Degree project in computer engineering

First cycle, 15 credits

Determining Important Features for Melanoma Classification Through Feature Selection

VILMER JONSSON

TOR STRIMBOLD

Stockholm, Sverige 2023

Determining Important Features for Melanoma Classification Through Feature Selection

VILMER JONSSON
TOR STRIMBOLD

Bachelor's Degree in Computer Science
Date: May 24, 2023
Supervisor: Kevin Smith
Examiner: Pawel Herman
School of Electrical Engineering and Computer Science
Swedish title: Värdering av attribut för klassificering av hudcancer

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 Machine Learning

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 borders of the lesion are identified. This is done to be able to identify which pixels in the image constitutes the lesion. 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. Similarly to random forest, it performs a majority voting and then chooses the most occurring class from neighbors as the final prediction. [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 the data is fed into the network. The last layer is the output layer, where the final prediction is made. The layers in between are called hidden layers. The nodes in the hidden layers are connected to all nodes in the previous layer, and the output of each neuron is calculated by a function. 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. The model consists of several decision trees taking different factors into consideration. When classifying, all individual trees try to classify the data. Then, the class that most of the trees predicted is chosen as the final classification. [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 of 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 the each of the remaining features are evaluated together with the added feature. This is repeated until all features are added or until som criterion is met.[20]

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 as well as 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 Coldman 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 locations, 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

The aim of this study was not to develop 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 published by [name] available on GitHub.

3.1.3 Feature extraction

As mentioned in the background, the two feature extraction methods used were ABCD and SIFT.

Morphological

The asymmetry and border features are calculated through the method proposed by [23] and the color features by [24]. Both computations showed promising results reaching an accuracy of 82% and 83%. Furthermore, other morphological features are calculated using the scikit skimage measure package. The implementation is originally done in the way described in this Github repository [25].

To calculate the asymmetry and border features the measure library from sci-kit is used, with the `regionprops_table` function. The color features are calculated using a numpy array to represent the image's RGB value.

SIFT

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

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 [27].

Although all ML models' parameters can be fine-tuned to optimize performance and run-time, the default parameters will be used since it is not the aim of the study to create an optimal implementation. The focus is instead on the feature selection's effect on the performance of the classifiers.

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. [28]

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 `Scikit-sklearn` library `feature_selection`.

The manner in which the feature selection was performed differs slightly due to the selection method employed. For the SFS, the feature set was initially empty. The method then adds all features, one at a time. For the SBS the feature set at first contained all features, whereupon features were removed one at a time, until the feature set consisted of only one feature. In each iteration, the accuracy and which features constituted the feature set were 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 optimal 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 the 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

4.1 Determining most important features

Below is a table showing the features that were included in the optimal set the most and least amount of times. The features are sorted in descending order based on how many times they were included in the optimal feature set.

In the following table the thirteen most prominent, and twelve least prominent features are presented, overall.

Table 4.1: The most and least prominent features in all 40 tests.

Feature	Occurence(s)	Probability
151	35	0.875
160	29	0.725
80	24	0.6
100	24	0.6
158	24	0.6
27	23	0.575
165	22	0.55
26	21	0.525
29	21	0.525
32	21	0.525
40	21	0.525
91	21	0.525
162	21	0.525
8	11	0.275
61	11	0.275
72	11	0.275
96	11	0.275
115	11	0.275
116	11	0.275
117	11	0.275
156	11	0.275
78	10	0.25
135	10	0.25
86	9	0.225
143	8	0.2

In the table it is observable that two features stand out: number 151 and 160. These are both morphological features.

Feature 151 is called "Extent" and measure how many of the pixels in an image are within the lesion. The extent feature was also the only feature that was included in the optimal feature set for all classifiers.

Feature 160 is a color feature that measures the red color of the image. It calculates the mean of the red color within the lesion and divides it by the mean of the red color in the parts of the image outside the lesion.

We can also see that there is a large spread, only two features are in the optimal feature set less than 25% of the tests. Furthermore, the occurrences quickly deteriorate where the most prominent feature has 35 occurrences whereas the seventh most prominent feature only has 22 occurrences, which is a total difference of 13 occurrences. The downwards trend does however flatten out quickly as there are 52 features that have the occurrence of 20-17 occurrences. Lastly, the final span of 17 to 8 occurrences entails all the remaining features.

