Original paper

# Evaluation of a smartphone application for diagnosis of skin diseases

**Maksym Mikołajczyk[1], Sebastian Patrzyk[2], Mariusz Nieniewski[3], Anna Woźniacka[2]**

[1]Student Research Circle at the Department of Dermatology and Venereology, Medical University of Lodz, Lodz, Poland
[2]Department of Dermatology and Venereology, Medical University of Lodz, Lodz, Poland
[3]Faculty of Mathematics and Informatics, University of Lodz, Lodz, Poland

Adv Dermatol Allergol 2021; XXXVIII (5): 761–766
DOI: https://doi.org/10.5114/ada.2020.101258

## Abstract

**Introduction:** Artificial intelligence (AI) could offer equal, or even more accurate, diagnoses of melanoma than most dermatologists. However, the value of popular smartphone applications for diagnosing unpigmented skin lesions remains unclear.

**Aim:** To compare the diagnostic accuracy of a popular, free-to-use web application for automatic dermatosis diagnosis against expert diagnosis of selected skin diseases.

**Material and methods:** Skin lesion images of patients with verified diagnosis were collected using a smartphone and were diagnosed by the application. The AI provided five diagnoses of varying probability. For each patient, accuracy of the diagnosis was evaluated by three criteria, i.e. whether the expert diagnosis was matched by the most probable automated diagnosis, one of the top three diagnoses or one of the top five diagnoses. Reliability was analysed using intraclass correlation coefficients.

**Results:** The chance of a correct diagnosis increased when more outcomes were considered and more samples of a skin condition were included. However, the probability of a diagnosis repeating for the same patient was below 25%. Reliability, sensitivity and specificity were insufficient for clinical purposes.

**Conclusions:** Although AI diagnostics are encouraging, there is also a large margin for improvement, and AI is not yet an adequate replacement for medical professionals.

**Key words:** artificial intelligence, smartphone application, web application, skin diseases diagnosis, psoriasis, new technology.

## Introduction

Learning the correct diagnosis of numerous skin diseases takes years of medical training. Even then, diagnostics are often a difficult, time-consuming procedure. In many fields, including dermatology, the demand for experts, especially in rural and remote areas, far exceeds the available supply. However, diagnostics have recently become cheaper and more accessible thanks to huge advances in machine learning, particularly deep learning algorithms, which have shown great progress in automatically diagnosing diseases [1].

Dermatology is a visual specialty and clinical imaging is now considered to be an essential part of documentation and follow up. Nowadays there are more than 800 mobile dermatology applications in use by patients, medical students and physicians [2]. Unfortunately, their validity and reliability is not generally established and patients are not usually able to critically evaluate their medical worth.

It has been proposed that some artificial intelligence (AI) algorithms could be equal or even better at diagnosing skin cancers, mainly melanoma, than qualified dermatologists [1]. However, no comprehensive evaluation of the ability of popular smartphone applications to diagnose unpigmented skin lesions has yet been performed.

## Aim

The aim of this study is to compare the diagnostic accuracy of a popular, free-to-use web application for diagnosing dermatoses, with that of verified, expert di-

**Address for correspondence:** Maksym Mikołajczyk, Department of Dermatology and Venereology, Medical University of Lodz, pl. Hallera 1, 90-647 Lodz, Poland, phone: +42 6393092, fax: +42 6884565, e-mail: maksym.m@onet.eu
**Received:** 31.03.2020, **accepted:** 19.04.2020.

agnosis of selected skin diseases. The name of the web application is known to the editors and not mentioned in the article on purpose as the aim is not to fully assess a given product.

## Material and methods

The study group consisted of 217 patients aged 18 or above (110 men and 107 women), all of whom had been treated at the Department of Dermatology and Venereology, Medical University of Lodz, in 2018–2019. All were volunteers and their disease was confirmed by at least two certified dermatologists as well as by histopathological examination. One hundred patients from the study group suffered from psoriasis. The protocol of the study was approved by the Bioethics Committee of the Medical University (RNN/186/17/KE). Written informed consent to take part was obtained from all subjects.

