How Users Perceive Content-Based Image Retrieval for Identifying Skin Images

Mahya Sadeghi[1(✉)], Parmit K. Chilana[1], and M. Stella Atkins[1,2]

[1] School of Computing Science, Simon Fraser University, Burnaby, BC, Canada
mahyas@sfu.ca
[2] School of Dermatology and Skin Science, University of British Columbia, Vancouver, BC, Canada

Abstract. Content-Based Image Retrieval (CBIR) is an application of computer vision techniques for searching an existing database for visually similar entries to a specific query image. One application of CBIR in the dermatology domain is displaying a set of visually similar images with a pathology-confirmed diagnosis for a given query skin image. Recently, CBIR algorithms using machine learning with high accuracy rates have gained more attention since researchers have reported they have the potential to help physicians, patients, and other users make trustworthy and accurate classifications of skin diseases based on visually similar cases. However, we do not have many insights into how interactive CBIR tools are actually perceived by end users. We present the design and evaluation of a CBIR user interface and investigate users' classification accuracy on dermoscopy images and explore users' perception of confidence and trust. Our study with 16 novice users for a given set of annotated dermoscopy images indicates that, in general, CBIR enables novices to make a significantly more accurate classification on a new skin lesion image from four commonly-observed categories: Nevus, Seborrheic Keratosis, Basal Cell Carcinoma, and Malignant Melanoma.

Keywords: Dermatology · Skin cancer · CBIR · Machine learning
Artificial intelligence · Evaluation · Human-computer interaction

1 Introduction

Skin cancer is one of the most common cancers, and the number of skin related patient visits in primary care is considerable. Melanoma, the deadliest type of skin cancer, is curable if it is diagnosed early. Basal cell carcinoma, another type of skin cancer, also needs early detection to be properly treated. Considering significant number of dermatology related visits in clinics, supporting non expert physicians in their diagnostic decision can improve patient outcomes and at the same time can save costs for healthcare systems by reducing unnecessary referrals and providing early diagnosis. This can also lead to a better resource management where there is a limited access to specialists and support general

D. Stoyanov et al. (Eds.): MLCN 2018/DLF 2018/iMIMIC 2018, LNCS 11038, pp. 141–148, 2018.
https://doi.org/10.1007/978-3-030-02628-8_16

physicians as an educational tool. Recent advances in computer-aided diagnostic methods can aid self-examining approaches based on images, which can significantly improve early detection as the most important step to improve prognosis. In fact, modern machine learning classifiers are becoming increasingly capable of classifying skin cancer images with a level of competence comparable to dermatologists [1,2]. Although medical imaging diagnostics can benefit from intelligent computer vision and machine learning techniques, most AI algorithms provide a black box diagnosis based on percentages which clinicians do not trust [3] and most of the knowledge contained in visual data is barely extracted and applied to deliver an accurate diagnostic decision.

With recent advances in machine learning algorithms, there has been renewed interest in content-based image retrieval (CBIR) approaches where computer vision methods can be used to visually search for images to a "query" image in large databases based on the content of the image and visual clues such as shape, color, and pattern [4]. CBIR provides similar images where user can interpret the results and determine whether they are reliable. Furthermore, within the dermatology context, this technology is designed to assist with identifying and comparing skin lesions using percentage-based classifiers. CBIR-based tools can be a safe and effective implementation and integration of artificial intelligence and machine learning algorithms in clinical workflow to be validated in a low risk clinical setting. Modern CBIR systems offer powerful possibilities for lowering the overall search time and increase retrieval accuracy and are being used in a number of scientific endeavors [5,6]. Although designing and evaluating such systems in direct collaboration with users has received only limited attention, findings in a study on CT images suggest that when interpretation was supplemented with an image retrieval tool, diagnostic accuracy was improved [7]. Therefore, there are several open questions about how these tools can be safely integrated and accepted in real-world settings to support the diagnostic work of medical professionals.

In our research, we are examining how a CBIR decision support tool can be used by non-dermatologists in classification of dermoscopic skin lesion images. In this paper, we use an intuitive and scalable method on CBIR as an explainable artificial intelligence application, and investigate to what extent a CBIR system can help a non-dermatologist make an accurate classification of a given skin lesion image. We also explored to what extent the use of CBIR affects the confidence levels of these users. Our findings shed new insights into how user-centered design techniques can improve non-expert user interaction with CBIR systems and open up new opportunities for non-experts to explore, trust, and learn from medical image collections.