Furthermore, we can also compare the two different sets of features, the handcrafted ABCD features and the automatically generated SIFT features.

Table 4.2: Average number of occurrences in optimal set for ABCD and SIFT features.

Feature Type	Feature count	Occurences	Average Occurences
ABCD	19	358	18.8
SIFT	150	2340	15.6

From the results, 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 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

To determine which features are important for different classifiers, the optimal feature set was examined in every test to see how many times a feature was included in the set. The results for each classifier is listed in the rubrics below.

4.2.1 K-nearest neighbors

Below are the two tables showing the most important features for KNN using SFS and SBS respectively. The features are sorted in descending order based on how many times they were included in the optimal feature set.

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

Feature	Occurrence(s)	Probability	Probability in SFS	Probability in SBS
151	9	0.9	1.0	0.8
20	5	0.5	1.0	0.0
27	5	0.5	0.0	1.0
37	5	0.5	1.0	0.0
32	5	0.5	0.8	0.2
34	5	0.5	0.8	0.2
111	5	0.5	0.8	0.2

The results show that the SFS preferred more features than the SBS counterpart. Furthermore, there is very little overlap between the preferred features when comparing the two methods. The only feature that has a probability of 80 percent or more is number 151, which were the most preferred across all classifiers. However, apart from the most common feature, the probability that a feature is included in the optimal set, is at most 50%. Furthermore, this probability is shared by the following six features. Lastly, there was very little overlap in which features were included for the two methods.

4.2.2 Neural Network

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

Feature	Occurrence(s)	Probability	Probability in SFS	Probability in SBS
80	10	1.0	1.0	1.0
0	9	0.9	1.0	0.8
84	9	0.9	1.0	0.8
26	9	0.9	0.8	1.0
140	9	0.9	1.0	0.8
165	9	0.9	1.0	0.8
66	9	0.9	1.0	0.8
58	9	0.9	1.0	0.8
166	9	0.9	0.8	1.0
106	9	0.9	1.0	0.8
160	9	0.9	1.0	0.8
40	9	0.9	1.0	0.8
158	9	0.9	1.0	0.8
35	9	0.9	1.0	0.8
157	9	0.9	0.8	1.0
136	9	0.9	1.0	0.8
6	9	0.9	1.0	0.8
147	9	0.9	1.0	0.8
81	9	0.9	1.0	0.8

The results show that the classifier included more features in the optimal set compared to other classifiers. The most prominent feature was 80, which is a SIFT feature which occurred in the optimal feature set in all tests. Furthermore, every feature was present in at least three of the tests. This can be seen in the appendix where the full table is presented.

4.2.3 Random Forest

Table 4.5: The features with at least 8 occurences in the 10 tests.

Feature	Occurence(s)	Probability	Probability in SFS	Probability in SBS
18	9	0.9	1.0	0.8
94	9	0.9	1.0	0.8
112	9	0.9	1.0	0.8
131	9	0.9	1.0	0.8
100	9	0.9	0.8	1.0
132	8	0.8	0.6	1.0
58	8	0.8	0.6	1.0
95	8	0.8	0.6	1.0
80	8	0.8	0.8	0.8
75	8	0.8	0.8	0.8
26	8	0.8	0.6	1.0
37	8	0.8	0.8	0.8
33	8	0.8	0.8	0.8
29	8	0.8	0.8	0.8
128	8	0.8	1.0	0.6
151	8	0.8	1.0	0.6

Random forest, similar to Neural Network, included a lot of features in the tests. The most prominent features were 18, 94, 112, 131 and 100, which were SIFT features. They were present in the optimal feature set in all tests except one. In general the classifier seemed to prefer SIFT features where the highest ranked ABCD feature was 151 with a probability of 80%.

4.2.4 Support Vector Machine

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

Feature	Occurence(s)	Probability	Probability in SFS	Probability in SBS
151	10	1.0	1.0	1.0
161	10	1.0	1.0	1.0
160	9	0.9	1.0	0.8
162	7	0.7	0.6	0.8
164	6	0.6	0.8	0.4
163	6	0.6	1.0	0.2
145	5	0.5	0.4	0.6
27	5	0.5	0.4	0.6
158	5	0.5	0.4	0.6
165	5	0.5	0.8	0.2
100	5	0.5	0.4	0.6

The SVM differs from the other classifiers, with a preference for morphological features. Consequently, all except number 151 were color features from the ABCD method. Furthermore, the occurences of features decrease quite rapidly where the seventh to eleventh most prominent features only have a probability of 50% of being present in the optimal set.

