mean CNN ACC (f & m)		0.0006
Mean CNN AUROC (f & m)		0.007

Table 3: The mean of the metrics for the LR and the CNN, along with their null hypotheses

nificantly larger for males than for females, which was not the case for the LR.

7.2 Analysis of results

We could not reject the null hypothesis (that $m = 0$), and that is not a surprise given previous research (Sies et al., 2022), which has shown that sex does not matter in the context of classifying skin lesions. Even though, we could not reject the null hypothesis, it is worth mentioning that the slope for females for the CNN's ACC scores is positive, even when accounting for uncertainty. The positive slope might have been more prominent if the CNN had time to be trained on all 125 training-validation sets. However, surprisingly in contrast to previous studies, we also found that for the CNN, the performance was better for males in general. We hypothesise that one of the reasons why we found this difference and Sies et al., 2022 did not, could be because they only processed dermoscopic images, which might have led to less difference between the sexes, since the images are better quality and the lesions are more visible than with smartphone images.

7.3 Limitations

We want to highlight that the results we obtained should be interpreted carefully, since our CNN and LR did not reach a state-of-the-art (SOTA) performance, and therefore we cannot guarantee that if the models reached a SOTA performance the results would be same. We speculate that the relatively low ACC could be due to several factors. First of all, the data that the models were trained on were all photos taken with a smartphone, rather than with a dermatoscope. Research show that dermatologists have a 65%-80% ACC rate in diagnosing melanoma, but with technical support, such as using dermoscopic images, the ACC of skin lesion diagnosing could be increased by a further 49% (Brinker et al., 2018). This could suggest that our models would improve their performance with dermoscopic images. It is also worth mentioning that even though melanoma is the most lethal form of skin cancer, it is also very rare. The distribution of the diagnoses in PAD-UFES-20 is small for melanoma. LR uses features based on medical guidelines used specifically to identify melanoma, and even though some of these characteristics between the different skin cancers are the same, it is worth mentioning that the LR uses methods developed for particularly melanoma (Corewell Health; DermLite).

Another point is that even though our data was split carefully to utilise as much data as possible, the amount of training data was relatively small, compared to SOTA models, giving us another disadvantage in the training of our model.

As for the LR, the handcrafted methods from the former first-year data science students, have a lot of features that rely on their own segmentations or masks of the lesions. The segmentations were used to calculate features such as asymmetry-, border- and colour-features. However, it is important to note that both we and the students lack expertise in dermatology. Therefore, the segmentation process was primarily guided by intuition rather than scientific reasoning.

Due to the lack of skin types ranging from 3-6 in regards to the Fitzpatrick scale, it was not possible to conduct this experimental setup, to investigate whether the LR and CNN would be robust to a dataset shift in regards to skin types with the PAD-UFES-20 dataset. We did not find it feasible to augment such underrepresented groups of skin types, due to the distribution being so skewed.

7.4 Ethical considerations

Like most of the skin lesion datasets out there, PAD-UFES-20 is also underrepresented in skin types, with the most of the skin types ranging from 1-3. People in the range of 4-6 are less likely to develop melanoma, so by including more people in the dataset that have a darker complexion, it could also combat the imbalance in datasets regarding cancerous skin lesions vs. non-cancerous skin lesions. Nonetheless, there is a lack of medical attention in regards to patients who are people of colour (POC) and skin cancers due to a lack of awareness of how skin cancer can behave on a darker complexion. So, even if there were more data subjects with a darker skin type, it would be difficult to conclude that they are less likely to have skin cancer, as healthcare professionals are lacking in diagnosing POC, meaning a skin lesion could be diagnosed as non-cancerous in the dataset, but it might be a cancerous skin lesion in actuality (Balch, 2022).

Deep learning today has a significant carbon footprint, and as of now we are heading towards an ecological collapse (Global Challenges Foundation). Therefore, even though our project's scale is rather small, we still have a responsibility of trying to minimize our projects' carbon footprint.

We established that we would use the feature sex instead of the feature gender, to get a binary classification problem, but in reality sex is not binary either. Calling sex binary erases intersex people, although there is research which shows that intersex does not violate sex being binary (Rehman, 2023). We will also acknowledge that by simply changing the feature from gender to sex, might cause problems for transgender individuals whose gender identity does not align with their biological sex. The data, with our change from gender to sex, might not correctly reflect the individuals' biological sex, which was our focus.