2 Method

2.1 Study Design

We used an experimental approach to answer our key research question: to what extent, does using a CBIR system affect user's ability to make a more accurate

classification on a new skin image? The key concepts we are using to answer this question are decision accuracy, confidence level, and user trust. Our study used a within-subjects design where all the participants went through the same tasks and questionnaires. The experiment consisted of two conditions (without CBIR and with CBIR). Each user was presented with the query images one at a time and was asked to choose one best category by clicking on the appropriate button. The same normal lighting condition with a large screen was provided for all users.

2.2 Dataset

All the images are from publicly available datasets, including The International Skin Imaging Collaboration (ISIC) archive [8] and a dermoscopy atlas [9]. Since the number of skin lesion classification categories is very large (over 100 commonly observed), we had to limit our study to 4 common skin lesion categories, similar to those used in the ISIC classification challenge 2017 [10] i.e. Nevus, Seborrheic Keratosis (SK), Basal Cell Carcinoma (BCC) and Malignant Melanoma (MM). All the images were approved by an expert dermatologist who had experience working with dermoscopic images. To simplify complex medical terms for general users, we used simple terminologies for each skin lesion category. From the 1021 images in our dataset, 40 query images were chosen: 20 without CBIR and 20 with CBIR. We selected 5 query images from each category for each condition to provide an equal disease distribution. Among the 20 images in both conditions, 4 of the images were repeated, one from each category so there were 36 unique query images.

2.3 System Description

We used an existing classifier and built a user interface on top. This system was designed as a decision support tool to read and retrieve all the similar images for each query image based on a list of classification probabilities from a classifier trained on the 4 classes of interest for each image. All the images were presented to the user based on a file that stored a dictionary where the key was the name of the query image and the value is a list of tuples (imagename, cosinedistance) of the top 20 closest images inclusive. The number represented the cosine distance between that image and the query image computed using the deep feature of the query image and the image being compared. All the retrieved similar images were sorted by their cosine distance in ascending order, so the first similar image was the most similar image to the query image based on our machine learning algorithm [11]. Figure 1 shows a screen capture of the interactive user interface with our CBIR system. During the CBIR condition, the 15 most visually similar images of the collection were returned for each query image, sorted from top left row to the bottom right row. Our user-interface software for the study was written using HTML, CSS, JavaScript, NodeJS and MongoDB.

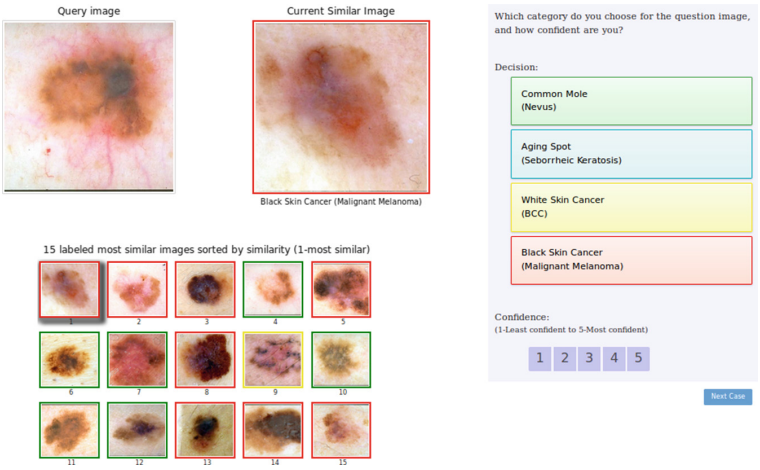

Fig. 1. Sample screenshot with CBIR algorithm results.

2.4 Protocol

We used the following protocol: After signing a consent form, participants were given a pre-task questionnaire about their past experience in medical image search. For the next 10 min, they went through a brief tutorial to learn about 4 different skin lesion categories (presented as educational slides). Next, participants started the study by classifying 20 query skin lesion images in the first condition, followed by classifying 20 query skin lesion images in the second condition. To reduce possible bias resulting from fatigue or learning effects, which are common in within-subject studies, each participant was randomized to start without CBIR or with CBIR condition. In addition, the order of "query" images was randomly selected by a shuffle algorithm inside the system, and was varied from user to user. Once the study ended, they were provided with total feedback on their performance. Finally, they filled out a post-task questionnaire about their experience.

2.5 Data Collection

We used multiple methodologies to gain insights from the different data types obtained in the study and recorded by the system. Qualitative data was obtained from interviews and questionnaire, and quantitative data such as decision accuracy and confidence level were recorded in a computer log captured in our decision support tool interface.

