Very Deep Convolutional Networks for Skin Lesion Classification

M. Alnowami

Department of Nuclear Engineering, King Abdulaziz University, 80204 Jeddah 21589, Saudi Arabia

malnowaimi@kau.edu.sa

Abstract. In medical imaging and diagnostic radiology, visual examination of a skin lesion is one of many potential applications of Computer-Aided Diagnosis (CAD). In this paper, a classification algorithm was developed and trained based on Densely Connected Convolutional Networks (DenseNets) for skin lesion classification. This work combines the use of dermatoscopic images pre-processing and deep convolutional network. Experimental studies were conducted using more than 30,000 dermatoscopic images from multiple open-access archives. Initially, the paper examined the impact of proposed image pre-processing and tuning on the accuracy of a binary classification CNN. The result showed that the binary classifier achieves a higher area under the curve (AUC), 0.93, when trained on processed data compared to 0.85 before. In the second stage, two different data-sets were created: a 3-ary classes and 9-ary classes. The validation results showed that the proposed work achieved validation accuracy of 81.2 ± 1.1% in the 3-ary and of 60.1 ± 1.3%. In the 9-ary classification studies. The proposed combination of dermatoscopic images pre-processing and deeper convolutional network can achieve better performance by learning more complex features of the input data with a more efficient memory implementation.

Keywords: Classification, Skin, Lesion, Computer-aided diagnosis, Neural network.

1. Introduction

Skin cancer is increasingly recognized as a worldwide public health concern [1]. The incidence and mortality rates of skin cancer are rapidly growing worldwide [2-4]. There are many factors may affect the survival rate such as stage of the tumor, thickness, anatomic site, type of melanoma and ulceration [5].

Early detection plays a crucial role in preventing melanoma mortality. Several studies and clinical evidence have shown the thickness of the tumor at diagnosis is the factor most predictive of melanoma survival that emphasize the belief that early detection of skin cancer results in better chances of survival [6]. To be able to detect cancer in its early stages, a screening process needs to be used. The conventional skin screening approaches require the expertise of dermatologists and carry with them high costs and delays [7]. The process of detecting skin cancer starts with examining the patient visually, and, from there, a decision can be made to either take further steps in scanning and diagnostics or release the patient with a clean bill of health. In many cases, the initial screening followed by more detailed dermoscopic analysis, a biopsy, and a pathological examination.
This process requires time and well-trained dermatologists to analyze the skin. Moreover, the amount of experience the team of dermatologists has plays a vital role in diagnostic accuracy. Computer aided diagnosis (CAD) may become a new way to handle these challenges and enhance dermatologic diagnosis.

In medical imaging and diagnostic radiology, a computer-aided diagnosis (CAD) is fast becoming a key tool. Recent developments in the field of computer vision have led to a renewed interest in computer applications in biology and medicine. A visual examination of a skin lesion is one of many potential applications of CAD. To fully utilize the potential benefits of a computer-vision-based diagnostic system, CAD requires a classification model that can rapidly and realistically process different images of clinical conditions. Moreover, a feature selection technique is one of the major factors that affect the accuracy of the machine learning classification.

The new emerging area of artificial intelligence in the field of deep learning has enabled machines to automatically determine features-of-features needed for data classification. The deep convolutional neural network (CNN) has led to a proliferation of studies in Computer-Aided Systems [8-10]. Several studies compared the accuracy of deep learning convolutional neural networks to dermatologists when classifying images of skin lesions. The main differences between these studies were either the classification model type or the number of labeled images used for training.

Esteva et al. [11] used 129,450 images to train GoogleNet Inception v3 CNN architecture [12]. The study was validated using 3-label and 9-label classification skin disease partition. The CNN achieved 72.1 ± 0.9% and 55.4 ± 1.7% overall accuracy when the algorithm was validated using a three-class and nine class respectively. In both cases, the deep learning CNN outperformed the dermatologists. The dermatologists achieved lower overall accuracy at 65.56% and 55.0% at three-class and 9 class respectively.

