Skin Lesion Classification using Machine Learning Algorithms

Ilker Ali OZKAN #1, Murat KOKLU 1

1 Computer Eng., Faculty of Technology, Selcuk Uni., Konya –TURKEY
# Corresponding Author: Email: ilkerozkan@selcuk.edu.tr

Abstract: Melanoma is a deadly skin cancer that breaks out in the skin’s pigmented cells on the surface. Melanoma causes 75% of the skin cancer-related deaths. This disease can be diagnosed by a dermatology specialist through the interpretation of the dermoscopic images in accordance with ABCD rule. Even if dermatology experts use dermatological images for diagnosis, the rate of the correct diagnosis of experts is estimated to be 75-84%. The purpose of this study is to pre-classify the skin lesions in three groups as normal, abnormal and melanoma by machine learning methods and to develop a decision support system that should make the decision easier for a doctor. The objective of this study is skin lesions based on dermoscopic images PH2 datasets using 4 different machine learning methods namely; ANN, SVM, KNN and Decision Tree. Correctly classified instances were found as 92.50%, 89.50%, 82.00% and 90.00% for ANN, SVM, KNN and DT respectively. The findings show that the system developed in this study has the feature of a medical decision support system which can help dermatologists in diagnosing the skin lesions.

Keywords: Skin lesion, classification, Medical Decision Support System, Machine Learning Methods

1. Introduction

Cancer is a disease that occurs through multiplying of the body cells in an uncontrolled manner and occupying the peripheral tissues. Although the skin cancer disease occurs less frequent than many other cancer types, it is highly important because of its high mortality. The skin cancer has different of types, such as Malignant Melanoma, Squamous cell carcinoma, and Basal cell carcinoma [1]. Melanoma incidence is being reported to increase more rapidly than the other forms of cancer. Melanoma is responsible for 4% only of all skin cancers, whereas it is responsible for 75% total of skin cancer deaths [2]. Melanoma, which is thought to be stimulated by ultraviolet rays, is more commonly occurring in areas where exposure to sunlight is relative higher. In Europe are detecting 62,000 new cases each year [3]. According to the American Cancer Society's report 2016 was for the year 2016 foreseen that 76,380 cases will be diagnosed with melanoma in the United States and 10.130 people will die from it [4].

The natural development of melanoma takes place in two stages except for the nodular type. The horizontal or radial development stage that progresses along the epidermal surface, is defined as “single cancer melanoma” which has a critical importance for the early diagnosis. The melanoma progresses to the vertical development stage, if it couldn't be diagnosed in this period. It obtains at this stage a potential for metastatic spread [5]. The melanoma is a disease with a survival rate of approximately 90% on condition that it must be diagnosed early enough, whereas no effective treatment is available for the delayed cases [2, 6].

Dermoscopy is a non-invasive diagnostic method that allows us to examine in more detail the morphological structure of the pigmented skin lesions. The melanoma diagnosis is performed by interpreting the images obtained with the dermoscopy device. The dermoscopy device allows through the harsh light the detailed visualization of the morphological structures and patterns. Dermatologists usually perform the diagnosis of melanoma through these images by ABCD (Asymmetrical Shape, Border, Color and Diameter) rule. ABCD is a highly subjective assessment that is dependent on the experience and knowledge of the related doctors [4].

Despite the use of dermoscopy to examine the subcutaneous tissue without surgery, successful results are dependent on intensive dermatology training and experience. A reliable diagnosis of the melanoma with this method is unfortunately often not possible, especially in early stages. For that reason, an automatic diagnostic tool becomes an inevitable need [7].

The melanoma diagnosis can be improved with the ABCD rule based and computer assisted systems. These systems usually consist of the separate units for the image segmentation, feature extraction and classification respectively [8-12]. Studies conducted in this field are as follows:

Baldrick et al. compared in their study the expert opinion and artificial neural networks when they classify the lesions. They obtained from the computer program a sensitivity of 95% and a specificity of 88%, while they measured the expert dermatological sensitivity and specificity as 95% and 90% respectively [13].

Moataz et al. practised upon a genetic algorithm with an artificial neural network technique for early detection of the skin cancers and obtained a sensitivity of 91.67% and a specificity of 91.43% [14].

Kamasak et al. classified dermoscopic images by extracting the Fourier identifiers of the lesion edges after dividing the dermoscopic images. They obtained an accuracy of 83.33% in diagnosing of the melanoma [15].

Fidan et al. succeeded in an exact classification of 93.33% according to data extracted from the PH2 data set by using an artificial neural network that was formed for the abnormal and melanoma skin cancers [17]. Baştürk et al. used a new method of detecting melanoma skin cancers called Deep Neural Network (DNN) in their study and reached an accuracy of 91.85% in disease diagnosis [18].
In the second part of this study were information given about data set, machine learning methods and performance measurement methods. In the third part, the classification studies and the results obtained from these studies were given by comparing with the studies in the literature.