### Study protocol

The study was performed in two stages: good quality smartphone images of selected well-documented skin diseases were collected in the first stage, and then were classified by a free-to-use web application in the second stage. Each patient was asked to present their full body. An exhaustive clinical evaluation was conducted, and three distinct regions of each skin lesion were selected for further evaluation. The described web application uses two images to identify a particular anomaly: a full view of the lesion and a close-up image. Therefore, six images were taken for each patient, i.e. three pairs of pictures. In the present study, each pair of images are classified as a single sample. Patient anonymity was maintained throughout the image acquisition process. Formal consent was collected from each patient before data were gathered.

In line with the guidelines provided by the web application developers, a 5 Mpixel smartphone camera was used, natural lighting was provided, and camera focusing was maintained. Before uploading into the application, the images were checked and irrelevant artefacts, such as jewellery or tattoos, were removed.

When a sample was submitted to the web application, the user was provided with five diagnostic outcomes, with the certainty of each diagnosis specified as a percentage. For example, the outcome could be presented as follows: psoriasis 36%; eczema 23%; acne vulgaris 11%; unspecified dermatitis 11%; urticaria 3%. These output data were saved, coded, and collected.

It was found that some diseases were not recognized by the web application; therefore changes to the dataset had to be made. Although it is stated that the algorithm can classify 33 skin conditions, no such list of conditions is available. Only 27 possible conditions could be identified in our group of patients by the application. Therefore, a new database was formed by removing patients with skin conditions that were non-classifiable by the application and more data of patients with classifiable conditions were collected. The final study group consisted of a total of 150 patients (900 images), among which 100 patients (600 images) were diagnosed with psoriasis.

### Data analysis

For each patient the automated diagnosis was evaluated based on three options: whether the verified diagnosis was matched by the most probable outcome from the web application, whether it was matched by one of the three most probable outcomes from the application, or whether it was matched by one of the five most probable outcomes.

A two-way model of intraclass correlation coefficients (ICCs) with 95% confidence intervals (95% CI) was used as a measure of precision or *relative reliability* (following the definition of the International Vocabulary of Metrology (2008): "closeness of agreement between indications or measured quantity values obtained by replicate measurements on the same or similar objects under specified conditions") [3, 4]. The results of the ICCs were analysed according to two interpretations: the first suggests results from 0.00 to 0.39 as a poor correlation, 0.40 to 0.59 as a fair correlation, 0.60 to 0.75 as a good correlation and 0.75 to 1.00 as an excellent correlation [5], while the second, more clinically-oriented criterion suggests values below 0.75 to be a poor to moderate correlation, 0.75 to 0.87 as a good correlation and values above 0.87 as clinical measures [6].

As a number of patients were diagnosed with psoriasis, this merited particular attention in the analysis. It was assumed that positive (P) denotes the number of verified psoriasis cases, and negative (N) denotes the number of cases of any other disease as given by a human expert. True positive (TP) denotes the number of cases correctly identified as psoriasis by the application, i.e. matching the human diagnosis, while true negative (TN) denotes the number of cases correctly classified as non-psoriasis. False negative (FN) denotes the number of incorrect negative diagnoses by the application, and false positive (FP) denotes the number of incorrect positive diagnoses.

Sensitivity and specificity for psoriasis diagnosis were analysed. Each sample, i.e. pair of images, was considered independently to simulate authentic patient interaction with the application, assuming no repeated tests. In this case, sensitivity was calculated as the ratio of correct diagnoses of psoriasis by the application to the total number of psoriasis diagnoses by medical experts, or TP/(TP + FN) = TP/P. The specificity is the ratio of correct diagnoses of non-psoriasis by the application to the total number of diagnoses of non-psoriasis by medical experts, or TN/(TN + FP) = TN/N.

### Statistical analysis

The described calculations were performed using Statistica ver. 13 (Dell Inc., Round Rock, TX) and Microsoft Excel 2013 (Microsoft Corp., Redmond, WA). *P*-value = 0.05 was considered as the threshold of statistical significance.

