Degree project in computer engineering

First cycle, 15 credits

Determining Important Features for Melanoma Classification Through Feature Selection

VILMER JONSSON

TOR STRIMBOLD

Determining Important Features for Melanoma Classification Through Feature Selection

VILMER JONSSON
TOR STRIMBOLD

Abstract

Skin cancer is a common disease in both Sweden and the United States. Although common, the survival rate of melanoma patients is high if the diagnosis is made at an early stage. Computer aided diagnostics has been shown to have potential in accurately diagnosing the disease utilizing machine learning. Thus, machine learning algorithms can be used to effectively classify a skin lesions as either benign or malignant. These algorithms can be made more accurate and efficient by applying feature selection since it decreases the dimensionality of the feature space. The aim of this study is to apply feature selection on four different classifiers to compare morphological and SIFT features in order to determine which features are important for classifying melanoma.

The results show that morphological features in general increased the accuracy more than the SIFT features, although it varied between different classifiers. Furthermore, forward selection was more effective than backward selection in terms of accuracy for two of the classifiers. Lastly, certain features were more effective than others, the most effective feature was "Solidity".

Sammanfattning

Hudcancer är en vanligt förekommande sjukdom i både Sverige och USA. Dock är sannolikheten att en patient kan botas från sjukdomen hög om diagnosen sker i ett tidigt stadium. Datordriven diagnostisering har visat sig kunna diagnostisera sjukdomen på ett effektivt och med hög säkerhet genom att tillämpa maskininlärning. På så vis kan maskininlärningsalgoritmer användas för att klassificera hudutslag som godartade eller inte. Dessa algoritmer kan effektiviseras genom att utföra attributurvalsmetoder då det minskar antalet dimensioner som behöver beräknas. Syftet med denna studie är att undersöka vilka attribut som är viktiga för klassificeringen. Detta gjordes genom att tillämpa attributurvalsmetoderna Sekventiell Framåt- och Bakåtselektion på fyra olika maskininlärningsalgoritmer med indata i form av morfologiska och SIFT-attribut.

Resultaten visar att morfologiska attribut generellt föredrogs i större utsträckning än SIFT-attribut, detta varierade dock mellan olika klassificeringsmodeller. Vidare var Framåtselektion mer effektiv än bakåtselektion sett till träffsäkerhet för två av klassificeringsmodellerna. Slutligen var vissa attribut mer effektiva än andra, det mest effektiva attributet var Soliditet".

Contents

Chapter 1

Introduction

The frequency of malignant melanoma has in the last decade been on the rise in Sweden with approximately 60 000 people being diagnosed with the disease yearly. [1]. A similar trend has been observed in the United States where the number of people diagnosed with the disease has doubled during the period 1982-2011 [2].

Skin cancer, although common, has a 5-year survival rate of 99% if discovered at an early stage [3]. This fact makes the diagnosis of malignant melanoma an essential part of the treatment of the disease. In the light of diagnosis, machine learning has become a subject of interest in many medical fields, dermatology included. Multiple studies have been conducted to evaluate the potential of the algorithms' accuracy and applicability through so-called computer aided diagnostics (CAD). Concerning malignant melanoma, some studies have reached an accuracy of up to 90% which competes with the experts in the field.

The accuracy of machine learning algorithms is however heavily reliant on the feature extraction methods used, where relevant features in an image are extracted and then used in the classification of the skin lesion. An effective feature extraction in turn relies on determining which features to use. Which is the subject of feature selection where different features are compared in terms of their effect on the diagnostic accuracy. It has been shown that effective feature selection make classifiers more effective, accurate and cost-effective. [4]

1.1   Research Question

In our thesis we studied the effect of two different feature selection methods upon 169 features and using 4 different machine learning models. We aim to answer:

1. Which features are the most important in terms of classification accuracy?

2. Is there a difference between different models concerning which features are most effective?

3. Is there a difference between selection methods in terms of performance enhancement?

1.2   Approach

To study the proposed research questions two different selection methods were used, namely Forward and Backwards Selection. To further widen the base for comparison four different machine learning models will be used: K-Nearest Neighbors, a one-layered Neural Network, Random Forest and Support Vector Machine. These models are widely used in the research of automatic skin lesion diagnosis which makes them relevant to study.

Two types of features were examined, morphological and SIFT features. The morphological features consisted primarily of so-called ABCD features but also of features calculated using different morphological formulas. The ABCD features are based on the clinical practice where a skin lesion is classified based on its asymmetry, border irregularity, color and diameter. The ABCD method is widely used when diagnosing skin lesions using machine learning due to its effectiveness and simplicity to implement. [5]

1.3   Scope

The study is limited by the number of features, selection methods and models being used.

Firstly the models are limited by the fact that they consist of only shallow machine learning models. An extension of the study could include a deep learning model such as a convoluted neural network.

Secondly, the study is limited by the feature selection methods being used where only feature selection by wrapper methods are used through the sequential backward and forward selection. Other methods such as filter methods and embedded methods could be used to find synergies between features and widen the scope of the study.

Thirdly, the features were only of two types. Including for example deep features would give further insight into how the different types of features affect the accuracy of the classifiers.

Chapter 2

Background

2.1   Skin Cancer

During the last decades the frequency of malignant melanoma in Sweden has been on the rise and the trend does not seem to be going down. Yearly, approximately 60 000 persons are diagnosed with skin cancer and around 500 die from the disease. [1]

Likewise in the United States, skin cancer is the most common type of cancer to the degree that every fifth American will develop the disease at some time during their lives. The rates of melanoma have on a general trend increased, nearly doubling between 1982-2011. On some levels the rates have decreased, as in for people that are under 30. [2]

Malignant melanoma is a type of skin cancer that is the result of melanocyte cells in the skin starting to grow uncontrollably. Although melanoma is a less common type of skin cancer, it is at the same time more prone to spreading to other parts of the body therefore making it more dangerous. [2]

Melanoma skin cancer has five different stages, stage 0-4, where in stage 0 the cancer is contained within the top layer of the skin while in the final stage the cancer has spread to other parts of the body [6]. According to UK researchers the survival rate of patients diagnosed with melanoma differs greatly depending on when during the four different stages the diagnosis is made [7]. If the cancer is detected in the first stage the survival rate is almost 100%. However, if the diagnosis is done in the last stage the survival rate drops to 30%, although the statistic does not take into account the age of the patients. [7]

With the development of artificial intelligence (AI) and machine learning (ML), great strides within dermatology in areas such as melanoma detection

and classification has been made. In one study, a machine learning algorithm was compared against experts, where the ML system outperformed the experts. [8]

2.2   Computer Aided Diagnostics

Machine learning is a field of computer science focusing on automated and intelligent data analysis. An ML algorithm is fed problems and answers whereupon the system tries to predict the answer. Based on the response the algorithm then makes adjustments to improve the predictions. [9]

In the case of skin cancer detection, an ML algorithm would receive images with a ground truth for each image, meaning either the skin lesion is malignant or benign. The algorithm would then try to classify the lesion. The correct answer would then be revealed whereupon the algorithm makes corrections based on whether it was successful in its diagnosis.