Nylund [13] presented a similar study. He used around 5000 images to train AlexNet [14], a common convolutional neural network for image classification. The study was validated using 23-labeled classification skin disease partition. The overall accuracy was 55.0%.

Ridell et al. [15] performed a binary classification between benign nevus and malignant melanoma. They used 200 and 1600 number images to train Google Inception v3 which resulted in an accuracy of 70.8% and 77.5% respectively. The study concludes that the size of the training data sets has a great impact on the accuracy of the deep learning model. Although there are several studies investigate the accuracy of different types of convolutional neural networks for skin lesion classification, far too little attention has been paid to the impact of image pre-processing and tuning on the accuracy of the network.

This paper combines the use of dermatoscopic images pre-processing and tuning and deep convolutional network. In addition, recent work suggested that a deeper convolutional network achieves better performance with a better efficiency to train [16, 17]. In order to provide such a classification model, this paper uses a classification algorithm that was developed and trained based on Densely Connected Convolutional Networks (DenseNets) [16, 17]. Initially, the paper examined the impact of proposed image pre-processing and tuning on the accuracy of a binary classification CNN. In the second stage, two different data-sets were created. The first data set included three different classes for a 3-
ary classification study. The second data set included nine classes for a 9-ary classification study. Experimental studies were conducted using more than 30,000 dermatoscopic images from multiple open-access archives \cite{18-29}. In order to evaluate the impact of both the deeper network and image pre-processing, the classification accuracy was measured with and without image pre-processing. In addition, we conduct evaluations with the state of the art. When comparing the results with those of older studies, it must be pointed out that higher accuracy was achieved.

The paper is organized as follows: Section 2.1 describes the data set used in this paper followed by Section 2.2, which introduces the data preparation and the use of dermatoscopic images pre-processing and tuning. Section 2.4 describes the used CNN architecture followed by the training and evaluation methods Section 2.5 and 2.6 respectively. Section 3 presents the results. Finally, Sections 3 and 4 present the discussion and conclusions of this work respectively.

2. Methodology

2.1. Dataset

In this study, more than 30,000 dermatoscopic images from multiple open-access archives were collected \cite{18-29}. A python program was used to access the files and to collect the content. We iterated through each file and automatically extracted and stored the source image and the corresponding metadata. The diagnoses for the dermatoscopic images were extracted from image metadata.

The data contains images of multiple classes of skin conditions, such as nevus, seborrheic keratosis (SK), Solar lentigo, basal cell carcinoma (BCC), and psoriasis. Each open-access archive contained dermatoscopic images reviewed by experienced dermatologists and the resulting diagnoses were annotated and each image was labeled. Images which were labeled as unknown or other were not used in this study.

2.2. Data Pre-processing

Since the data were collected from multiple sources, preprocessing stage was required to prepare the data. The preprocessing stage is one of the fundamental factors that could affect the performance of the classification model. Noisy data, missing information, duplication or any outliers in the training data set can result in an inaccurate training and unreliable result. Therefore, a data pre-processing was used to convert the raw data into a clean data set. The aim is to build a data with the most important features in order to simplify the model and classify more precisely.

The database contained images with illumination variation. This variation in the illumination create an irrelevant variation in the images classification levels to address this, images were undergone two processing steps.

Firstly, Images were normalized, to represent pixels in the range 0 to 1, and resized into 224 × 224 pixels. The purpose of image normalization is to increase contrast for improved feature extraction. Secondly, the contrast limited adaptive histogram equalization (CLAHE) filtering algorithm was used for enhancing the local contrast of the images \cite{30}. This was performed by converting the RGB image into the L*a*b* color space. The L*a*b* color space is derived from the CIE, International Commission on Illumination, XYZ tristimulus values \cite{30}. The coordinate L* defines a luminosity axis, a* chromaticity layer defines a redness greenness axis, and a chromaticity layer b* defines a blue-yellow axis. Then, CLAHE was adopted for the enhancement of the L component in the L*a*b* color space. The L* component image
was divided into several non-overlapping tiles, 88 tiles, of almost equal sizes. Then, a histogram of each tile is calculated and redistributed so its height does not exceed a specific limit for contrast expansion. The aftermath histograms cumulative distribution functions (CDF) [31] are determined for grayscale mapping. Bilinear interpolation then used to combine neighboring tiles in order to eliminate artificially induced boundaries.

