Received: 3 January 2022 | Accepted: 10 November 2022

DOI: 10.1111/jdv.18853

JEADV JOURNAL OF THE EUROPEAN ACADEMY OF DERMATOLOGY & VENEREOLOGY

ORIGINAL ARTICLE

Outlier detection in dermatology: Performance of different convolutional neural networks for binary classification of inflammatory skin diseases

Maximilian C. Schielein[1,2] | Joshua Christl[3] | Sebastian Sitaru[1] | Anna Caroline Pilz[1] | Robert Kaczmarczyk[1,2] | Tilo Biedermann[1] | Tobias Lasser[3] | Alexander Zink[1,2]

[1]Department of Dermatology and Allergy, School of Medicine, Technical University of Munich, Munich, Germany

[2]Unit of Dermatology and Venerology, Department of Medicine, Karolinska Institutet, Karolinska University Hospital, Stockholm, Sweden

[3]Department of Informatics and Munich School of BioEngineering, Technical University of Munich, Munich, Germany

Correspondence
Alexander Zink, Department of Dermatology and Allergy, School of Medicine, Technical University of Munich, Biedersteiner Str. 29, 80802 Munich, Germany.
Email: alexander.zink@tum.de

Abstract

Background: Artificial intelligence (AI) and convolutional neural networks (CNNs) represent rising trends in modern medicine. However, comprehensive data on the performance of AI practices in clinical dermatologic images are non-existent. Furthermore, the role of professional data selection for training remains unknown.

Objectives: The aims of this study were to develop AI applications for outlier detection of dermatological pathologies, to evaluate CNN architectures' performance on dermatological images and to investigate the role of professional pre-processing of the training data, serving as one of the first anchor points regarding data selection criteria in dermatological AI-based binary classification tasks of non-melanoma pathologies.

Methods: Six state-of-the-art CNN architectures were evaluated for their accuracy, sensitivity and specificity for five dermatological diseases and using five data subsets, including data selected by two dermatologists, one with 5 and the other with 11 years of clinical experience.

Results: Overall, 150 CNNs were evaluated on up to 4051 clinical images. The best accuracy was reached for onychomycosis (accuracy = 1.000), followed by bullous pemphigoid (accuracy = 0.951) and lupus erythematosus (accuracy = 0.912). The CNNs InceptionV3, Xception and ResNet50 achieved the best accuracy in 9, 8 and 6 out of 25 data sets, respectively (36.0%, 32.0% and 24.0%). On average, the data set provided by the senior physician and the data set provided in accordance with both dermatologists performed the best (accuracy = 0.910).

Conclusions: This AI approach for the detection of outliers in dermatological diagnoses represents one of the first studies to evaluate the performance of different CNNs for binary decisions in clinical non-dermatoscopic images of a variety of dermatological diseases other than melanoma. The selection of images by an experienced dermatologist during pre-processing had substantial benefits for the performance of the CNNs. These comparative results might guide future AI approaches to dermatology diagnostics, and the evaluated CNNs might be applicable for the future training of dermatology residents.

Maximilian C. Schielein and Joshua Christl contributed equally as shared first authors.

Tobias Lasser and Alexander Zink contributed equally as shared last authors.

INTRODUCTION

Artificial intelligence (AI) refers to the simulation of human intelligence processes by machines.[1] It represents a rising trend in modern medicine, with numerous publications describing AI-based identification of diseases based on clinical images, especially using deep learning and convolutional neural networks (CNNs).[2–5] CNNs are special types of neural networks in die field of machine learning that are especially beneficial when working with visual data and have high potential for applications in medical fields that utilize imagery data.[6] CNN architectures are based on neural networks and have capabilities to recognize visual patterns inside images, independent of their position. While all architectures are based on the same key principle, different architectures from various researchers have been developed, differing in their depth, structure and connections between neurons. It is not inherently clear which architectures are the best choice for which use case, particularly in medical applications. Recently, the combination of AI and human intelligence was promoted for the detection of skin cancer in parallel with a growing acceptance of AI in medicine among patients.[7,8] AI may therefore soon become part of day-to-day dermatological practice and clinical routine.[8–10] All technical terminology that is made use of can be found in Table S1.