This process consists of four main steps: 1) pre-processing, 2) segmentation, 3) feature extraction and 4) classification. All the named steps are described below.

2.2.1   Pre-processing

The main goal with pre-processing is to improve the quality of the images. Improvement can be made in many ways, but one common method is artifact removal [10]. In general, artifact removal entails removing misleading objects in the image such as hair pixels and air bubbles which can lead to inaccurate segmentation and feature extraction, and thus harming the quality of the classification. [11]

2.2.2   Segmentation

In the segmentation phase the pixels that constitute the lesion are identified. The result of the segmentation is a binary image where white pixels represent the lesion and black pixels represent the surrounding skin. Similar to pre-processing, segmentation is vital for the classification and feature extraction since parameters such as the lesion's border cannot be correctly evaluated if the segmentation is poorly executed. At the same time, the segmentation of skin lesions is one of the most challenging steps due mainly to four reasons. Firstly, most lesions do not have a stark contrast compared to the surrounding skin. Secondly, the lesion's border may be irregular. Thirdly, the color variegation

within the lesion can be vast where multiple colors are present at the same time. Finally, there may be artifacts such as hairs or lens flare present within the image which hinders the segmentation, this is however meant to be removed during the pre-processing. [11]

2.2.3 Feature Extraction

Clinical classification is done based on the characteristics of the lesion. Likewise, for an algorithm to be able to classify a lesion, the features needs to be quantified, which is the goal of feature extraction. Which features to consider relevant depends on the base of classification, the same holds for the way to extract the features. The two different types of features used in this study are Scale-invariant Transform Features (SIFT) and morphological features.

Morphological features

Morphological features describe the physical attributes of the lesion. A common way to describe a lesion using its morphological features is the ABCD-rule which is a mimic of the clinical practice related to diagnosing melanoma. The method is based on four different aspects of a lesion: the asymmetry, border, color and diameter of the lesion. The higher the asymmetry of the lesion, in terms of either shape or color, the more likely it is to be malignant. The border aspect refers to whether the border of the lesion is irregular. One way to compute this is by dividing the lesion into eight slices and for each slice, if the border is irregular a point is awarded. Regarding the color, the lesion is scanned for the occurrence of six colors (white, red, light brown, dark brown, blue-grey and black). For each color a point is awarded [12]. The last aspect is often evaluated using measurements in the images. However, our dataset do not have any measure present in the images, therefore the diameter is not explicitly evaluated. However, other morphological features used in this study can be seen as a substitute for the diameter. For example the "extent" of the lesion, meaning the quotient between the area of the lesion and the area of the image [13]. [14]

Scale-invariant feature transform (SIFT)

Scale-invariant feature transform (SIFT) is a popular method to extract key points out of an image. These key points are detected by an algorithm and given descriptors based on the pixels near the key point. These descriptors can then be used in a classifier. [15] Although SIFT descriptors have a wide scope

of application, the research where the features are used to classify melanoma is quite limited.

2.3   Classification

The classification of the lesions is based on the features extracted and can be done through several machine learning algorithms. The ones used in this study are listed below.

2.3.1   K-Nearest Neighbors (KNN)

K-nearest-neighbors is used for classifying data points. The ground principle is to find the nearest neighboring data points from the training set for all the test data points. This is based on the idea that similar data points can be found close to one another. When classifying, the model regards the k the nearest data points and classifies the test data point based on which class the majority of these neighbors have. [16]

The algorithm is simple to implement and does not have many hyperparameters to tune. However, one drawback is that it suffers from the curse of dimensionality, meaning that the model does not perform well with many features. [16]

2.3.2   Neural Network (NN)

Neural networks are a type of machine learning algorithm that is inspired by the human brain. The network consists of several layers of nodes, where each node is connected to all nodes in the previous layer. The first layer is the input layer, where one node corresponds to one feature from the input data. The last layer is the output layer, where the final prediction is made. The layers in between are called hidden layers. All nodes in one layer is connected to all nodes in the next layer. Each connection between two nodes has a weight, which is adjusted during the training of the model. The weights are adjusted in order to minimize the loss function, which is a measure of how well the model performs. [17]

When classifying, the input data is fed into the input layer. The data is then passed through the network, where each node in the hidden layers calculates a weighted sum of the inputs and applies an activation function to the result. The output of the neurons in the last hidden layer is then fed into the output layer, where the final prediction is made. [17]

2.3.3   Random Forest (RF)

Random forest is a tree-based method based on the idea of bagging. Bagging is a method where several models are trained on different subsets of the training data. This is done in order to reduce the correlation between different trees. The final prediction is made by a majority vote between the individual models. [18]

2.3.4   Support Vector Machine (SVM)

Support vector machines are common and effective ML algorithms first introduced in the 1990s. SVM is an extension of the support vector classifier, which in turn is an extension of the maximal margin classifier.

The maximal margin classifier aims to separate data into two classes using a hyperplane. The support vector classifier builds on the same principle, but is more adjusted to the case when the data is not easy to separate. Lastly, the SVM enlarges the feature space of the data in order to be able to separate data in a non-linear fashion. [19]

2.4   Feature selection

Feature selection is a process which aims to reduce the number of features, or detect the most important features, passed to the classifier. Reducing the number of features results in a shorter runtime and some features may be irrelevant and thus confusing for the model which impedes the performance. [20]

In this study the greedy methods Sequential Backward Selection (SBS) and Sequential Forward Selection (SFS) were used.

In both methods the goal is to determine the subset of features which leads to the best performance. In SBS this subset includes all the features in the feature space. The process then consists of removing each feature individually which creates several subsets. These subsets are then evaluated and the one with the best performance is considered the new subset. This procedure is continued until only one feature is left or until some criterion is met. In SFS the opposite process is performed. At first the subset is empty and each feature in the original feature set is evaluated. The feature with the best performance is then added to the subset, then each of the remaining features are evaluated together with the added feature. This is repeated until all features are added or until some criterion is met. [20]

In this study, the subset of features that lead to the highest accuracy is considered the *optimal feature set*.

2.5 Related Works

The automatic classification of skin lesions as being either malignant or benign has been a subject of great interest and there is a plethora of publications investigating the manner. Hence, there are several studies focused around classifying lesions, using the feature set and ML models which this thesis aims to investigate. However, there is not such an abundant pool of studies which focus on feature selection methods and exploration of which types of features are the most important.

[21] studied the minimal number of necessary features for effective and accurate skin lesion classification using a KNN-model. The features that were studies were based on the geometric (A), color (C) and boundary (B) aspects of the ABCD rule of melanoma.([21]) A feature selection method which implemented the Sequential Backward Selection (SBS) was used. When all features (15 in total) were used, the average accuracy was around 78%. The highest accuracy of 91% was obtained when there were only 6, 8 and 9 features. Hence, [21] showed that an effective feature selection method both improves the accuracy of the classifier and the computation and storage cost. Though they did not highlight which of these features were abundant and which were important to include. Additionally, they only limited the study to one type of ML model, namely the KNN.

[21] studied the effect of feature selection using the NCA method with an M-SVM on three different data sets. Like [21] their method showed an increase in accuracy and runtime. However, as well as [21], they did not investigate which features were the ones leading to the increased accuracy. Furthermore, only one classifier was used with the M-SVM and therefore no results were generated using some other type of classifier.