2.3. Data Distribution and Partitioning

The dermoscopic images were distributed according to their classes. Three different data-sets were created. The first data-set included a binary labeled data (melanoma vs benign) for binary model classification evaluation experiment. The second data set included three different classes for a 3-ary classification study. The third data set included nine classes for a 9-ary classification study. Each class contained 1000 images. Figure 1 shows the first three levels of the taxonomy used in this study. The goal is to build up a deep learning model to classify multi-category dermoscopic images. The rules of classification and the number of classes was based on the disease taxonomy proposed by Andre et al. [11].

It is common that training a CNN with a balanced data-set will increase the overall performance of the network, compared to CNN trained with imbalanced data-set [32]. Although many studies provide a thorough review of learning from an imbalanced data, this was beyond the scope of this paper.

The collected data-set was used to build subsets with balanced distributions of data between the classes, and these subsets were used to train a CNN. The number of dermatoscopic images used in this study is presented in Table 1. All labeled dermatoscopic images were randomly separated into 80% training, 10% validation, and 10% testing sets.

2.4. CNN Architecture

A deep learning classification model has two phases: Training and testing. Extracting features from images is a common process in both phases. The training phase is characterized by the extraction of features using a large number of images as training data. The training data are used together with the correct/expected output. The testing phase is characterized by the extraction of features from previously unseen images (testing data). Figure 2 represents the block diagram for the training and testing phases of the deep-learning model.

Recent work suggests that a deeper convolutional network can achieve better performance by learning more complex features of the input data [16, 17]. However, adding more layers may lead to a negative effect and degrade the optimum performance of the model. The proposed algorithm is based on the Densely Connected Convolutional Networks (DenseNets) code package [16]. DenseNets connects each layer feature map to every successive layer in a feed-forward fashion. This allows the propagation of the feature maps of all prior layers to the input of subsequent layers, thus eliminating the degradation problem of deeper convolutional networks and allowing deeper networks to converge at optimum performance.

In this work, a DenseNets structure with four dense blocks on 224 × 224 input images was used as illustrated in Fig. 3. The model was made up of four dense blocks with a total of 58 layers. Each layer contained a convolutional filter, a rectified linear unit (ReLU) activation, and batch normalization (BN) layers. The convolution and pooling layers are defined as transition layers. Each transition layer contains a batch normalization
layer and a convolutional layer followed by an average pooling layer. The exact network architecture is shown in Table 2.

Fig. 1. The first three levels of the taxonomy used in this study. The dermoscopic images were distributed according to their classes. Three different datasets were created for 2-ary a 3-ary and 9-ary classification studies.

Table 1. Datasets used in this study.

<table>
<thead>
<tr>
<th>Dataset Type</th>
<th>Number of Classes</th>
<th>Number of Images</th>
<th>Number of Images</th>
<th>Number of Images</th>
<th>Number of Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-ary classification</td>
<td>2</td>
<td>2500</td>
<td>2000</td>
<td>250</td>
<td>250</td>
</tr>
<tr>
<td>3-ary classification</td>
<td>3</td>
<td>3000</td>
<td>2400</td>
<td>300</td>
<td>300</td>
</tr>
<tr>
<td>9-ary classification</td>
<td>9</td>
<td>9000</td>
<td>7200</td>
<td>900</td>
<td>900</td>
</tr>
</tbody>
</table>

Fig. 2. The block diagram for the training and testing phases in the deep-learning model.
Fig. 3. Schematic representation of a deep DenseNet with four dense blocks. The transition layers represent the layers between two adjacent blocks.