4.3 Difference between selection methods

To examine eventual differences between the methods, two type of results were gathered. Firtly, the average number of features included in the optimal set. This was calculated by recording the number of features present in the optimal set for every test and then calculating the average of these. Secondly, the average difference in performance between the methods for every classifier. This was done by calculating the average accuracy of every iteration in the tests. The results are then plotted for every classifier.

Following are the comparison between the two different 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
Neural Network	SFS	126.2
Neural Network	SBS	98.2
K-Nearest Neighbors	SFS	43.4
K-Nearest Neighbors	SBS	21.4
Support Vector Machine	SFS	32.6
Support Vector Machine	SBS	25.4
Random Forest	SFS	86.2
Random Forest	SBS	106.2

The results show that the SFS method used on average more features than 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 twice the number of features than the SBS method. However, no test used all of the available features to achieve the highest accuracy.

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

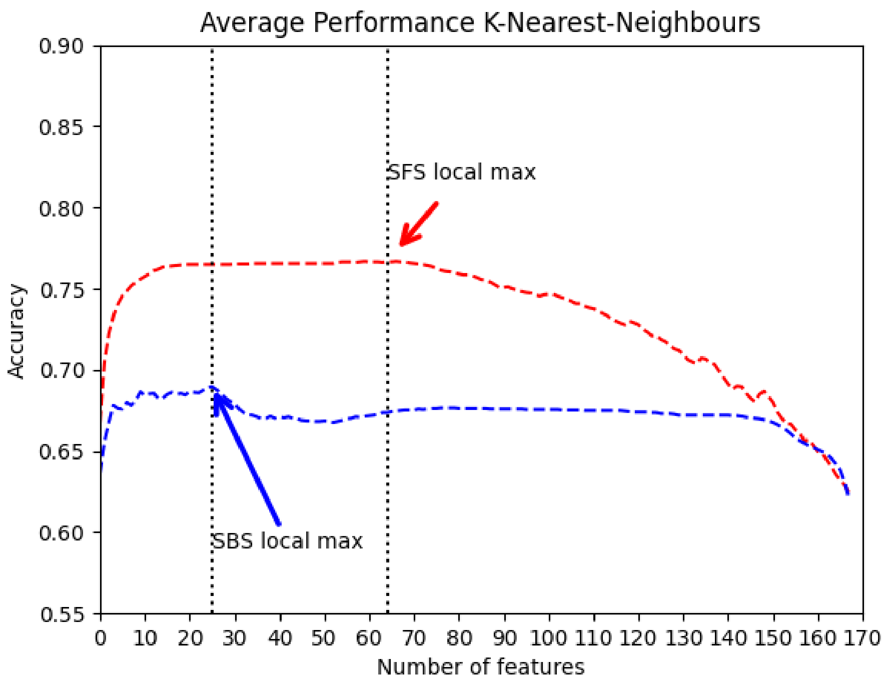

Figure 4.1: Difference in performance for the KNN classifier

The results show that the SFS achieved a greater increase in accuracy compared to the SBS method with an accuracy of about 78% compared to 70%. However, 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.