Chapter 3

Methods

3.1 Machine learning methods

3.1.1 Preprocessing

For removing artifacts such as hairs, the DullRazor® software was used. The software is created by Tim Lee, Vincent Ng, Richard Gallagher, Andrew Cold-man and David Mclean and is used frequently for preprocessing of skin lesion images. The method removes the hairs in two steps: 1) using generalized grayscale morphological closing operations, it identifies the dark hair loca-tions, 2) verifies the shape of the hair pixels as long and thin, thereafter it replaces the verified pixels by bilinear interpolation and lastly 3) smooths the replaced hair pixels. [22]

3.1.2 Segmentation

After the preprocessing step, the segmentation of the image is calculated in order to be used to extract the features. The aim of this study was not to de-velop a new segmentation method, but rather to compare the effect of different feature selection methods on the accuracy of the classification. Therefore, the segmentation was done using open-source software available on GitHub [23].

3.1.3 Feature extraction

As mentioned in the background, the two feature extraction methods used were ABCD and SIFT. In total 169 features were used where the SIFT and morpho-logical features consisted of 150 and 19 features respectively.

Feature type	Indexes in feature set	Number of features
SIFT	0 - 149	150
Morphological	150 - 168	19

Morphological

The morphological features could be divided into four different categories: morphological formulas, and three of the aspects of ABCD, namely the asymmetry, border and color aspects. The calculations for each feature is presented in table 3.1

The indexes 150 - 156 are morphological formulas that describe the lesion in different ways.

Index 157 through 159 were based on the research of [24]. The first two (index 157 and 158) describe the asymmetry of the lesion where a high value indicates high asymmetry. In table 3.1, A_x, A_y and A denotes the asymmetry along the x- and y-axis and the area of the lesion respectively. The third, index 159, describes the border irregularity where P denotes the perimeter of the lesion, and d and D denotes the minor and major diameter of the lesion.

The remaining features are based on the research of [25] and describe the colour attributes of the lesion. These are based on the RGB-values of the pixels in the lesion and draws either comparisons to the lesion itself or to the surrounding skin. The indices 160-162 describe the relative R, G and B ratio. In table 3.1, R_L, G_L and B_L describe the average value of red, green and blue in the surrounding skin's pixels and R_L, G_L and B_L denotes the average red, green and blue values of the lesion's pixels. Hence, these features measure the contrast between the lesion and the surrounding skin terms of the three colors. The indices 163-165 measure the relative R, G and B difference between the lesion and the surrounding skin, the same notations are used. Lastly, features 165-168 measure the normalized relative R, G and B differences between the surrounding skin and the lesion. [25]

Table 3.1: Description of the morphological features

Index in feature set	Description
150	Area of the lesion/Area of bounding rectangle
151	Area of lesion /Area of the lesion's convex hull
152	d/D
153	$4A/(\pi d^2)$
154	$\pi d/P$
155	$4\pi A/P^2$
156	$P/(\pi D)$
157	$min(A_x, A_y)/A$
158	$(A_x + A_y)/A$
159	$P * (1/d - 1/D)$
160	$F_4 = R_L/R_S$
161	$F_5 = G_L/G_S$
162	$F_6 = B_L/R_S$
163	$F_{10} = R_L - R_S$
164	$F_{11} = G_L - G_S$
165	$F_{12} = B_L - B_S$
166	$F_{13} = F_{10}/(F_{10} + F_{11} + F_{12})$
167	$F_{14} = F_{11}/(F_{10} + F_{11} + F_{12})$
168	$F_{15} = F_{12}/(F_{10} + F_{11} + F_{12})$

The implementation is originally done by [23], the documentation is available in the GitHub repository.

SIFT

The SIFT features are extracted using the OpenCV class SIFT which implements the algorithm by David G. Lowe. [26] [15].

The implementation is based on a GitHub repository which uses the "Bag of Words" technique to transform the key points descriptors into 150 different features [27]. The Bag of Words technique aims to capture the presence or absence of specific patterns or visual cues within the lesion images that are indicative of melanoma [28].

3.1.4 Classification

All classifiers used in this study are from the Scikit-sklearn library which is a free-to-use library for machine learning in Python [29].

With the aim to improve the accuracy of the models, hyperparameters were tuned in order to better fit the data. It is however important to note that a complete parameter sweep was not performed.

The `n_neighbors` hyperparameter were modified for the KNN classifier and was set to 8.

For the NN classifier two hyperparameters were modified from the default value. `hidden_layer_sizes` was set to $[8]$, which means that the classifier used one hidden layer with eight nodes. `max_iter` was set to 100 in order to speed up the testing time.

Regarding the RF classifier, the only modified hyperparameter was `n_estimators` which was set to 5. The standard value is 100 but because of the testing time scaling rapidly it was lowered.

For the SVM only one hyperparameter were modified, namely `C`, which was set to 100 instead of the usual 1.

3.2 Dataset

The dataset used in this study is a variant of the one produced by the ISIC Archive. The difference with the original one is that the one used only has two classes: Malignant and Benign. The dataset contains around 3400 images, each being 224x224 pixels. The dataset is balanced, which has positive effects on the classifier's accuracy. The larger, but unbalanced, HAM10000 dataset was also considered. However, during initial testing it was discovered that the balanced dataset resulted in a better accuracy. It was also easier to work with the balanced dataset, as it already only had two classes while HAM10000 has seven.

The dataset was downloaded and structured in order to have every image in the same folder. In order to store the ground truth information indicating whether a given image is malignant or benign, a one-dimensional array was created storing boolean values.

3.3   Metrics

3.3.1   Accuracy

Accuracy is a metric which is widely used and simple to calculate. It is calculated by dividing the total number of correctly classified images with the total number of images. It therefore measures the probability that the model will classify a given random image correctly. [30]

Accuracy was the metric used to determine the performance of the classifiers and the effect of the feature selection methods.

3.4   Feature selection

All feature selection methods used in the study originates from the `mlxtend` library `feature_selection` [31].

For SFS, all features were added one at a time, starting with an empty feature set. For SBS, all features were removed one at a time, starting with all features. In each iteration, the accuracy and which features constituted the feature set were recorded. The subset of features with the highest accuracy over all iterations is considered the *optimal feature set* and is also recorded.

This process was repeated five times for every selection method and classifier. Hence, 40 tests were conducted in total.

3.5   Evaluation

The method for evaluating the three different research questions are stated below.

3.5.1   Which features are the most important in terms of classification accuracy?

Based on the findings from the previous research question, the results were summarized to give insight on a more general level where the occurrences of each feature in the optimal feature set was divided by the total number of tests performed. This gives an insight into which features are most likely to constitute the optimal feature regardless of the classifier. Furthermore, the two different types of features can be compared which gives insight into which type of features is more important.

3.5.2   Is there a difference between different models concerning which features are most effective?

Since in every iteration of the selection methods, the included features were recorded, the features included in the feature set generating the highest accuracy was determined. Onwards, this feature set will be referred to as the optimal feature set.