3 Results

16 participants successfully completed the lab experiment, including 10 males and 6 females, all non-expert adults (graduate students). From the pre-task

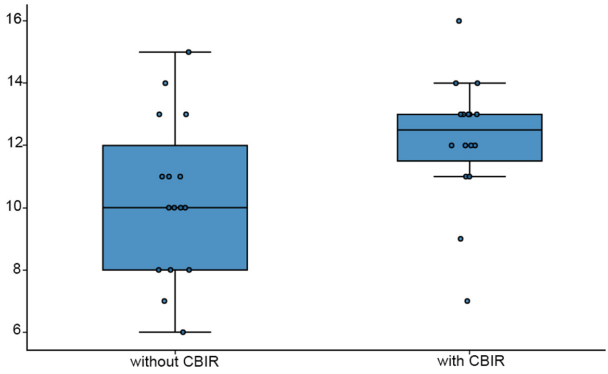

Fig. 2. Total accuracies in each condition (without CBIR and with CBIR) are shown. x axis represents the condition and y axis represents number of correct answers in each condition (out of 20). By incorporating CBIR, the mean accuracy is increased from 10.31(51.56%) to 12.19(60.94%)

questionnaire, we learned that 11 participants (68.57%) had experience in medical image search previously. They were mainly looking for photographs. The key motivation for them was personal-diagnosis (73%) and self-education (73%). However, most of the participants only found their previous searches somewhat useful and somewhat trustworthy. Irrelevant and untrustworthy images were stated as the major problems encountered during the search process.

Accuracy: For accuracy calculations, user decisions were compared to the diagnosis for every query. Overall, there was a significant improvement in mean classification accuracy from 51.56% (165 of 320) without CBIR to 60.94% (195 of 320) with CBIR as shown in Fig. 2. Corresponding null hypothesis was that there is no difference in the means, and difference in mean accuracy between conditions was tested by the two-tailed t test for paired samples. The improvement was greatest for the Nevus and MM categories, as shown in Table 1. For the Seborrheic Keratosis Category, although the accuracy decreased, no significant difference was found.

Confidence and Trust: To determine the change in users confidence in decisions without vs with CBIR, we used the Likert scale [12] score in scale of 1 (least confident) to 5 (most confident) for every query. Our null hypothesis was that there is no difference in the means. The difference in mean confidence between conditions was tested by the two-tailed t test for paired samples. The overall mean user confidence score was 3.47 without using CBIR and 3.7 with using CBIR ($P < 0.05$). Users mean confidence in TP cases was improved by 6.59% ($P < 0.05$) which shows showing similar cases is effectively increases users confidence. However when the classification result was incorrect, the impact of showing similar cases was not as significant in increasing users confidence, and was only increased by 2.52%. In addition, significant difference between confidence on correct classifications (78.16%) and incorrect classifications (69.74%)

Table 1. TP (True Positives) and Percentage of Correct classifications With and Without CBIR. Significant results where $P < 0.05$ (paired two-sided t test) are shown in bold.

Skin lesion category	Total correct classifications without CBIR (N = 80)	Total correct classifications with CBIR (N = 80)
Nevus	**50(62.5%)**	**72(90%)**
Seborrheic Keratosis	29(36.25%)	19(23.75%)
Basal Cell Carcinoma	49(61.25%)	57(71.25%)
Malignant Melanoma	**37(41.25%)**	**49(61.25%)**
Total	**165(51.56%)**	**195(60.94%)**

using CBIR was found ($P < 0.05$). Table 2 reveals mean confidence level with and without using CBIR, as well as standard deviation errors in parentheses. Trust as another critical factor was also measured on the Likert scale score in scale of 1 (least confident) to 5 (most confident) in pre-task and post-task questionnaire. Our null hypothesis was that there is no difference in the means, and the difference was tested by the two-tailed t test for paired samples. 11 of the users had previous experience with medical image search and reported a mean of 54.5% (SD = 0.98) trust on their previous findings. After the study, these users self-reported a mean of 59.29% (SD = 1.08) trust to the CBIR results; however, the difference was not significant ($p = 0.65$).

Table 2. Confidence level and SD of classifications With and Without CBIR. Significant results where $P < 0.05$ (paired two-sided t test) are shown in bold.

Classification	Average confidence without CBIR	Average confidence with CBIR
Correct	**71.57%(0.66)**	**78.16%(0.52)**
Incorrect	67.22%(0.61)	69.74%(0.48)
Total	**69.4%(0.63)**	**74%(0.54)**

4 Discussion

Although role of AI image classifiers in medicine are undeniably positive, their inner structures are often hard to comprehend and they are not usually used in the real-world settings. CBIR decision support tools can be seen as transparent applications of AI and are likely to play a growing role in the clinical practice of dermatology since this field heavily relies on the training level and expertise of medical professionals in visual inspection of skin diseases. In our user-centered design approach, we tried to tackle the problem of skin lesion classification and users' perceptions in using CBIR. Our initial results indicate that CBIR can indeed be effective for users based on the number of correct classifications they made and the increase in their confidence levels when using a CBIR interface.