Table 2. Densely Connected Convolutional Networks (DenseNets) architecture.

<table>
<thead>
<tr>
<th>Layers</th>
<th>Output Size</th>
<th>DenseNet-121</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convolution</td>
<td>112 × 112</td>
<td>7 × 7 conv, stride 2</td>
</tr>
<tr>
<td>Pooling</td>
<td>56 × 56</td>
<td>3 × 3 max pool, stride 2</td>
</tr>
<tr>
<td>Dense Block</td>
<td>56 × 56</td>
<td>(\begin{array}{c}1 \times 1 conv \ 3 \times 3 conv\end{array}) × 6</td>
</tr>
<tr>
<td>Transition Layer</td>
<td>56 × 56</td>
<td>1 × 1 conv</td>
</tr>
<tr>
<td>Dense Block</td>
<td>28 × 28</td>
<td>(\begin{array}{c}1 \times 1 conv \ 3 \times 3 conv\end{array}) × 12</td>
</tr>
<tr>
<td>Transition Layer</td>
<td>28 × 28</td>
<td>1 × 1 conv</td>
</tr>
<tr>
<td>Dense Block</td>
<td>14 × 14</td>
<td>(\begin{array}{c}1 \times 1 conv \ 3 \times 3 conv\end{array}) × 24</td>
</tr>
<tr>
<td>Transition Layer</td>
<td>14 × 14</td>
<td>1 × 1 conv</td>
</tr>
<tr>
<td>Dense Block</td>
<td>7 × 7</td>
<td>(\begin{array}{c}1 \times 1 conv \ 3 \times 3 conv\end{array}) × 16</td>
</tr>
<tr>
<td>Classification Layer</td>
<td>1 × 1</td>
<td>7 × 7 global average pool fully-connected, softmax</td>
</tr>
</tbody>
</table>

2.5. Training

The training was undertaken by carefully controlling the learning rate using a stochastic descent algorithm [33]. In this study, to optimize the training time, an initial training rate was set at 0.0001 at the beginning of training and gradually reduced every 5 epochs by a factor of 0.2 with 90 epochs, and mini-batches of 64 observations were used at each iteration. As the training progressed, this gradually changes enabled smaller a finer search towards the optimum value.

2.6. Evaluation Metrics

As mention earlier, all labeled dermoscopic classes were randomly split into 80% training, 10% validation, and 10% testing subsets. Then the model was trained for the evaluation of the CNN model. This iterate 10 with a different subset reserved for evaluation purposes each time.

To compare the performance of the classification model, the accuracy of the data set was measured. It is the most common evaluation metric for classification problems. It is a reliable measure when the target variable classes in the data are nearly balanced. It is the ratio of the number of correct predictions made to all predictions made.

In addition, sensitivity and specificity measure was used. The sensitivity of a clinical
classification model can be defined as the ability of the classifier to correctly identify the disease. While the specificity measures the ability of the classification model to correctly distinguish the healthy data. For the binary classification problems, Area under ROC, Receiver Operating Characteristics, Curve or AUC was used also as a performance metric.

The AUC represents a model’s ability to distinguish between positive and negative classes, in this study between melanoma and benign. A perfect model, a model that made all predictions correct, will have an area of 1. The AUC of 0.5 indicates a classifier with no power, i.e. random predication.

3. Results and Discussion

3.1. Dermoscopic Image Processing

As mentioned earlier, Images were normalized to represent pixels in the range 0 to 1. Then, the contrast limited adaptive histogram equalization (CLAHE) filtering algorithm was used for enhancing the local contrast of the images. This was performed by converting the RGB image into the L*a*b* color space and perform CLAHE on the L channel. Finally, converting the resulting image back into the RGB color space. Results from this preprocessing step are presented in Fig. 4.