Based on this statistic, the probability that a feature would be included in the optimal feature set, was calculated by dividing the number of times the feature was included in the optimal feature set with the number of tests performed. The features with the highest probability can therefore be seen as the most important.

3.5.3   Is there a difference between selection methods in terms of performance enhancement?

To determine the performance of a selection method two aspects needs to be considered. First and foremost, the eventual gain in accuracy a method gives a classifier. This can be evaluated by calculating the average accuracy in every iteration for a selection method and classifier. For a selection method and a classifier this would entail calculating the average accuracy when the feature set consisted of one feature, two features and so on. For the same classifier, the same process is done for the other selection method. These two lists of averages can then be plotted and the difference in performance can be observed. From this plot the highest average accuracy can be determined.

The second aspect is the average number of features used to obtain the highest accuracy. This can be calculated by first calculating the highest accuracy in every test and the number of features that were in the feature set. This information is then used to calculate the average number of features used to achieve the highest accuracy.

Chapter 4

Results

In this chapter, the results obtained are presented in the order of the research questions. Note that the full tables can be found in the appendix.

4.1   Determining the most important features

In this section the overall results are presented which describes a ranking of all the features, as well as the difference between SIFT and morphological features. The results concerning the overall difference between features are presented below in figure 4.1.
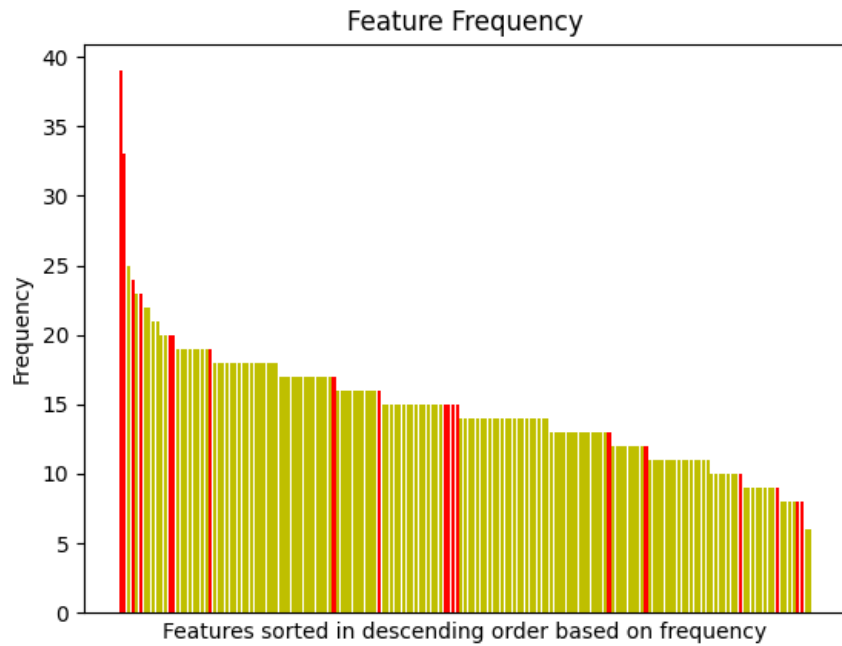
Figure 4.1: Prominence of each feature in all 40 tests sorted in descending order. Morphological features are colored red and SIFT features are colored green.

From the figure it is observable that two features had a higher frequency in the optimal set compared to the other features. Both of these features were morphological features. The most prominent feature was 'Solidity', which was calculated as the area of the lesion divided by the area of the lesion's convex hull. In simpler terms, the measurement gives an indication of how compact the lesion is. The second feature describes the contrast between the lesion and the surrounding skin in regard to the color red. It was calculated as the average red value of the pixels inside the lesion, divided by the average red value of the pixels outside the lesion. Thus, a value above '1' indicates that the lesion contains more of the color, while a lesser value indicates that the surrounding skin is redder than the lesion.

Concerning the difference between morphological and SIFT features table 4.1 show the results.

From table 4.1 it is observable that the SIFT features have more occurrences overall than the morphological features. However, if the distribution between the two types is considered, the morphological features have a higher

Table 4.1: Average number of occurrences in optimal set for Morphological and SIFT features.

Feature Type	Feature count	Occurrences	Average Occurrences
ABCD	19	331	15.8
SIFT	150	2157	14.6

occurrence rate per feature than the SIFT features, which can imply that the morphological features are more likely to be included in the optimal set.

4.2   Determining important features for each classifier

In this section results concerning features' importance for different classifiers are presented. To begin with the difference between the importance of SIFT and morphological features are presented in table 4.2.

Table 4.2: Average number of occurrences in optimal set for Morphological and SIFT features grouped by classifier.

Classifier	Morphological Occurrences	Occurrences per Morphological Feature	SIFT Occurrences	Occurrences per SIFT Feature
KNN	32	1.7	232	1.5
NN	118	6.2	792	5.3
RF	95	5.0	933	6.2
SVM	86	4.5	200	1.3

The results show no particular difference between the two types of features for the KNN, NN and RF classifiers. For the SVM, however, the morphological features were more common in the feature set.

In the following subsections the results for every classifier is presented through both a figure showing the distribution of feature occurrences and a table giving more details into which features were the most preferred.

4.2.1   K-nearest neighbors

Figure 4.2: Prominence of each feature in all 10 tests using KNN sorted in descending order. Morphological features are colored red and SIFT features are colored green.

Figure 4.2 shows that two features had a high frequency compared the others, and amongst these two there was a gap between the most and second most prominent.

Table 4.3 presents the features which had at least five occurrences in the optimal set. Note that in total ten tests were run so five occurrences amount to a frequency of 0.5.

Table 4.3: The features with at least 5 occurrences in the 10 tests.

Feature	Frequency	Frequency using SFS	Frequency using SBS
151	0.9	0.8	1.0
160	0.6	0.8	0.4
62	0.5	1.0	0.0
112	0.5	1.0	0.0
153	0.5	1.0	0.0

The results show that the feature 'Solidity' (index 151) was prominent in the KNN classifier, which is as previously mentioned a morphological feature. Likewise, the second most prominent feature was '161' which was the second most prominent feature in all tests. A final observation is that the frequency differs greatly depending on the selection method.
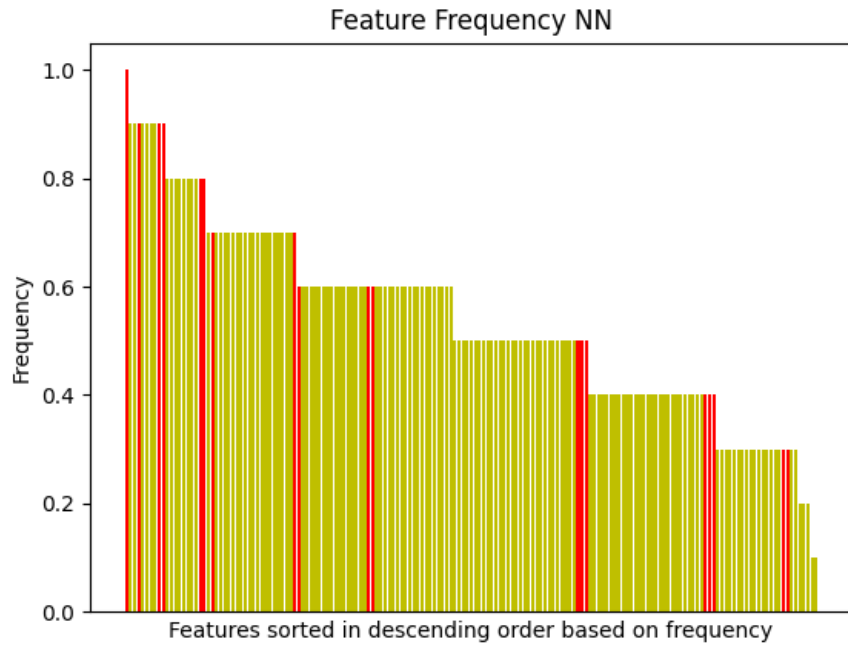
4.2.2   Neural Network

Figure 4.3: Prominence of each feature in all 10 tests using NN sorted in descending order. Morphological features are colored red and SIFT features are colored green.