Skin diseases affect up to 30% of the general population but vary widely regarding their epidemiology and pathophysiology.[11,12] While chronic inflammatory skin diseases, including psoriasis and atopic dermatitis, are common and exhibit a point prevalence of up to 8.7% among the European population, dermatoses like lupus erythematosus and bullous pemphigoid are rare.[12–14] Furthermore, skin diseases are highly diverse regarding their clinical appearance, which can include scaling, blistering, discoloration and redness, and can be difficult to differentiate.

In addition to their diversity, skin diseases are habitually diagnosed and treated in interdisciplinary settings and by physicians who potentially have limited dermatological experience.[15,16] Elaborated classification aids for diagnosis or disease prediction are already widely applied in the field of melanoma diagnostics.[7,9,17–21] While first approaches presented promising results using different frameworks of AI, data on the best-performing CNNs and direct comparisons of their performance in various dermatological diagnoses are rare and if available, mostly based on the classification of pigmented skin lesions and dermatoscopic images.[5,19–22] Scientists developing new AI approaches in the field of dermatology usually have to rely on performance parameters established using unspecific data sets with no context to dermatological imaging.[23] Additionally, each neural network depends on the data provided for its training. However, the exact influence of data quality and pre-processing methods for the use in neural networks regarding clinical non-dermatoscopic images remains unidentified.

Therefore, the objectives of this study were not merely to develop an AI-based system for outlier detection of dermatological diseases (the detection of images differing significantly visually from other images in the same data set for each respective disease), but also (i) to directly compare existing neural network architectures and evaluate their performance in different dermatological diagnoses as well as (ii) to determine the performance of AI-based classification depending on the quality of training data as provided by clinicians with different experience levels. The developed AI system was examined for five pathologies selected because of either their high prevalence or diverse visual appearance: psoriasis, atopic dermatitis, lupus erythematosus, bullous pemphigoid and onychomycosis.

MATERIAL AND METHODS

Study data

Clinical images of each of the investigated heterogeneous diagnoses (psoriasis, atopic dermatitis, lupus erythematosus, bullous pemphigoid and onychomycosis) were retrieved from an image database featuring professional photos taken during routine care at a university hospital. The administered protocol was developed in a consortium of one senior dermatologist, one junior dermatologist, two epidemiologists and two members of the Department of Informatics of the Technical University of Munich. No approval of the ethics commission was needed due to (i) all data entries being handled locally and (ii) all data entries being secondary data collected during clinical routine care.[24] Images were chosen from the database at random and based on their absolute frequency. Subsequently, the amount of data for each diagnosis was not equally distributed. Inclusion criteria were (i) secured diagnosis or first-named differential diagnosis by a physician and (ii) the presence of a skin alteration in the image. The retrieved data set contained 1200 images for psoriasis, 1038 images for atopic dermatitis, 726 images for lupus erythematosus, 881 images for bullous pemphigoid and 276 images for onychomycosis. There was a total of 4051 images for the initial data set (Figure 1).

Selection process

One junior dermatologist (5 years of experience at a university dermatology hospital) and one senior dermatologist (11 years of experience at a university dermatology hospital) individually reviewed all included images and were instructed to identify and delete incongruous photos. Thereby, exclusion criteria were (i) reasonable doubt about the accuracy of the stated diagnosis, (ii) other skin diseases in a particular image, (iii) and the presence of widespread coloured antiseptic agents. Subsequently, five data sets were defined: the full study data, data from the full study data selected by a young dermatologist (junior dermatologist), data from the full study data selected by a more experienced dermatologist

