

Less is more: Finding key features for skin cancer diagnosis

Vilmer Jonsson, Tor Strimbold

May 19, 2023

Our work focuses on determining what information is important when diagnosing skin cancer using different machine learning algorithms.

Traditionally, diagnosing skin cancer has been done by clinicians, but with the advancement in computer science it has become possible to create algorithms that perform this task. This is done by feeding the computer images of skin lesions, one at a time. The computer makes a prediction concerning the lesions' nature. Based on whether the prediction was correct or not, adjustments to the evaluation method are made.

However, computers cannot simply look at a picture of a skin lesion and understand what the picture depicts, let alone make a judgment concerning its medical nature. Hence, the image needs to be quantified where information is extracted and given a numerical value, which is called a feature. It is using these features that the computer can make comparisons.

Why is it important to determine what information is necessary for an accurate diagnosis?

Reducing the amount of information fed to the algorithms is like cleaning your workspace. Just as a tidy desk allows for better focus and productivity, minimizing the amount of information to process allows the computer to make better predictions faster.

Apart from the technical perspective, increasing the efficiency of algorithms has a value from a medical point of view. The survival rate of skin cancer patients depends greatly on when the disease is diagnosed and treatment can begin. Hence, if effective and cheap algorithms can be implemented, automated diagnostic tools can be made more available. Thus making it easier to diagnose skin cancer at an earlier stage, increasing the chance of survival.

How do you determine which information is important?

Common methods for determining important features are forward and backward selection. Forward selection adds one of the features and observes the impact it has on the performance. The feature with the best effect is added and the process is repeated. Backward selection does the opposite, it removes the features with the worst impact one at a time. The processes are done when all features have been added or removed. In this manner we can find the important

features as the ones that resulted in the best performance.

These methods were applied on two types of features: ABCD and Scale-Invariant Feature Transform (SIFT).

The ABCD method is a common way of mimicking and simulating the practice of clinicians where the lesion's Asymmetry, Border, Color and Diameter is the basis for the diagnosis. For example, if the lesion has many different and high-contrast colors, it is more likely to be malignant.

SIFT features are generated by an algorithm which finds high-contrast spots in the image. We can then get information about these spots, and count how many times we find some specific features in the image.

What did we find?

All of the examined algorithms achieved their highest accuracy using fewer features. Furthermore, the ABCD features were more likely to have a greater impact on the performance than the SIFT features.

We also found great variance between how many features different algorithms preferred. Furthermore, forward selection led to better performance in all tests but also resulted in more features being used than when using backward selection. This information could be used to further improve and research the automated diagnosis of skin cancer by only extracting relevant features.