From figure 4.3 it is observable that the classifier had a quite even distribution overall where almost all features were present in the optimal set at least once. Furthermore, the distribution between the morphological and SIFT features was slightly in the favor of the first-mentioned.

Table 4.4: The features with at least 9 occurrences in the 10 tests.

Feature	Frequency	Frequency using SFS	Frequency using SBS
151	1.0	1.0	1.0
80	0.9	1.0	0.8
100	0.9	0.8	1.0
160	0.9	0.8	1.0
0	0.9	1.0	0.8
62	0.9	1.0	0.8
96	0.9	1.0	0.8
109	0.9	1.0	0.8
157	0.9	1.0	0.8
158	0.9	1.0	0.8

The results show that the NN classifier utilized several features where 10 features had a frequency of 0.9 or more. Additionally, the top features were preferred by both selection methods. Lastly, the only feature that was included in the optimal set in every test was the 'Solidity' feature.
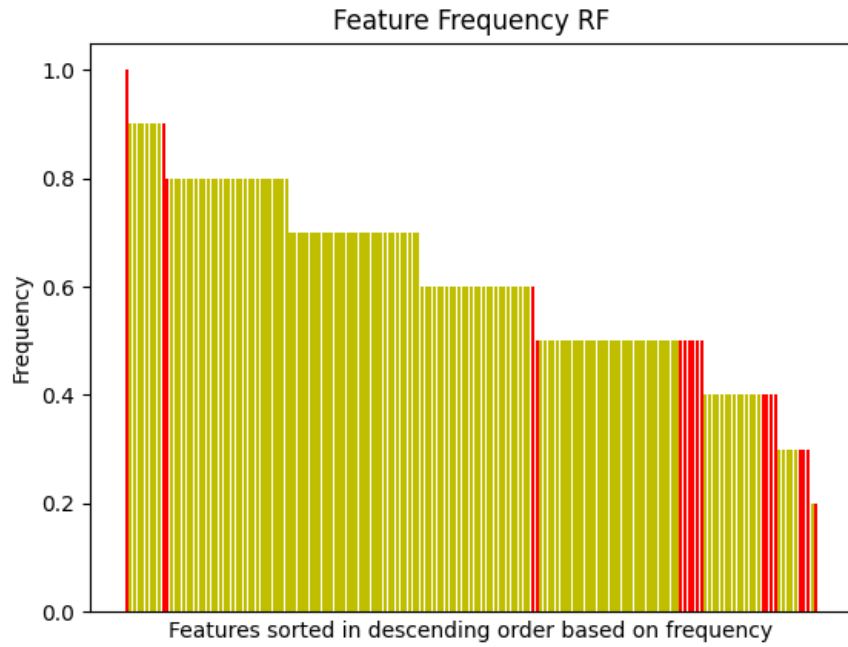
4.2.3 Random Forest

Figure 4.4: Prominence of each feature in all 10 tests using RF sorted in descending order. Morphological features are colored red and SIFT features are colored green.

The RF classifier preferred the SIFT features which is presented in table 4.2. This is also observable in figure 4.4 where most morphological features have fewer occurrences compared to the SIFT features. Lastly, there was one feature that had a comparatively high frequency. The upper part of the distribution is presented in table 4.5.

Table 4.5: The features with at least 9 occurrences in the 10 tests.

Feature	Frequency	Frequency using SFS	Frequency using SBS
151	1.0	1.0	1.0
80	0.9	1.0	0.8
100	0.9	1.0	0.8
23	0.9	1.0	0.8
24	0.9	1.0	0.8
72	0.9	0.8	1.0
73	0.9	0.8	1.0
94	0.9	1.0	0.8
111	0.9	0.8	1.0
166	0.9	1.0	0.8

Random forest, similar to Neural Network, utilized several features to a high degree. The most prominent feature was 151, which is the morphological feature 'Solidity'. Apart from this feature, the only other morphological feature with a frequency of 0.9 was feature 166 which, like feature 160, describes a difference between the lesion and the surrounding skin in terms of the color red.

4.2.4 Support Vector Machine

Figure 4.5: Prominence of each feature in all 10 tests using SVM sorted in descending order. Morphological features are colored red and SIFT features are colored green.

The SVM classifier preferred the morphological features which was shown in table 4.2 which is observable in figure 4.5 where almost all the morphological features had more occurrences than the SIFT features.

Table 4.6: The features with at least 5 occurrences in the 10 tests.

Feature	Frequency	Frequency using SFS	Frequency using SBS
151	1.0	1.0	1.0
160	1.0	1.0	1.0
161	1.0	1.0	1.0
162	1.0	1.0	1.0
145	0.7	0.6	0.8
165	0.6	0.8	0.4
163	0.6	1.0	0.2
167	0.6	1.0	0.2
164	0.5	0.8	0.2
166	0.5	1.0	0.0

As mentioned earlier the SVM preferred morphological features. From table 4.2.4 it is observable that out of the features that had a frequency above 0.5, only one was a SIFT feature. Consequently, all except number 151 were color features from the ABCD method. Furthermore, the occurrences of features decrease quite rapidly where there are only 10 features with a frequency of 0.5 or above. Lastly, the frequency of the features did depend on the feature selection method to a certain degree.

4.3   Difference between selection methods

Table 4.7: The average number of features in the optimal set for the different classifiers and selection methods.

Classifier	Selection Method	Average Number of Features	Average Accuracy
KNN	SFS	47.4	0.76
KNN	SBS	5.4	0.71
NN	SFS	100.0	0.73
NN	SBS	82.0	0.72
RF	SFS	101.4	0.73
RF	SBS	104.2	0.74
SVM	SFS	45.8	0.81
SVM	SBS	11.4	0.77

The results show that the SFS method used more features on average compared to the SBS method in all cases except for the Random Forest classifier. The biggest relative difference was in the case of the KNN classifier where the SFS method used more than eight times more features than the SBS method. However, no test used all the available features to achieve the highest accuracy. The highest average accuracy was achieved by the combination of SVM and SFS.