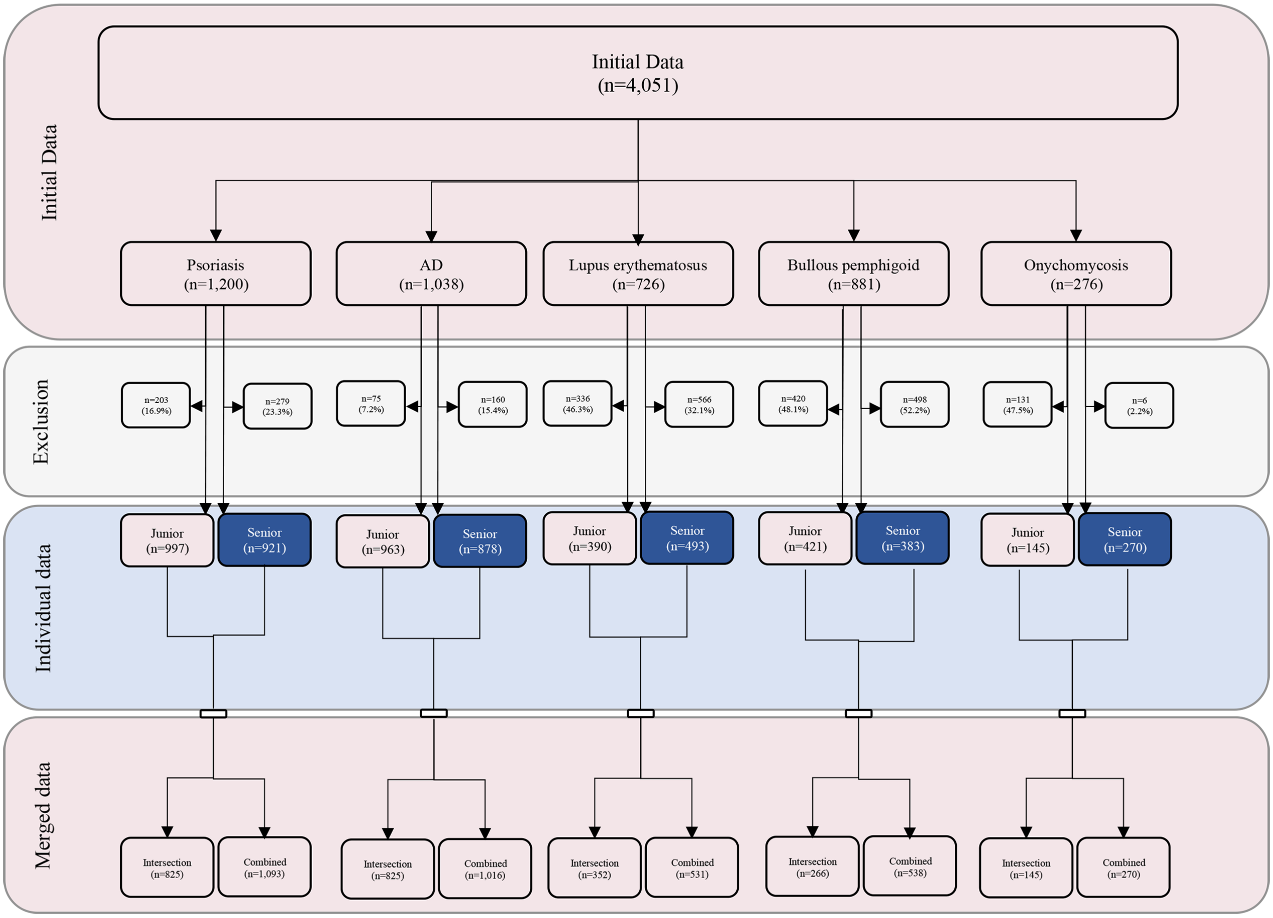

FIGURE 1 Data flow diagram. Junior = dermatologist with 5 years of experience; Senior = dermatologist with 11 years of experience; Intersection = images included by both dermatologists; Union = images included by at least one dermatologist.

(senior dermatologist), data included in both of the latter (intersection) and data included in at least one of the dermatologists' data sets (union).

Procedure and modifications

Due to their high applicability regarding image data, convolutional neural networks (CNNs) were leveraged to classify the data. Six different state-of-the-art network architectures were utilized: VGG-16,[25] VGG-19,[25] Xception,[26] InceptionV3,[27] ResNet50[28] and MobileNetV2.[29] The choice of these architectures was based on their previous usage in medical image classification as well as their performance in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), a large-scale competition for the classification of images in which different network architectures are measured regarding their accuracy at predicting image classes.[23] Especially VGG-16, ResNet50 and MobileNetV2 already demonstrated reliable performance for classification tasks of mammary carcinoma.[30]

Each network was set up to conform to a binary classification task. Hence, one event corresponds to the classification 'normality', and its complementary event corresponds to the classification 'outlier'.