As mentioned early in the paper, multiple mistakes were found in the PAD-UFES-20 dataset, and we cannot guarantee that there are no more mistakes than the ones we have found. This means that we are working with a dataset, where we are not sure if the data is exactly as described by the authors, and therefore potentially compromising the basis of our research.

8 Conclusion

To conclude, this research investigates the robustness of machine learning models, specifically an LR model and a CNN based on ResNet-50, in the context of skin cancer detection. The PAD-UFES-20 dataset was used to evaluate the robustness of these models under varied ratios of sex compositions in the training data. Our findings indicate that both the LR and the CNN displayed robust performance across the different datasets, suggesting it is resilient to shifts in data related to sex. However, the overall ACC and AUROC scores for females and males achieved by the CNN showed, when conducting a Mann-Whitney U test that the CNN had consistently higher performance for both metrics for males than females. Therefore, even though both models showed robustness regarding the dataset shift, the CNN's higher performance on males suggests a need for further research into the field of dataset sex composition for skin cancer detection.

9 Acknowledgements

We would like to thank Petersen et al. for investigating bias in medical imaging, giving us inspiration to apply their experimental setup to our own research for skin cancer detection. We would also like to thank Pacheco and Krohling for their valuable dataset. Special thanks to our supervisor, Veronika Cheplygina, who helped provide the dataset, the data from the students, guidance and much more. Additionally, thanks to the first-year data science students from 2023 for their handcrafted methods, and the first-year data science

students from 2024 who provided segmentations for the lesions.

References

Ameri A. 2020. A Deep Learning Approach to Skin Cancer Detection in Dermoscopy Images. *Journal of Biomedical Physics & Engineering*, 10(6):801–806.

Bridget Balch. 2022. Why are so many Black patients dying of skin cancer?

Saulo Barreto. 2022. Data Augmentation | Baeldung on Computer Science.

Titus Josef Brinker, Achim Hekler, Jochen Sven Utikal, Niels Grabe, Dirk Schadendorf, Joachim Klode, Carola Berking, Theresa Steeb, Alexander H. Enk, and Christof von Kalle. 2018. Skin Cancer Classification Using Convolutional Neural Networks: Systematic Review. *Journal of Medical Internet Research*, 20(10):e11936. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

Cancer Council Victoria. Skin Cancer Guide: Types, Symptoms, and Prevention | Cancer Council Victoria - Cancer Council Victoria.

M. Emre Celebi, Hitoshi Iyatomi, William V. Stoecker, Randy H. Moss, Harold S. Rabinovitz, Giuseppe Argenziano, and H. Peter Soyer. 2008. Automatic Detection of Blue-White Veil and Related Structures in Dermoscopy Images. *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, 32(8):670–677.

Veronika Cheplygina. 2023. Submissions of students in First Year Project 2021-2023 at IT University of Copenhagen [Python code and PDF reports]. First Year Project, IT-University of Copenhagen.

Corewell Health. Skin Cancer | ABCDE Assessment for Melanoma | Beaumont Health.

Council Of Europe. Sex and gender - Gender Matters - www.coe.int.

DermLite. 7-Point Checklist.

Global Challenges Foundation. Ecological collapse.

Vignesh Gopalakrishnan. 2023. Image segmentation using otsu threshold selection method.

Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. 2021. Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1820–1828. IEEE. Place: Nashville, TN, USA.

Lisa N. Guo, Michelle S. Lee, Bina Kassamali, Carol Mita, and Vinod E. Nambudiri. 2022. Bias in, bias out: Underreporting and underrepresentation of diverse skin types in machine learning research for skin cancer detection-A scoping review. *Journal of the American Academy of Dermatology*, 87(1):157–159.

Frederik Hvilshøj. 2022. Introduction to Balanced and Imbalanced Datasets in Machine Learning.

Imbalanced Learn. SMOTE — Version 0.12.2.

Label Studio. Open Source Data Labeling.

Xiaoxiao Li, Ziteng Cui, Yifan Wu, Lin Gu, and Tatsuya Harada. 2021. Estimating and Improving Fairness with Adversarial Learning. Issue: arXiv:2103.04243 arXiv: 2103.04243 [cs].