## Results

For the first variant, i.e. the first option given by the application matching the verified diagnosis, a correct diagnosis was achieved for 5.26% of all patients and for 4.44% of those with psoriasis. The percentage of correct diagnoses rose with each additional sample provided for a given patient, increasing to 10.33% for all patients, and 8.67% for those with psoriasis for two samples, to 20.67% for all patients and 26.00% for those with psoriasis when three samples were used (Table 1).

The percentage of correct diagnoses was also evaluated in the second variant, i.e. one of the top three most probable diagnoses generated by the web application matched the verified diagnosis. A correct diagnosis was obtained for 12.89% of all patients and for 9.41% of psoriatic patients when one sample was provided, 25.78% and 18.78% for two samples, and 51.56% and 56.33% for three samples (Table 1).

Regarding the third variant, where one of any five options matched the human diagnosis, the percentage of correct diagnoses also rose with each additional sample. For all patients, the percentage rose from 19.26% through 38.89% up to 77.78% as further samples were provided, while for psoriatic patients, the percentage rose from 13.85% through 28.22% up to 84.67% (Table 1).

The acquired results were also analysed in a cumulative approach, with regard to the number of times they repeated during successive attempts using different samples (i.e. pairs of images) for each patient.

The application provided the correct diagnosis as most probable in two out of three attempts for 9.33% of all patients and for 10.00% of psoriatic patients. Identical, correct diagnosis was presented three times in three tests for 3.33% of all patients and 5.00% of psoriatic patients (Table 2).

The correct diagnosis was given in the top three possible options for 20.67% of all patients and 25.00% of psoriatic patients when considering two out of three attempts. These values fell to 16.67% of all patients and 17.00% of psoriatic patients when all three attempts were considered (Table 2).

Finally, when all five possible options are analysed, 20.67% of all patients and 23.00% of psoriatic patients were presented with two correct diagnoses out of three tests. These values increased to 36.67% of all patients and 40.00% of psoriatic patients when the correct diagnosis was observed three times in three tests (Table 2).

It is worth noting that all the triple correct diagnoses were for patients with psoriasis. This may be due to over-representation of psoriatic patients in the dataset as psoriasis represented 66% of all records.

For all patients, the mean certainty of diagnosis (percentage provided by the application for each diagnosis option) was 14.5%. The mean certainty of the most certain diagnosis was 29.0%. Mean certainties of diagnosis for all patients and for psoriatic patients differed negligibly (0.1%).

**Table 1.** Percentage of patients with the correct diagnosis after one, two and three attempts. The correct diagnosis was evaluated in three variants: as the top diagnosis, among the top three and among the five possible conditions provided by the web application

| Number of samples | The most certain diagnosis (%) | | One out of three most certain diagnoses (%) | | One out of all five diagnoses (%) | |
|---|---|---|---|---|---|---|
| | All diseases (150 patients) | Psoriasis (100 patients) | All diseases (150 patients) | Psoriasis (100 patients) | All diseases (150 patients) | Psoriasis (100 patients) |
| 1 | 5.26 | 4.44 | 12.89 | 9.41 | 19.26 | 13.85 |
| 2 | 10.33 | 8.67 | 25.78 | 18.78 | 38.89 | 28.22 |
| 3 | 20.67 | 26.00 | 51.56 | 56.33 | 77.78 | 84.67 |

**Table 2.** Percentage of patients with appearance of the same, correct diagnosis in all three samples

| Number of appearances of the same, correct diagnosis | True diagnosis as the most probable outcome (%) | | True diagnosis included in three most probable outcomes (%) | | True diagnosis included in five most probable outcomes (%) | |
|---|---|---|---|---|---|---|
| | All diseases (150 patients) | Psoriasis (100 patients) | All diseases (150 patients) | Psoriasis (100 patients) | All diseases (150 patients) | Psoriasis (100 patients) |
| Two out of three | 9.33 | 10.00 | 20.67 | 25.00 | 20.67 | 23.00 |
| Three out of three (all attempts) | 3.33 | 5.00 | 16.67 | 17.00 | 36.67 | 40.00 |