For a technical description on the adaption of the utilized neural network architectures on the study's data, refer to Table S2.

For each pathology, the category 'normality' was filled with images of the actual pathology, and the category 'outlier' was filled with an equally distributed number of images from the other four pathologies. For example, the data set 'psoriasis initial' contained 1200 images for the classification 'normality' that were actual images of psoriasis and about 300 (<300 for onychomycosis due to lack of data) random images of each of the other pathologies (1200 random images in total) for the classification 'outlier'.

Statistics

The scarce amount of data in each class (especially for onychomycosis, lupus erythematosus and bullous pemphigoid) meant that each data set was split into 80% training data and 20% test data. The data sets were split randomly and automatically, preserving the respective class distributions. The images were strictly partitioned by the associated patient, meaning that one patient's images could either be in the training data set or in the test data set, but there could be no

TABLE 1 Accuracy, sensitivity and specificity of best performing convolutional neural networks (according to accuracy) for diseases and their respective data sets

	Psoriasis			Atopic dermatitis			Lupus erythematosus			Bullous pemphigoid			Onychomycosis			Average		
	Acc	Sens	Spec	Acc	Sens	Spec	Acc	Sens	Spec	Acc	Sens	Spec	Acc	Sens	Spec	Acc	Sens	Spec
Initial data	0.831	0.790	0.873	0.769	0.818	0.719	0.815	0.785	0.846	0.771	0.677	0.872	1.000	1.000	1.000	0.837	0.814	0.862
Junior dermatologist	0.847	0.776	0.909	0.820	0.851	0.790	0.737	0.787	0.688	0.845	0.859	0.831	0.947	0.897	1.000	0.839	0.834	0.844
Senior dermatologist	0.899	0.911	0.885	0.826	0.864	0.786	0.912	0.980	0.840	0.931	0.889	0.973	0.981	0.975	0.987	0.910	0.924	0.894
Intersection	0.871	0.903	0.841	0.873	0.821	0.924	0.873	0.838	0.912	0.951	0.982	0.917	0.984	0.972	1.000	0.910	0.903	0.919
Union	0.875	0.874	0.876	0.829	0.776	0.883	0.871	0.908	0.832	0.859	0.892	0.824	0.990	0.979	1.000	0.885	0.886	0.883

Note: Values per column are highlighted according to the respective performance, ranging from worst (intense red) to best (intense green).

Abbreviations: Acc, Accuracy; Sens, Sensitivity; Spec, Specificity.

subset of a patient's images in both training and test data set. This ensured that the network's predictions were not based on recognition of a certain patient. Every network architecture was trained on every training data set. The performance of the respective architectures was evaluated using achieved test accuracy (accuracy value on the test data set based on the network's predictions after training). For all data sets and models, sensitivity and specificity were calculated. In the case of multiple networks achieving the same accuracy and therefore being considered as best fit, the mean of the respective sensitivities and specificities of all tied networks were considered for the evaluation of different data sets.

RESULTS

Input

Overall, 4051 clinical images of the investigated dermatoses were identified. The respective number of images ranged from 267 for onychomycosis to 1200 for psoriasis (Figure 1). On average, both physicians excluded 27.7% of images per disease. The least images were excluded for atopic dermatitis (average: 11.3%) and the most were excluded for bullous pemphigoid (average: 50.4%). The senior physician eliminated 1106 images, resulting in a data subset of 2945 images. The junior physician eliminated 1135 images, resulting in a data subset of 2916 images. Combining both subsets, the intersection data subset comprised 2413 images (1638 images fewer than the initial data set). The high difference between the physicians' revised data sets and the intersection data set (δ_{senior} = 532, δ_{junior} = 503) shows low cohesion of both subsets, meaning that there was a substantial difference between the images selected by the two physicians.