Neural Network

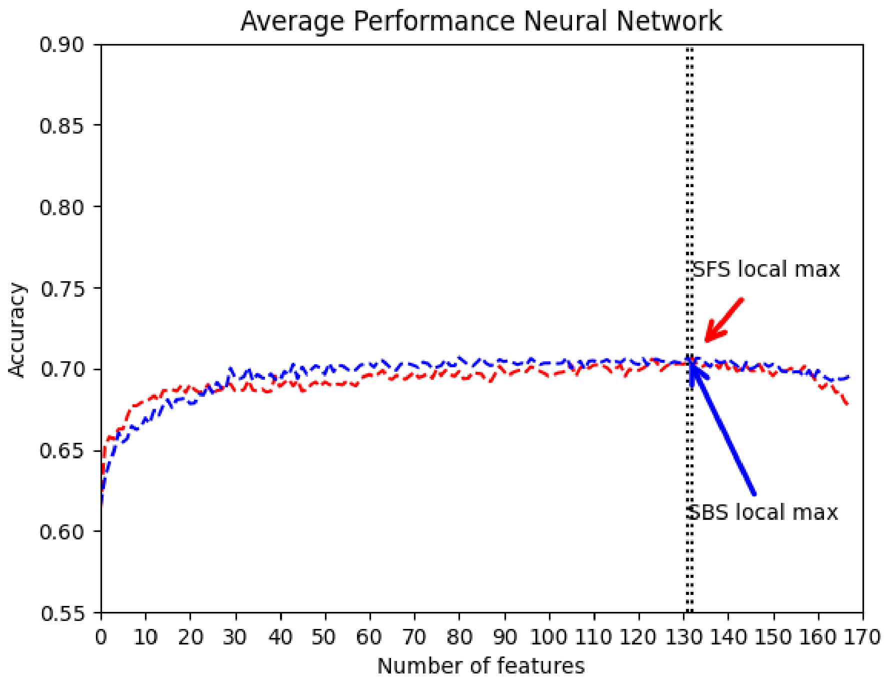

Figure 4.2: Difference in performance for the NN classifier

Figure 4.2 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.

Random Forest

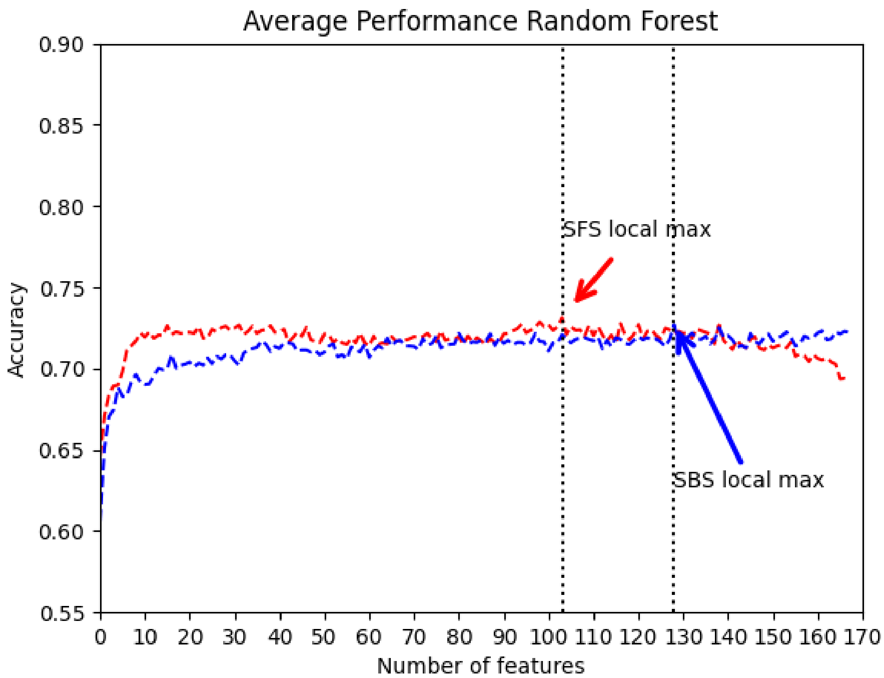

Figure 4.3: Difference in performance for the RF classifier

Similarly to the neural network, the difference between the two methods is not vast in terms of performance. The difference lies however within the number of features used where the SFS used fewer features than the SBS method. Furthermore, it is observable that there is minimal gain for the SFS method when using over 100 features compared to using between 30 and 10 features.