**Table 3.** Percentage of patients for whom a diagnosis was made with certainty equal to or above 50% throughout one, two and three attempts for all patients and for psoriatic patients

| Number of attempts | Certainty of correct diagnosis ≥ 50 (%) | | Certainty of incorrect diagnosis ≥ 50 (%) | |
|---|---|---|---|---|
| | All diseases (150 patients) | Psoriasis (100 patients) | All diseases (150 patients) | Psoriasis (100 patients) |
| 1 | 2.7 | 3.0 | 10.7 | 11.0 |
| 2 | 0.0 | 0.0 | 0.0 | 0.0 |
| 3 | 0.0 | 0.0 | 0.0 | 0.0 |

**Table 4.** Contingency table of the web application diagnosis in psoriatic patients

| | Condition positive | Condition negative | |
|---|---|---|---|
| **Test outcome positive** | True positive = 60 | False positive = 57 | Positive predictive value = 51.28% |
| **Test outcome negative** | False negative = 240 | True negative = 93 | Negative predictive value = 27.93% |
| | Sensitivity = 20.00% | Specificity = 62.00% | |

Certainty of diagnosis equal to or above 50% was only present in 2.7% of all diseases and it was not repeated across more than one attempt. For psoriasis, this value was 3.0% and also did not repeat. Regarding incorrect diagnoses, a certainty equal to or above 50% was observed for 10.7% of the whole dataset and 11.0% of psoriasis cases. No correct or incorrect diagnoses with a certainty of 50% or more were proposed two or three times for the same patient by the application (Table 3).

In an attempt to provide a patient-oriented analysis, a weighted average of correct diagnosis was calculated, attributing greater weight to correct diagnoses repeating across the three samples, thus providing an estimate of the likelihood of a correct diagnosis when using the web application in a non-scientific approach. The weighted average value of the correct classification was found to be 19.28% for all diseases and 20.78% for psoriasis. However, when taking only the highest certainty diagnosis into account, these values fell to below 5% for all patients and to 6.67% for psoriatic patients. The chance of correct diagnosis within the first three outcomes is 12.91% for all patients and 14.11% for patients with psoriasis.

The most common incorrect diagnoses were unspecified dermatitis ($n$ = 420, 195 times as top diagnosis), folliculitis ($n$ = 154, 40 times as top diagnosis) and acne vulgaris ($n$ = 79, 20 times as top diagnosis).

Regarding precision, single measures of the two-way model ICCs for three attempts were 0.298 (95% CI: 0.196–0.404) for all patients and 0.487 (95% CI: 0.391–0.579) for psoriatic patients, assuming only the most probable diagnosis by the application, 0.361 (95% CI: 0.259–0.463) and 0.383 (95% CI: 0.282–0.484) if any of the first three are analysed, and 0.413 (95% CI: 0.313–0.511) and 0.398 (95% CI: 0.298–0.498) when all five options are considered.

To verify the results from a clinical perspective for psoriasis diagnosis, a contingency table for medical tests was constructed based on the most certain diagnosis (Table 4). Out of all first (i.e. most likely) diagnoses provided by the web application for the data set of psoriatic patients, 60 true positive results and 57 false positive results were present, indicating a positive predictive value of 51.28%. In contrast, 93 true negative results and 240 false negative results were obtained, indicating a negative predictive value of 27.93%. The application was also found to have a sensitivity of 20.00% and specificity of 62.00%.

## Discussion

The diagnosis process combines an analysis of patient history and a physical examination. In dermatology, visual analysis of the skin is fundamental to this process, and a knowledge of different types of skin lesions, as well as their distribution and symmetry, is essential. Many helpful procedures, such as histopathological or immunofluorescence evaluation, dermoscopy and ultrasonography, are based on digital image analysis.