4.3.1   Difference in performance between selection methods

Following are the results related to the difference in performance between the two methods for every classifier.
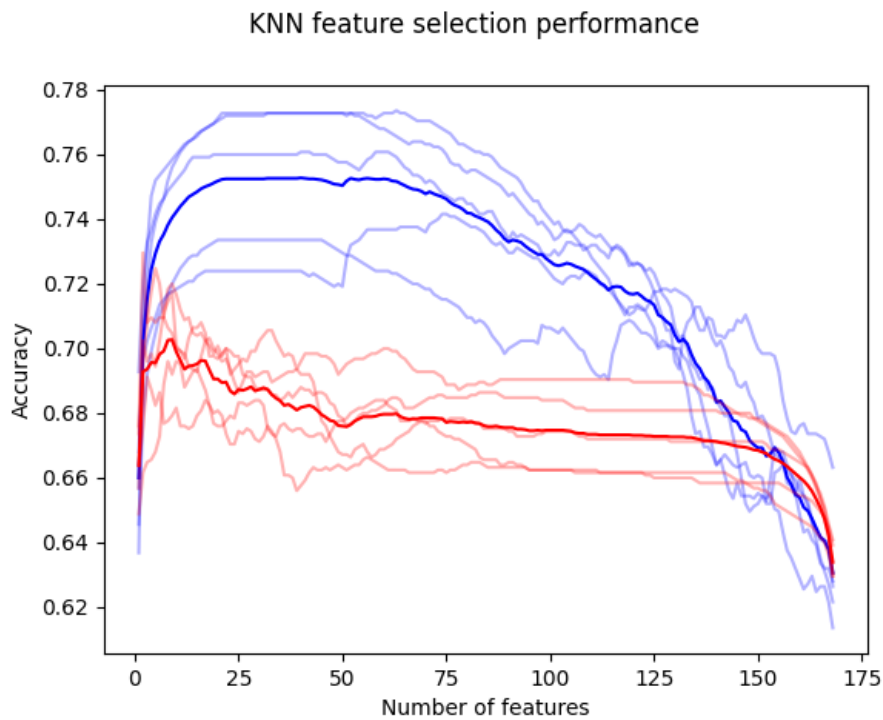
K-Nearest Neighbors

Figure 4.6: Difference in performance for the KNN classifier, where each individual test is plotted with the average accuracy in bold. Tests using the SFS method are in blue while the SBS are in red.

The results show that the SBS utilized less than half of the number of features compared to the SFS, where the former used around 35 features compared to the 65 features for the SFS. However, the SFS method achieved a higher accuracy than the SBS method. Another observation is a forming of plateaus where the gain in accuracy is marginal when adding features. Lastly, there was a variance in accuracy between the individual tests.
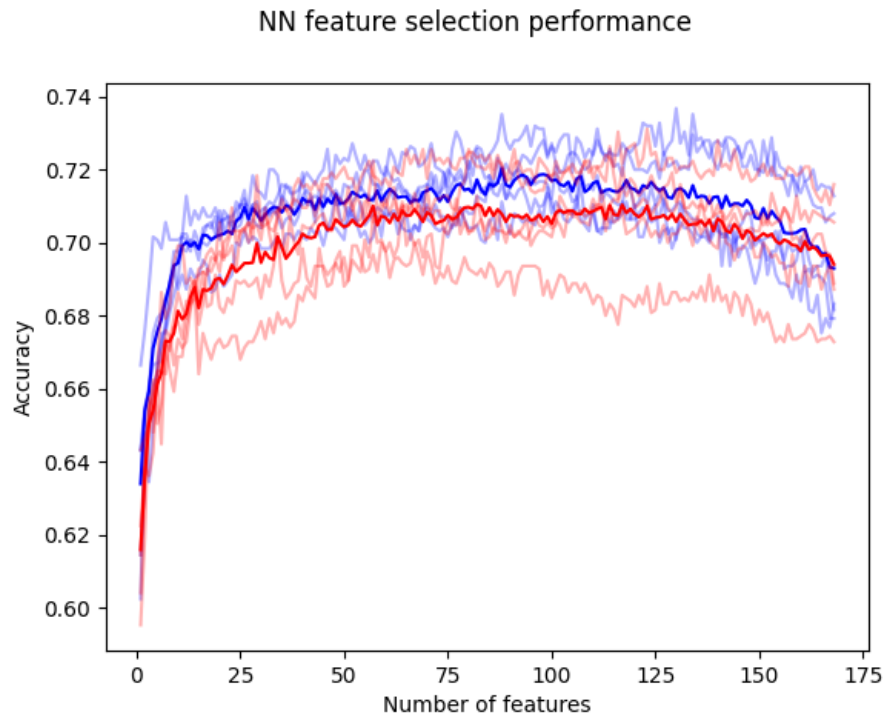
Neural Network

Figure 4.7: Difference in performance for the NN classifier, where each individual test is plotted with the average accuracy in bold. Tests using the SFS method are in blue while the SBS are in red.

Figure 4.7 shows that the two methods had similar results for the neural network, both in terms of performance and number of features used. None of the methods resulted in any major gains in terms of performance compared to using all features. Like in the case of the KNN there is some variance between the individual tests.
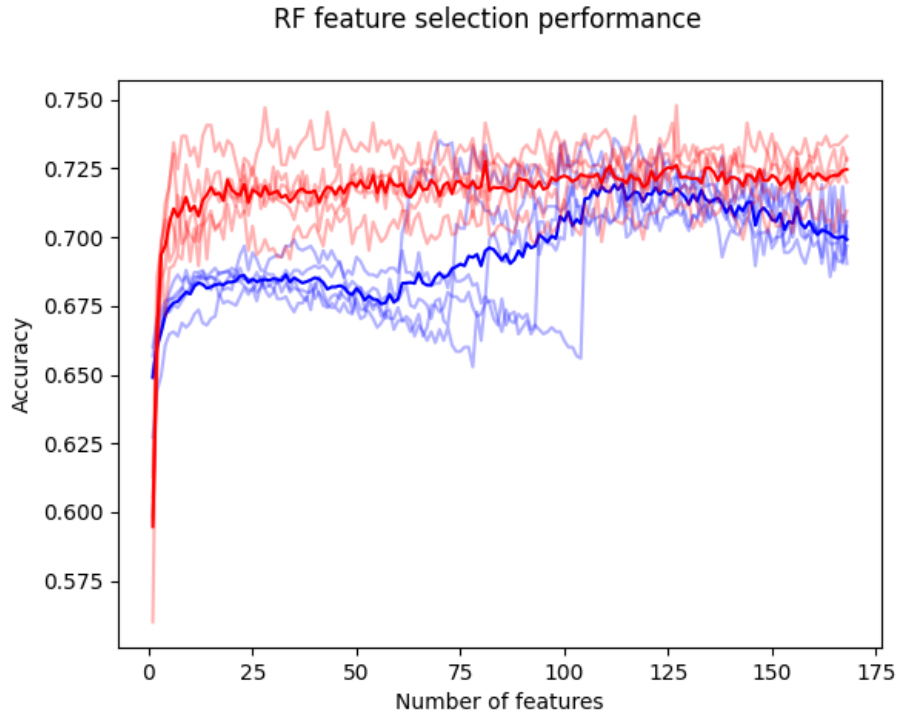
Random Forest

Figure 4.8: Difference in performance for the RF classifier, where each individual test is plotted with the average accuracy in bold. Tests using the SFS method are in blue while the SBS are in red.