Xiaoxuan Liu, Livia Faes, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, Joseph R Ledsam, Martin K Schmid, Konstantinos Balaskas, Eric J Topol, Lucas M Bachmann, Pearse A Keane, and Alastair K Denniston. 2019. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health*, 1(6):e271–e297.

Ali Madooei, Mark S. Drew, and Hossein Hajimirsadeghi. 2015. Learning to Detect Blue-white Structures in Dermoscopy Images with Weak Supervision. ArXiv:1506.09179 [cs].

MLflow. MLflow | MLflow.

Sarang Narkhede. 2021. Understanding AUC - ROC Curve.

Andre G. C. Pacheco and Renato A. Krohling. 2020. The impact of patient clinical information on automated skin cancer detection. *Computers in Biology and Medicine*, 116:103545.

Eike Petersen, Aasa Feragen, Maria Luise da Costa Zemsch, Anders Henriksen, Oskar Eiler Wiese Christensen, and Melanie Ganz. 2022. Feature robustness and sex differences in medical imaging: a case study in MRI-based Alzheimer's disease detection. Issue: arXiv:2204.01737 arXiv: 2204.01737 [cs, eess].

Tri-Cong Pham, Chi-Mai Luong, Van-Dung Hoang, and Antoine Doucet. 2021. AI outperformed every dermatologist in dermoscopic melanoma diagnosis, using an optimized deep-CNN architecture with custom mini-batch logic and loss function. *Scientific Reports*, 11(1):17485. Publisher: Nature Publishing Group.

PyTorch. Illustration of transforms — Torchvision main documentation.

Rashad Rehman. 2023. "Intersex" Does not Violate the Sex Binary. *The Linacre Quarterly*, 90(2):145–154.

resnet50. resnet50 — Torchvision main documentation.

RITHP. 2023. Logistic Regression and regularization: Avoiding overfitting and improving generalization.

scikit-learn. scikit-learn: machine learning in Python — scikit-learn 1.4.2 documentation.

SciPy. scipy.stats.mannwhitneyu — SciPy v1.13.0 Manual.

Katharina Sies, Julia K. Winkler, Christine Fink, Felicitas Bardehle, Ferdinand Toberer, Timo Buhl, Alexander Enk, Andreas Blum, Wilhelm Stolz, Albert Rosenberger, and Holger A. Haenssle. 2022. Does sex matter? Analysis of sex-related differences in the diagnostic performance of a market-approved convolutional neural network for skin cancer detection. *European Journal of Cancer*, 164:88–94.

Philipp Tschandl. 2023. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Version Number: 4.

Appendix

A Fitzpatrick scale distribution

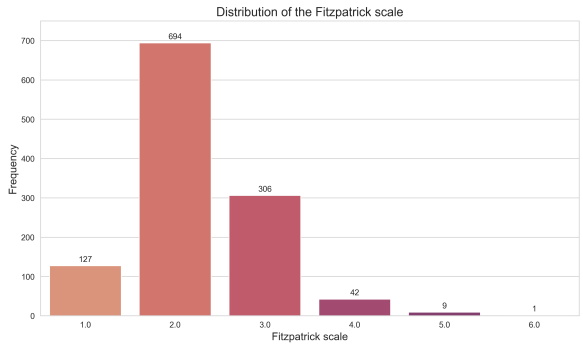

Figure 5: Distribution of the Fitzpatrick scale

B Exploring the PAD-UFES-20 dataset

1. When exploring the data, we encountered 45 instances of two patients sharing the same `lesion_id`.

2. Multiple patients have different `lesion_ids` but the images of those lesions are the same, just more zoomed in. Just for context, `img_id` is composed of: `PAT_id, lesion_id, random_number.png`

(a) Image for `img_id` `PAT_417_3537_45.png`

(b) Image for `img_id` `PAT_417_828_723.png`

Figure 6: Caption for both images

   (a) The only difference between those cases was the feature `changed` might say **`True`** instead of **`False`**

3. There are also cases where the `lesion_ids` are different for the same patient, but the images are fully identical. A case example are for the `img_ids`: `PAT_528_993_589.png` and `PAT_528_3072_615.png`

Figure 7: Image representing the aforementioned point