Correct diagnosis of skin disorders is a time-consuming process, preceded by years of medical training. Machine learning, predominantly deep learning algorithms, have made huge advances in automatic identification, making diagnostics cheaper and faster. In time, their diagnoses may even become more accurate than a general practitioner or even a specialist.

The popularity of mobile applications is growing among dermatology providers and patients. With 6.3 billion smartphone subscriptions estimated to be in use by 2021, the field of smartphone diagnosis looks very promising [7].

Nowadays, there is a wide and never-ending dispute regarding the value of smartphones and new technologies, and there is great pressure on physicians to remain up-to-date with new developments and take advantage of the opportunities and challenges that lie ahead.

Currently, a key area of technological innovation is being driven by the development of AI. AI, or machine intelligence, is in fact an algorithm that allows a system

to make independent decisions based on data inputs, or one that helps users make decisions. AI simulates human intelligence processes [8, 9].

It remains uncertain whether smartphone diagnoses can be relied upon. The implementation of new technology in medical practice is beset by challenges such as engaging sufficient processing power and datasets.

Current data suggest that AI algorithms might be equal or even better than human specialists at diagnosing pigmented skin lesions and cancer. A joint research team at Stanford University used a single Convolutional Neural Network (CNN) trained on general skin lesion classification to develop an AI system that can diagnose skin cancer with similar reliability to human specialists [1]. On the other hand, there are no literature data concerning the usefulness of popular smartphone applications in diagnosing unpigmented skin conditions [10].

The present study tests a free web and smartphone application developed for easy diagnosis of various skin conditions. The application was chosen due to its popularity and ease of use: the automatic diagnosis process consists of uploading an image from a smartphone or a computer. For the best results, the image must be in focus, in good lighting, and the region of interest must be centrally located. Although all our images were taken under such conditions, the results are not promising.

The percentage of correct diagnoses tended to increase with each additional sample provided, at all diagnosis variants. After three attempts, correct diagnoses were present among the five results presented by the application for more than 75% of all diseases and almost 85% for psoriatic patients. However, when analysing the data cumulatively, the number of correct diagnoses falls considerably. The percentage of patients receiving a correct diagnosis on all three samples is roughly 40% when assuming that the correct diagnosis is one of the five diagnoses provided by the application. This value is roughly half that of the non-cumulative percentages, indicating that the chance of a correct diagnosis is much higher than outcome coherence. However, this statement holds true only when considering all five web application diagnoses, irrespective of their certainty.

The obtained results show that a hypothetical patient making one attempt using the web application to self-diagnose a skin lesion is very unlikely to receive the correct diagnosis as the most certain option: the probability of a correct diagnosis is only about 5%. If patients were to perform three attempts, the probability of a correct diagnosis is roughly 20%. Assuming that a person would be more likely to choose a diagnosis that appears multiple times throughout their tests, the likelihood of a result repeating once when analysing another place on their skin is roughly 9%. It is also three times less likely that the same diagnosis will appear on all three tests.

Taking into account the three most certain diagnoses, the probability of correct diagnosis in one attempt is two times larger than the probability for the first, most certain diagnosis. There is 50% chance to acquire a correct diagnosis among three tests. Results would repeat once for one in 5 patients and two times for about 16% of patients.

What is more, if the patient were to take all five diagnoses into account, the probability of correct diagnosis being present in the results is 20%. There is nearly 80% likelihood of a correct diagnosis to be present when three attempts are made. Interestingly, in this scenario, the correct diagnosis would be seen in two out of three attempts for roughly 21% of patients and in all three attempts for nearly 37% of patients.

Even though the last results look promising, the scenario is the least likely to be followed by a patient, who does not have medical knowledge and cannot fully assess the diagnoses. It is doubtful that anyone would like to perform three tests and then compare a maximum of fifteen diagnoses. However, in such a scenario, the correct diagnosis is at roughly 40% probability to be present.