2. Material and Methods

2.1. PH² Dataset

A diagnostic study were performed with the machine learning algorithms formed for melanoma diagnosis in the PH² data set. This data set was established by a group of researchers from the Technical Universities of Porto and Lisbon in the dermatology service to Pedro Hispano Hospital. The PH² dataset contains 200 dermoscopy images at 768x560 resolution. Each image has 8-bit RGB channels [16].

![Views of the sample lesions in the PH² data set](Fig. 1)

In the PH² data set are available 80 images for the normal type, 80 images for the abnormal and 40 images for the melanoma respectively. Some examples of these are shown in Figure 1. Although the PH² data set was established by extracting the features according to the ABCD rule criteria, the criterion B was ignored hereby. For that reason, the features found in the dataset and used in the study are given in Table 1.

Table 1. Features of the PH² data set

<table>
<thead>
<tr>
<th>ABCD Criterion</th>
<th>Criterion Name</th>
<th>Criterion Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Asymmetry</td>
<td>0-Even-symmetric</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1- Axially symmetric</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-Asymmetric</td>
</tr>
<tr>
<td>B</td>
<td>Edge</td>
<td>Not available in the PH² data set</td>
</tr>
<tr>
<td>C</td>
<td>Colours (white, red, light brown, dark brown, blue-gray, black)</td>
<td>1-Applicable</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0-Not applicable</td>
</tr>
<tr>
<td>D</td>
<td>Pigment Network</td>
<td>0-Normal</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1-Abnormal</td>
</tr>
<tr>
<td></td>
<td>Dotted-beaded structure</td>
<td>0-Normal</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1-Abnormal</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-No observed</td>
</tr>
<tr>
<td>ABCD</td>
<td>Diagnosis</td>
<td>0-Normal Nevus</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1-Abnormal Nevus</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-Melanom</td>
</tr>
</tbody>
</table>

2.2. Machine Learning Methods

With this study, four different classifying techniques based on dermoscopic images of the data sets were applied on the skin lesions. Short information about each of the classifying techniques, i.e. ANN, SVM, KNN and DT are given in the following paragraphs.

Artificial Neural Network (ANN): ANNs are mathematical systems consisting of many process units (neurons) connected with each other in a weighted manner. The process unit receives signals from other neurons; combines, transforms them and generates a numerical result. In general, the process units are corresponding roughly to the real neurons and are interconnected in a network, so that this structure constitutes the artificial neural networks [19].

SVM: SVMs are nonparametric classifiers. Regarding their distribution is no preliminary information as a presupposition available. Inputs and outputs are paired in the training sets. Through the pairs, decision functions are obtained which classify the input variables in the test set and new data set. The task is here to be able to find out the line with the highest margin from the infinite number of lines that can classify the data, when a linearly separation were possible. It uses a non-linear mapping for transforming the original work data into a higher dimension, when a linearly separation were impossible. In the new transformed dimension is being investigated then the (optimal) separator plane with the maximum margin [20].

K-Nearest Neighbor (KNN): The KNN (K-Nearest Neighbor) algorithm is one of the most basic sample-based learning algorithms. In example based learning algorithms, the learning process is performed based on the data held in the training set. A new faced example is categorizing according to similarities with the examples in the available training set [21].

Decision tree (DT): The decision tree is a classifier algorithm in the structure form of a “tree”. Decision Trees are simple, but very commonly used methods by moving the inductive logic into a programming environment. It works with discrete valued parameters. The basic intuition about the inductive philosophy on which the decision tree algorithms are based is that a “good” decision tree to be constructed with learning characteristics should be small as possible [22].

2.3. The Commonly-Accepted Performance Evaluation Measures

This is the case we focus on in this study. Classification performance without focusing on a class is the most general way of comparing algorithms. It does not favor any particular application. The introduction of a new learning problem inevitably concentrates on its domain, but omits a detailed analysis. Thus, the most used empirical measure and accuracy does not distinguish between the numbers of correct labels of the different classes. [23]:

\[ TP = \text{true positives: number of examples predicted positive that are actually positive} \]

\[ FP = \text{false positives: number of examples predicted positive that are actually negative} \]

\[ TN = \text{true negatives: number of examples predicted negative that are actually negative} \]

\[ FN = \text{false negatives: number of examples predicted negative that are actually positive} \]

Accuracy: It refers to the total number of records that are correctly classified by the classifier. Accuracy of a classifier is defined as the percentage of test set tuples that are correctly classified by the model [24].

\[ Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \times 100\% \] (1)
Sensitivity: Refers to the true positive rate that means the proportion of positive tuples that were correctly identified [24].

\[ \text{Sensitivity} = \frac{TP}{TP+FN} \times 100\% \]  
(2)

Specificity: Indicates the rate at which a test or diagnostic method sets a correct (ie negative) diagnosis for a patient who is not ill.

\[ \text{Specificity} = \frac{TN}{TN+FP} \times 100\% \]  
(3)

Balanced accuracy (BACC): The balanced accuracy, which can be defined as the average accuracy obtained on either class [25].

\[ BACC = \frac{TP+TN}{2} \times 100\% \]  
(4)

Precision: The fraction of retrieved instances that are relevant [24].