According to the data collected in our study, applying CBIR models that deliver most visually similar images within the decision support tool will help users in decision making process where the final decision can be left to discretion of the user. It is noteworthy that users' accuracy scores on SK images actually decreased. Although it's not a significant difference, this may be related to the limited number of SK images in the dataset which resulted in fewer similar images from the SK category. We are currently limited to small and public datasets that often have low quality images; however, as the database for such systems grows, system accuracy is likely to increase. Our findings indicate that there should be enough representation of different disease in dataset for CBIR systems, regardless of malignancy status, when all diseases have an equal distribution balance. Other major decision making challenges for users are imperfect accuracy rates of algorithms, quality of the images(such as contrast, lighting, size), external objects in the images(such as ruler and hair), inconsistency in force and tilt while placing the dermoscopy device on the skin [13], and insufficiently magnified images.

Our findings also demonstrate that users confidence level with seeing similar images significantly increased. Hence, patient safety needs to be addressed in real clinical settings, and we need to investigate how primary physicians can adopt CBIR in clinical setting safely for better patient care outcome and more efficient workflow. Trust as another critical factor was measured in pre-task and post-task questionnaire. According to the data collected in our study, although trust is increased because of similar results, it's not a significant value. One reason may be due to the novelty of the system. Medical tools need to have long term impact, and trust can be increased overtime with personal experience, scientific evaluation, and publications. Another reason may be related to user expertise level and lack of medical knowledge.

In this study, we were limited to non-expert users as proxies for population of medical students, general physicians, and expert dermatologists. Primary care physicians have limited training in dermatology and in most cases no training in dermoscopy which is standard of diagnosis and management for skin cancer pigmented and vascular lesions. In our study we focused on novices to understand the implications of offering interactive CBIR tools to investigate their classification accuracy. According to our results, we believe this system can help users both in image interpretation and as an educational tool, since the user will be able to view pre-diagnosed similar images.

In future work, we will consider whether the results from novices transfer to other user groups. Initial informal feedback from general physicians shows their knowledge of dermoscopic skin lesions is as limited as the novices we tested, and we plan to perform user studies with experts and with physicians in future to confirm these findings. For establishing an effective interaction between a CBIR system and users, it is key to know how CBIR tools can be safely and effectively implemented, integrated, and customized for people with different levels of expertise.

References

1. Esteva, A., et al.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115 (2017)
2. Han, S.S., Kim, M.S., Lim, W., Park, G.H., Park, I., Chang, S.E.: Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. J. Invest. Dermatol. **138**, 1529–1538 (2018)
3. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144. ACM (2016)
4. Estrela, V.V., Herrmann, A.E.: Content-based image retrieval (CBIR) in remote clinical diagnosis and healthcare. In: Encyclopedia of E-Health and Telemedicine, pp. 495–520. IGI Global (2016)
5. Dhara, A.K., Mukhopadhyay, S., Dutta, A., Garg, M., Khandelwal, N.: Content-based image retrieval system for pulmonary nodules: assisting radiologists in self-learning and diagnosis of lung cancer. J. Digit. Imaging **30**(1), 63–77 (2017)
6. Benam, A., Drew, M.S., Atkins, M.S.: A CBIR system for locating and retrieving pigment network in dermoscopy images using dermoscopy interest point detection. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), pp. 122–125. IEEE, Location (2017)
7. Aisen, A.M., et al.: Automated storage and retrieval of thin-section CT images to assist diagnosis: system description and preliminary assessment. Radiology **228**(1), 265–270 (2003)
8. The International Skin Imaging Collaboration (ISIC) Archive. https://isic-archive.com/. Accessed 10 Dec 2017
9. Argenziano, G., Soyer, H.P., De Giorgi, V., Piccolo, D., Carli, P., Delfino, M.: Interactive Atlas of Dermoscopy (Book and CD-ROM). EDRA Medical Publishing & New Media, Milan (2000)
10. Codella, N.C.F., et al.: Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 168–172. IEEE (2018)
11. Tschandl, P., Argenziano, G., Razmara, M., Yap, J.: Diagnostic accuracy of content based dermatoscopic image retrieval with deep classification features, 28 p. (2018, under review)
12. Likert, R.: A Technique for the Measurement of Attitudes. Archives of Psychology. The Science Press, New York (1932)
13. Dreiseitl, S., Binder, M., Vinterbo, S., Kittler, H.: Applying a decision support system in clinical practice: results from melanoma diagnosis. In: AMIA Annual Symposium Proceedings, p. 191. American Medical Informatics Association (2007)