Similarly to the neural network, the difference between the two methods is not vast in terms of performance. The SBS outperformed the SFS method, utilizing fewer features for a greater gain in accuracy. Secondly, like the two aforementioned classifiers, there is a variance between the different tests for both selection methods.

Lastly, in the case of the SFS method, there are spikes in all the tests, where the performance is greatly increased when adding just a few features.
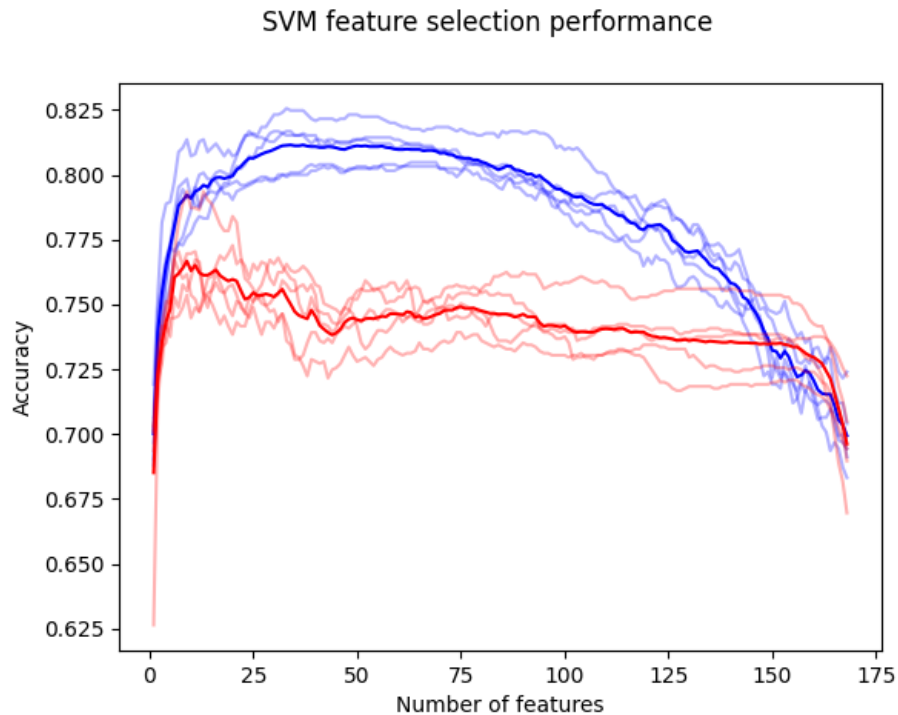
Support Vector Machine

Figure 4.9: Difference in performance for the SVM classifier, where each individual test is plotted with the average accuracy in bold. Tests using the SFS method are in blue while the SBS are in red.

Figure 4.9 shows that they are a significant difference between the two selection methods for the SVM. The SFS method achieved an average accuracy of over 80% while the SBS method only had one test which achieved an accuracy of over 78%. A final observation is that like all other classifiers there is a variance for both of the selection methods.

Chapter 5

Discussion

As mentioned in the result section, the morphological features were overrepresented in terms of being included in the optimal set. Similar results were achieved by [32] who compared deep features generated by a convoluted neural network against SIFT descriptors. They obtained the highest accuracy using only deep features. While mixing deep and SIFT features, the accuracy was decreased. Finally, only using SIFT features resulted in the worst observed performance. In our results, combining the SIFT and morphological features resulted in a higher accuracy than only using one of the types.

The most prominent feature was feature 151, which is a morphological feature. The feature measures the solidity or the compactness of the lesion. This might be due to the way the tests were conducted. In the forward selection, if a feature is selected early, then it is likely to be contained in the optimal set. Thus, the fact that the feature was so common might be an indicator that it was chosen early in the process. Which in turn indicates that the feature is a good predictor on its own. However, the feature was common in both SBS and SFS, thus it might simply be an effective predictor of malignant melanoma. This could be due to the fact that the compactness of a lesion is similar to describing the asymmetry, lesion which is one of the criteria in the ABCD method. The fact that this is a good indicator might be because malignant lesions usually grow in a more random manner than benign.

The second most prominent feature was feature 160, which is a color feature. As stated in the results, the feature measures the quotient of the average amount of red color within the lesion over the amount of red color in the surrounding skin. Thus, a value greater than 1,0 means that the lesion is ´´more red" than the surrounding skin and vice versa for a value lower than 1,0. It must be mentioned that a conclusion can not be made concerning whether a

high or low value, or that the lesion is more or less red than the surrounding skin, is a good indicator of melanoma.

An interesting result was the difference between the two selection methods for the SVM and KNN classifiers. In both cases the SFS method achieved higher maximum accuracies while at the same time using more features than the SBS. Furthermore, the results show that during the selection process, the SFS method had a higher average accuracy than the SBS when they utilized the same number of features. They do however converge toward the same accuracy in the beginning and end of the selection process. For the RF and SFS case there were spikes of accuracy in all tests when only a few features. These spikes occurred at different iterations of the selection process in all the tests which could be an indication of hidden synergies between some features. No conclusion can however be made in the matter since this would require further investigating which features were added and performing more tests to see if this is a reoccurring phenomenon.

Neural network and Random forest did not show any significant difference between the two selection methods. They also used more features on average in their most optimal set. Furthermore, these two models gained less in terms of performance compared to the SVM and KNN classifiers. Hence, the selection models seem to have difficulties to find effective features compared to the other classifiers.

Lastly, the graphs show that for all the methods, there was an early gain in terms of accuracy which then plateaus to then steadily decline when more and more features are added. The maximum accuracy was usually obtained towards the end of this plateau which results in a marginal gain in accuracy but a vast increase of features which increased the runtime of the algorithms. There might therefore be an interest in instead of solely measuring the accuracy, to instead introduce a two-factor system where the selection method also takes into account the increase of the feature space. Thus, the method can balance the optimal number of features dependent on both performance and runtime.

Chapter 6

Conclusions

From the results, this study concludes that morphological features were more important than SIFT features for our classifiers' accuracy. However, compared to solely using SIFT or morphological features, a combination of the two leads to a greater gain in accuracy.

Two features were of significant importance compared to all other and both of these features were morphological. The most prominent was a feature describing the compactness of the lesion and the second most important feature described the contrast between the lesion and the surrounding skin in terms of the color red.

There was some variance between the difference models regarding which features were the most important. The main difference between the models were how many features were used. The SVM and KNN classifiers used fewer features than the NN and RF classifiers.

Furthermore, the results show that the SFS method was more effective than the SBS method for the SVM and KNN classifiers. For the NN, SFS was slightly more effective while for the RF the SBS method lead to a greater gain. SBS used fewer features for the NN, SVM and RF classifiers while the opposite was true in RF case.

Bibliography

[1] Strålsäkerhetsmyndigheten. *Sverige har snart samma nivåer av hud-cancerfall som Australien.* `https://www.stralsakerhetsmyndigheten.` `se/press/nyheter/2021/sverige-har-snart-samma-` `nivaer-av-hudcancerfall-som-australien/`. Strålsäk-erhetsmyndigheten, June 2021.