Our clinical practice indicates that a patient is more likely to believe a diagnosis of certainty above 50%. Such certainty was present for roughly 11% of the incorrect diagnoses, and 3% of the correct diagnoses. Luckily, there were no triple incorrect diagnoses with more than 50% certainty. Such low certainties of diagnosis are much more beneficial when analysing the effect of automated self-diagnosis as a whole. A person without medical knowledge could be convinced that their skin condition is not dangerous when a highly certain, but false result occur.

Additionally, the applications' most frequent suggestion is an unspecified dermatitis. That is a vague outcome as most dermatological diseases have an inflammatory background. However, it is also a safe diagnosis as it does not give any specific pathology and may be more of an incentive for the patient to visit a specialist.

Regarding the precision of the web application algorithm, the relation between diagnosis probability and diagnosis repetition correspond to the precision values of the ICCs. The web application exhibits fair precision according to Cicchetti [5]; however, as the ICCs do not exceed 0.600, they are poor according to Portney and Watkins [6]. Interestingly, precision rises when taking three diagnosis options into account, and by nearly 50% when all five options are considered. However, for psoriasis, the highest precision is exhibited when only the first, most certain diagnosis is analysed.

As patients are subjected to an overabundance of information and applications, it is likely that some of it is incorrect. There is a great need to educate patients and inexperienced doctors regarding accurate dermatology resources and evaluate their reliability.

However, future usage of such applications could be beneficial for a non-dermatologist medical professional. A 20% sensitivity for psoriasis skin lesions makes the ap-

plication unsuitable for screening and positive predictive value of around 50% is nowhere near clinical acceptability [11]. However, the specificity of 62% could be a sign that the algorithm, with some work, could be used to rule out psoriasis. Also, an informed guess from fifteen diagnoses after three tests and a 40% probability of two repetitions of diagnoses may be of value. Unfortunately, it is impossible though for the user to access a full list of possible diagnoses.

We strongly believe that as technology evolves, mobile applications have the potential to progress as well. Some corrections or new ideas are needed to obtain more accurate outcomes, including a standardized scientific approach to evaluating the validity and reliability of an application, as well as more and a better quality of input data. AI will not replace physicians, but might support their work in near future. Knowing these upcoming trends will help understand the impact of new applications on patients and health care providers. Although the potential of AI applications is huge, the services currently available to patients free of charge require criticism and cannot serve as sole diagnostic tools.

## Acknowledgments

This work was supported be the Medical University of Lodz, grant number 503/1-152-01/503-11-001-19-00.

Maksym Mikołajczyk, Sebastian Patrzyk – equal contribution.

## Conflict of interest

The authors declare no conflict of interest.

### References

1. Tschandl P, Codella N, Akay BN, et al. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. Lancet Onkol 2019; 20: 938-47.
2. Tongdee E, Markowitz O. Mobile App rankings in dermatology. Cutis 2018; 102: 252-6.
3. International vocabulary of metrology – Basic and general concepts and associated terms (VIM). JCGM 2008.
4. Bruton A, Conway JH, Holgate ST. Reliability: what is it, and how is it measured? Physiotherapy 2000; 86: 94-9.
5. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess 1994; 6: 284-90.
6. Portney LG, Watkins MP. Foundations of clinical research: applications to practice. Prentice Hall Inc., New Jersey 2000.
7. Charalambides M. Dermatology skin cancer applications: the future of healthcare provision? J Natl Student Assoc Med Res 2018; 1: 45-9.
8. Brinker TJ, Hekler A, Enk AH, et al. Deep learning out performed 136 of 157 dermatologists in head-to-head dermoscopic melanoma image classification task. Eur J Cancer 2019; 113: 47-54.
9. Li CH, Shen CB, Xue K, et al. Artificial intelligence in dermatology: past, present, and future. Chin Med J 2019; 132: 2017-20.
10. Haenssle HA, Fink C, Schneiderbauer R, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol 2018; 1: 1836-42.
11. Lalkhen AG, McCluskey A. Clinical tests: sensitivity and specificity. Continuing Education in Anaesthesia Critical Care Pain 2008; 8: 221-3.