3.2. Binary Model Classification

To examine the strengths and weaknesses of the classification model after the CLAHE application, DenseNets was trained on a binary-labeled (melanoma vs benign). This implementation was performed using two sets of training data. Firstly, the model was trained with the data before CLAHE application. In the second experiment, the contrast limited adaptive histogram equalization (CLAHE) filtering algorithm was used for enhancing the local contrast of the images. Then, the model was retrained using binary-labeled data after the CLAHE application. The purpose of these two experiments was to explore the advantage of the proposed image preprocessing technique.

The accuracy of both binary classifiers was measured by the area under the ROC curve, see Fig. 5. The figure shows that the binary classifier achieves a higher AUC, 0.93 when trained on the data after CLAHE application when compared to 0.85 before. The sensitivity and specificity improved from 0.82 and 0.74 to 0.91 and 0.86 respectively.

3.3. Multi-class Classification

As mentioned in the previous section, performing the CLAHE filtering algorithm improves the binary classifier performance. In this section, to examine a multiclass Classifier performance, the DenseNets were trained on two different multi-labeled data (3-labeled and 9-labeled). Two studies were performed. Similar to the binary classifier study, both types of training data sets were used in this study, before and after performing CLAHE filtering.

Firstly, DenseNets was trained on three-labeled data (benign, malignant and non-neoplastic skin lesions). In the second study, the classifier was trained on nine-labeled data, see Section 2.1 for details. The performance of the DenseNets algorithm was validated in the 3-ary and 9-ary classifications. Figure 6 shows the multi-class Classification accuracies for 3-ary and 9-ary classifiers.

After following the pre-processing steps, the 3-ary classifier accuracy improved from 72.3±0.9% to 81.2±1.1%. The accuracy of the 9-ary classifier before and after preprocessing steps was 56.6±1.9% and 60.1±1.3% respectively. Given the interesting measurement increase in both multiclass classification accuracies, it is evident that preprocessing steps has a promising impact on the model performance.
Two of the published works in skin lesion classification performed multi-class classification Boman and Volminger [34] and Esteva et al. [11]. The results, as shown in Table 3, indicate that in the 3-ary and 9-ary classifications. The validation results showed that the proposed work has a higher validation accuracy of \( 81.2 \pm 1.1\% \) in the 3-way classification when compared to \( 72.1 \pm 0.9 \) \([11]\) and \( 68.30\% \) \([34]\). In the 9-way classification, the proposed work achieved an accuracy of \( 60.1 \pm 1.3\% \). This is higher than what was achieved in Esteva et al. [11], \( 55.4 \pm 1.7\% \).

![Fig. 4. Shown are fundoscopic illustrations before and after the CLAHE application.](image)

![Fig. 5. ROC curves for the model examined on two separate training data sets.](image)

![Fig. 6. Multi-class Classification accuracies for 3-ary and 9-ary classifiers.](image)
Table 3. Comparison of the accuracy of different studies.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Number of classes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Esteva et al. [11]</td>
<td>72.1±0.9%</td>
</tr>
<tr>
<td>BOMAN and Volminger [34]</td>
<td>68.30%</td>
</tr>
<tr>
<td>This work</td>
<td>78.3±0.4%</td>
</tr>
</tbody>
</table>

The proposed combination of dermatoscopic images pre-processing and the deeper convolutional network was capable of achieving better performance by learning more complex features of the input data with a more efficient memory implementation.

DenseNet connects each layer feature map to every successive layer in a feed-forward fashion. This will allow the propagation of the higher-order feature to the subsequent layers, thus allowing DenseNet to converge at optimum performance.

4. Conclusion

This work combines the use of dermatoscopic images pre-processing and deep convolutional network to design a skin lesion classification algorithm. Experimental studies were conducted using more than 30,000 dermatoscopic images from multiple open-access archives. The collected data were passed through several steps of image processing and cleaning to convert the raw data into a clean data set. Firstly, Images were normalized and resized. Then, a contrast limited adaptive histogram equalization (CLAHE) filtering algorithm was used for enhancing the local contrast of the images. The resultant data were used to train Densely Connected Convolutional Networks (DenseNets). DenseNets allows the propagation of the feature maps of all prior layers to the input of subsequent layers, thus eliminating the degradation problem of deeper convolutional networks and allowing deeper networks to converge at optimum performance.