\[ \text{Precision} = \frac{TP}{TP+FP} \times 100\% \]  
(5)

F- measure: The F- measure also refers to F measures that combined both the measures Precision and Recall as the harmonic mean [24].

\[ F - \text{measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100\% \]  
(6)

ROC Curve: Receiver Operating Characteristics curved showed both sensitivity and specificity of the test. The comparison of TPR (True Positive Rate) and FPR (False Positive Rate) is defined as ROC curve. The TPR is the proportion of positive tuples that are correctly labeled by the model whereas FPR is of negative tuples misclassified as positive [24].

i.e. TPR = TP (TP+FN) and FPR = FP (FP+TN)

3. Results and Conclusions

In this study for a melanoma diagnosis; Artificial Neural Network (ANN), Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Decision Tree (DT) classifiers were compared with each other within a PH² data set. Categorical values are coded by “one-of-N coding” for entries performed in this study.

The experimental studies have shown that the optimum value for “k” amounted to 5 and 10 in the k-fold cross-validation method. [19, 26]. As shown in Fig. 2, in this study the data set divided into 10 parts by using a 10-layer cross-validation method. The system is trained and tested with “k” different training and test clusters whereas for each case the “k” performance measures could be obtained. Thus, the arithmetic mean of the obtained “k” performance measures is calculated to determine the success of the cross validation [27].

This study has been performed using the functions of MATLAB Statistics and Machine Learning Toolbox and MATLAB Neural Network Toolbox [28, 29]. The ANN structure formed consists of three layers. All the input parameters (12 units total) in the data set are establishing the input vector at Layer 1. Layer 3 is the layer that indicates the output of the classification and the number of neurons in this layer depended on the number of classes in the output. Scaled conjugate gradient backpropagation algorithm is used as the learning algorithm [30]. Back propagation training parameters are given in Table 2.

![Fig. 2. 10-Fold cross validation method](image)

Table 2. Backpropagation training parameters

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning Rate</td>
<td>0.01</td>
</tr>
<tr>
<td>Momentum Constant</td>
<td>0.9</td>
</tr>
<tr>
<td>Epochs</td>
<td>1000-10000</td>
</tr>
<tr>
<td>Performance Function</td>
<td>Crossentropy</td>
</tr>
<tr>
<td>Minimum Performance Gradient</td>
<td>1/(e⁻¹⁰)</td>
</tr>
</tbody>
</table>

Network structures containing from 2 to 50 neurons were trained on the given network structure to find out the number of the hidden layer neurons with the best result. The established network structures in the study were examined and the ANN architecture with 18 neurons was used in the hidden layer with the best accuracy (92.50%).

Table 3 shows the performance results for classifying PH² data set obtained with Artificial Neural Network (ANN), Support Vector Machine (SVM), K-nearest neighbors (KNN) and decision tree (DT) classifiers by using the 10-fold cross-validation.

Table 3. Performance rates of the application

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy (%)</th>
<th>Balanced Accuracy (%)</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Precision (%)</th>
<th>F1-Score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN</td>
<td>92.50</td>
<td>93.49</td>
<td>90.86</td>
<td>96.11</td>
<td>92.38</td>
<td>90.45</td>
</tr>
<tr>
<td>SVM</td>
<td>89.50</td>
<td>90.35</td>
<td>86.25</td>
<td>94.44</td>
<td>89.09</td>
<td>87.31</td>
</tr>
<tr>
<td>KNN</td>
<td>82.00</td>
<td>85.04</td>
<td>79.58</td>
<td>90.49</td>
<td>81.45</td>
<td>80.33</td>
</tr>
<tr>
<td>DT</td>
<td>90.00</td>
<td>90.97</td>
<td>87.08</td>
<td>94.86</td>
<td>88.58</td>
<td>87.70</td>
</tr>
</tbody>
</table>

Table 3 shows that ANN has an accuracy of 92.50%, SVM of 89.50%, KNN of 82.00% and DT of 90.00%. This suggests that the proposed ANN has a clearly better classification performance for the PH² data set. The accuracy values of the ANN, SVM, KN and DT algorithms for each classifier output are given in Fig. 3 for the classification of skin lesions according to the data from the PH² data set. The ANN classifier appears to be more successful in classifying each skin lesion than other algorithms.
that ANN has more successful classified the PH² data set than SVM, KNN and DT. An accuracy of 92.50% achieved with the ANN classifier reveals that this classifier is a medical decision support system which could help dermatologists to diagnose the skin lesions.

Previous studies including upper mentioned on the related data set are summarized with the accuracy ratios in Table 4. Compared with the literature studies given in Table 4, the ANN structure, obtained in this study for classification of the skin lesions, has only in the sensitivity of the skin lesions has a lower value than that of Barata et al. and Marques et al., whereas our study has a higher specificity value. Additionally this study has distinguished with its higher accuracy, specificity and balanced accuracy ratio compared to all other studies. This study maybe further progressed by using the different preliminary data processing techniques and hybrid classification algorithms. In addition, this study can be combined with the related image processing techniques also to be able to make autonomous decisions in several medical issues.

### References