[2] American Academy of Dermatology Association. *SKIN CANCER.* `https:` `//www.aad.org/media/stats-skin-cancer`. American Academy of Dermatology Association, Apr. 2022.

[3] *Skin Cancer Facts & Statistics.* `https://www.skincancer.` `org/skin-cancer-information/skin-cancer-facts/`. Accessed: May 23, 2023. 2019.

[4] Esra Mahsereci Karabulut, Selma Ayşe Özel, and Turgay İbrikçi. "A comparative study on the effect of feature selection on classification accuracy". eng. In: *Procedia technology* 1 (2012), pp. 323–327. ISSN: 2212-0173.

[5] Shivangi Jain, Vandana jagtap, and Nitin Pise. "Computer Aided Melanoma Skin Cancer Detection Using Image Processing". In: *Procedia Computer Science* 48 (2015). International Conference on Computer, Communication and Convergence (ICCC 2015), pp. 735–740. ISSN: 1877-0509. DOI: `https://doi.org/10.1016/j.procs.2015.` `04.209`. URL: `https://www.sciencedirect.com/science/` `article/pii/S1877050915007188`.

[6] Cancer Research UK. *Stages and types.* `https://www.cancerresearchuk.` `org/about-cancer/melanoma/stages-types`. Cancer Research UK, May 2020.

[7] Cancer Research UK. *Survival.* `https://www.cancerresearchuk.` `org/about-cancer/melanoma/survival`. Cancer Research UK, Mar. 2020.

[8]   N. C. F. Codella et al. "Deep learning ensembles for melanoma recognition in dermoscopy images". In: *IBM Journal of Research and Development* 61.4/5 (2017), 5:1–5:15. DOI: 10.1147/JRD.2017.2708299.

[9]   K. Das et al. "Machine Learning and Its Application in Skin Cancer". In: *International Journal of Environmental Research and Public Health* 18.24 (2021), p. 13409. DOI: 10.3390/ijerph182413409. URL: https://doi.org/10.3390/ijerph182413409.

[10] Catarina Barata, M. Emre Celebi, and Jorge S. Marques. "A Survey of Feature Extraction in Dermoscopy Image Analysis of Skin Cancer". In: *IEEE Journal of Biomedical and Health Informatics* 23.3 (2019), pp. 1096–1109. DOI: 10.1109/JBHI.2018.2845939.

[11] J. Jaworek-Korjakowska and P. Kłeczek. "Automatic Classification of Specific Melanocytic Lesions Using Artificial Intelligence". In: *BioMed Research International* 2016 (2016), p. 8934242. DOI: 10.1155/2016/8934242. URL: https://doi.org/10.1155/2016/8934242.

[12] Reda Kasmi and Karim Mokrani. "Classification of malignant melanoma and benign skin lesions: implementation of automatic ABCD rule". In: *IET Image Processing* 10.6 (2016), pp. 448–455. DOI: https://doi.org/10.1049/iet-ipr.2015.0385. eprint: https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/iet-ipr.2015.0385. URL: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-ipr.2015.0385.

[13] L.-M. Sánchez-Reyes et al. "A High-Accuracy Mathematical Morphology and Multilayer Perceptron-Based Approach for Melanoma Detection". In: *Applied Sciences* 10.3 (2020), p. 1098. DOI: 10.3390/app10031098. URL: https://doi.org/10.3390/app10031098.

[14] N. Smaoui and S. Bessassi. "A developed system for melanoma diagnosis". In: *International Journal of Computer Vision and Signal Processing* 3.1 (2013), pp. 10–17.

[15] David G. Lowe. "Distinctive Image Features from Scale-Invariant Keypoints". In: *International Journal of Computer Vision* 60.2 (2004), pp. 91–110. DOI: 10.1023/B:VISI.0000029664.99615.94. URL: https://doi.org/10.1023/B:VISI.0000029664.99615.94.

[16]   *KNN Algorithm.* IBM, n.d. URL: https://www.ibm.com/topics/knn.

[17]   *What are Neural Networks?* https://www.ibm.com/topics/neural-networks. [Accessed 22-May-2023]. IBM.

[18]   *Random Forest Algorithm.* IBM, n.d. URL: https://www.ibm.com/topics/random-forest.

[19]   Gareth James et al. *An Introduction to Statistical Learning.* New York, NY: Springer, 2013.

[20]   R. Chaganti et al. "Thyroid Disease Prediction Using Selective Features and Machine Learning Techniques". In: *Cancers* 14.16 (2022), p. 3914. DOI: 10.3390/cancers14163914. URL: https://doi.org/10.3390/cancers14163914.

[21]   Suleiman Mustafa. "Feature selection using sequential backward method in melanoma recognition". eng. In: *2017 13th International Conference on Electronics, Computer and Computation (ICECCO).* IEEE, 2017, pp. 1–4. ISBN: 1538624990.

[22]   *DullRazor.* http://www.dermweb.com/dull_razor/. [Accessed 22-May-2023]. dermweb.

[23]   antocad. *Melanoma-Classifier.* https://github.com/antocad/Melanoma-Classifier. GitHub Repository.

[24]   Sameena Pathan et al. "Classification of Benign and Malignant Melanocytic Lesions: A CAD Tool". In: Sept. 2017. DOI: 10.1109/ICACCI.2017.8126022.

[25]   M. Emre Celebi et al. "Automatic detection of blue-white veil and related structures in dermoscopy images". In: *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society* 32.8 (2008), pp. 670–677. DOI: 10.1016/j.compmedimag.2008.08.003. URL: https://doi.org/10.1016/j.compmedimag.2008.08.003.

[26]   *OpenCV Documentation: cv::SIFT Class.* https://docs.opencv.org/4.x/d7/d60/classcv_1_1SIFT.html. [Accessed 22-May-2023].

[27]   docongminh. *Classification-SIFT-SVM.* https://github.com/docongminh/Classification-SIFT-SVM. GitHub Repository.

[28]   Erdem Okur and Mehmet Turkan. "Patch Enhancement for Melanoma Detection with Bag of Visual Words". In: *2022 Medical Technologies Congress (TIPTEKNO)*. 2022, pp. 1–4. DOI: `10.1109/TIPTEKNO56568.2022.9960177`.

[29]   *scikit-learn Documentation: Supervised Learning*. `https://scikit-learn.org/stable/supervised_learning.html\#supervised-learning`. [Accessed 22-May-2023].

[30]   Ahmed Takiddin et al. "Artificial Intelligence for Skin Cancer Detection: Scoping Review". In: *Journal of Medical Internet Research* 23.11 (2021), e22934. DOI: `10.2196/22934`. URL: `https://doi.org/10.2196/22934`.

[31]   Sebastian Raschka. "MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack". In: *The Journal of Open Source Software* 3.24 (Apr. 2018). DOI: `10.21105/joss.00638`. URL: `https://joss.theoj.org/papers/10.21105/joss.00638`.

[32]   Sheng Zhang, Xiaoning Lu, and Xiaojun Du. *Optimization of Data Reconciliation Based on Sensitivity and Gross Error Analysis*. Cham, 2018.