Initially, the paper examined the impact of proposed image pre-processing and tuning on the accuracy of a binary classification CNN. Both processed data and unprocessed data were used to train the network in two separate experiments. The accuracy of both binary classifiers was measured by the area under the ROC curve. The result showed that the binary classifier achieves a higher AUC, 0.93, when trained on processed data compared to 0.85 before. The sensitivity and specificity improved from 0.82 and 0.74 to 0.91 and 0.86 respectively.

In the second stage, two different data-sets were created. The first data set included three different classes for a 3-ary classification study. The second data set included nine classes for a 9-ary classification study. The validation results showed that the proposed work has higher validation accuracy of 81.2 ± 1.1% in the 3-way classification when compared to72.1 ± 0.9 [11] and 68.30% [34]. In the 9-way classification, the proposed work achieved an accuracy of 60.1 ± 1.3%. This is higher than what was achieved in Esteva et al. [11], (55.4 ± 1.7%).

The proposed combination of dermatoscopic images pre-processing and deeper convolutional networks can achieve better performance by learning more complex features of the input data with a more efficient memory implementation. Considerably more
work will need to be done to evaluate the performance of the model in a real clinical environment.

5. Compliance with Ethical Standards

5.1. Conflict of Interest

M. Alnowami declares that he has no conflict of interest.

5.2. Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

References


الشبكة العصبية الملتفة العميقة جدًا لتصنيف الأورام الجلدية

مجدي راشد سالم النويامي
قسم الهندسة النووية، كلية الهندسة، جامعة الملك عبد العزيز، 2004، المملكة العربية السعودية
malnowaimi@kau.edu.sa

المستخلص. في التصوير الطبي والأشعة التشخيصية، بسرعة أصبح التشخيص بمساعدة الحاسوب آداة رئيسية، وبعد الفحص البصري للأمراض الجلدية أحد التطبيقات المحتملة. في هذه الورقة، تم تطوير خوارزمية تصنيف وتدريبها استنادًا إلى التعليم العمق والشبكات الممتفقة المرتبطة بكثافة (DenseNets) لتصنيف الأمراض أو الأورام الجلدية. ويجمع هذا العمل بين استخدام الصور التنظيرية للمختمف الأورام الجلدية، وأجريت الدراسة باستخدام أكثر من 3000 صورة لمختلف الأورام الجلدية، وتم الحصول عليها من مصادر مختلفة. في البداية، بحثت الورقة تأثير المعالجة المسبقة للصورة وضبطها على دقة التصنيف الثنائي. التعليم العمق أظهر أن المصنف الثنائي يحقق مساحة أقل تحت المنحنى (AUC)، عند التدريب على البيانات المعالجة مقارنة بـ 0.85 دون معالجة الصورة. في المرحلة الثانية، تم إنشاء مجموعتين مختلفتين من البيانات من 3 فئات و 9 فئات، حيث إن كل فئة تميز لمرض معين أو ورم ذي تصنيف مختلف. وأظهرت نتائج التحقق من صحة كون العمل المفترض حقق دقة التحقق من 0.81 ± 0.1% في التصنف الثلاثي و 0.6 ± 0.3% في حالة تصنيف نسعة أمراض. ويمكن للمزيج المقترح من الصور الجلدية والشبكة التلافيفية الأممية تحقيق أداء أفضل من خلال تعلم مؤثرات أكثر تعقيدًا لتصنيف الأمراض، مع ذاكرة أكثر كفاءة.

كلمات مفتاحية: التصنيف، الجلد، الأورام، التشخيص بمساعدة الحاسوب، الشبكة العصبية.