Support Vector Machine

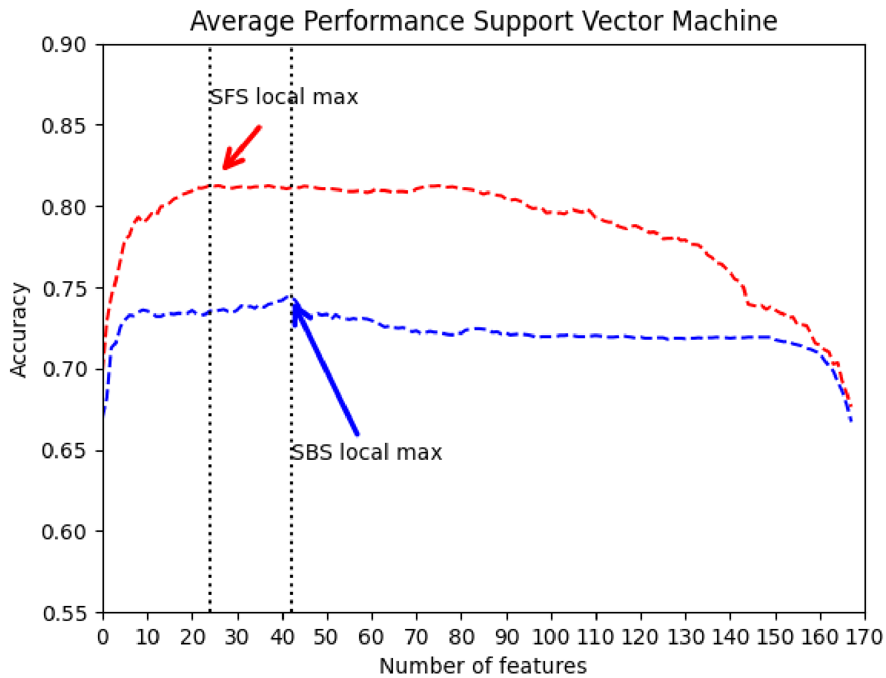

Figure 4.4: Difference in performance for the SVM classifier

Figure 4.4 shows that they are a significant difference between the two selection methods for the SVM. The SFS method achieved an accuracy of over 80% while the SBS method never achieved an accuracy of over 75%. Furthermore, the SFS method used less features than the SBS method.

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 [29] 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.

As mentioned in the results, the most prominent feature was feature 151, which is a morphological feature. The feature measures the extent of the lesion, i.e. how many of the pixels in the image are within 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 strong indicator on its own. However, the feature was common in both SBS and SFS, thus it might simply be an effective indicator of malignant melanoma.

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 high or low value indicates a contrast in terms of red colors between the lesion and the surrounding skin. The prominence of this feature might indicate that a contrast in terms of red color between the surrounding skin and the lesion is a common indicator of malignant 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. This indicates that there might be some hidden synergies between features where some combinations give rise to higher performance. Furthermore, this can be an indicator of that it is easier for the selection method to find these synergies when adding features rather than removing them.

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.

Furthermore, the graphs depicting the average performance of the selection methods for every classifier, and the tables presenting the average number of features used to obtain the highest accuracy for every selection method for every classifier, does not match in terms of the number of features used to obtain the highest accuracy. For example, in the case of the SVM classifier, the SFS used on average 32,6 features and the SBS 25,4 features to obtain the highest accuracy. Meanwhile, the graph shows that the SFS obtained the highest accuracy around 24 features and the SBS around 42 features. This difference is most likely due to the method for obtaining the results. In the former case, the accuracy of every iteration of every test was recorded. The average accuracy for every iteration was then calculated, and the graph was plotted. In the latter case, the maximum accuracy and which features constituted the optimal set was recorded. The features were then counted and the average number of features used could hence be calculated. The difference between the two test might be reduced if the number of tests performed was increased.

Lastly, the graphs show that for all of the methods, there is an early gain in terms of accuracy which then plateaus to then steadily decline when more and more features are added. The maximum accuracy is usually obtained towards the end of this plateau which results in a marginal gain in accuracy but a vast increase of features which increases the runtime of the algorithms. This is especially true for the KNN classifier using the SFS method, where the accuracy plateaus at around 15 features but the maximum is achieved at around 65 features. 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, we can say that morphological features are more important than SIFT features for our classifiers' accuracy. However, there is a variance inside the morphological feature set and the best accuracy were achieved with a combination of morphological and SIFT features.

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 is more effective than the SBS method for the SVM and KNN classifiers. However, for the RF and NN classifiers, the two methods were equally effective.

SBS used fewer features for all models except NN, but SFS achieved better accuracy when using as many features as the SBS.

Bibliography

[1] Strålsäkerhetsmyndigheten. *Sverige har snart samma nivåer av hudcancerfall som Australien*. https://www.stralsakerhetsmyndigheten.se/press/nyheter/2021/sverige-har-snart-samma-nivaer-av-hudcancerfall-som-australien/. Strålsäkerhetsmyndigheten, 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] Sameena Pathan et al. "Classification of Benign and Malignant Melanocytic Lesions: A CAD Tool". In: Sept. 2017. DOI: 10.1109/ICACCI.2017.8126022.

[24] 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.

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

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

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

[28] 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.

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