Overall performance

The overall best accuracy was reached for onychomycosis (accuracy = 1.000), followed by bullous pemphigoid (accuracy = 0.951), lupus erythematosus (accuracy = 0.912) and psoriasis (accuracy = 0.899). All achieved accuracy values were at least 0.873. Therefore, for each disease about 9 out of 10 images were classified correctly, meaning that images labelled as ‘outlier’ were also classified as ‘outlier’ by the network, and images labelled as ‘normality’ were also classified as ‘normality’. For all best-performing models, sensitivity was at least 0.821 (atopic dermatitis), showing correct identification in more than 8 out of 10 outliers, while the lowest specificity was 0.840 (lupus erythematosus), indicating that 84.1% of images were correctly identified as being of type ‘outlier’ or ‘normality’ (Table 1).

Model performance

Six networks were trained for each disease in each of the five data sets, yielding a total of 150 CNNs. The accuracy

TABLE 2 Best performing architectures for each disease and data set according to accuracy values

	Psoriasis		Atopic dermatitis		Lupus erythematosus		Bullous pemphigoid		Onychomycosis	
	CNN	Accuracy	CNN	Accuracy	CNN	Accuracy	CNN	Accuracy	CNN	Accuracy
Initial data	VGG-16/ Xception/ ResNet50	0.831	InceptionV3	0.769	ResNet50	0.815	Xception	0.771	VGG-19/InceptionV3/ Xception/ResNet50	**1.000**
Junior dermatologist	VGG-16	0.847	ResNet50	0.820	MobileNetV2	0.737	VGG-19	0.845	InceptionV3	0.947
Senior dermatologist	**InceptionV3**	**0.899**	Xception	0.826	Xception	0.912	ResNet50	0.931	VGG-19/InceptionV3/ Xception	0.981
Intersection	ResNet50	**0.871**	Xception	0.873	MobileNetV2	0.873	**InceptionV3**	0.951	MobileNetV2	0.984
Union	Xception	0.875	InceptionV3	0.829	InceptionV3	0.871	InceptionV3	0.859	VGG-16	0.990

Note: For each column, the highest achieved accuracy value and its associated CNN is highlighted.

values of final CNNs ranged from 0.651 for the data provided by the junior physician on lupus erythematosus using the VGG-16 model to 1.000 for the initial data set of onychomycosis using VGG-19, Inceptionv3, Xception or ResNet50. All CNN performances and the data sets can be seen in Table S2. InceptionV3 achieved the best accuracy in 9 out of 25 data sets (36.0%), followed by Xception (8/25; 32.0%) and ResNet50 (6/25; 24.0%, Table 2). Overall, InceptionV3 showed the best average accuracy and sensitivity, while Xception outperformed the other CNNs regarding specificity (Figure 2, Table S3).

Sample performance

The performance among all networks was the highest in the intersection data set for atopic dermatitis (accuracy = 0.873) and bullous pemphigoid (accuracy = 0.951). Data provided by the senior physician performed best for psoriasis (accuracy = 0.899) and lupus erythematosus (accuracy = 0.912), while the initial data outperformed in images of onychomycosis (accuracy = 1.000). On average, the intersection data set performed the best (accuracy = 0.910), followed by data provided by the senior physician (accuracy = 0.910), the union data (accuracy = 0.885) and data of the junior physician (accuracy = 0.839). The average values of all data sets exceeded that of the initial data (accuracy = 0.837; Table 1), meaning that every revision of the data sets was beneficial to the average accuracy value (Figure 3).

Sensitivity ranged from 0.677 for bullous pemphigoid to 1.000 for onychomycosis, both utilizing the initial data set. Furthermore, specificity ranged from 0.688 for lupus erythematosus in the junior dermatologist data set to 1.000 for onychomycosis in all data sets except for the one provided by the senior dermatologist. The data set provided by the senior dermatologist performed best for average sensitivity (0.924) and second best for average specificity (0.894), while the intersection data set performed best for average specificity (0.919) and second best for average sensitivity (0.903). In all pathologies besides onychomycosis, either the senior dermatologist data set or the intersection data set performed the best regarding their achieved accuracy values (Table 2).

DISCUSSION

Overall, well performing AI systems for the detection of outliers in five heterogeneous dermatoses were established. InceptionV3, Xception and ResNet50 were identified as most likely to outperform other CNNs for the binary detection of outliers in dermatological images. Additionally, the selection of images by an experienced dermatologist or the combination of two dermatologists prior to training AI systems showed substantial benefits for the respective performances of the CNNs.

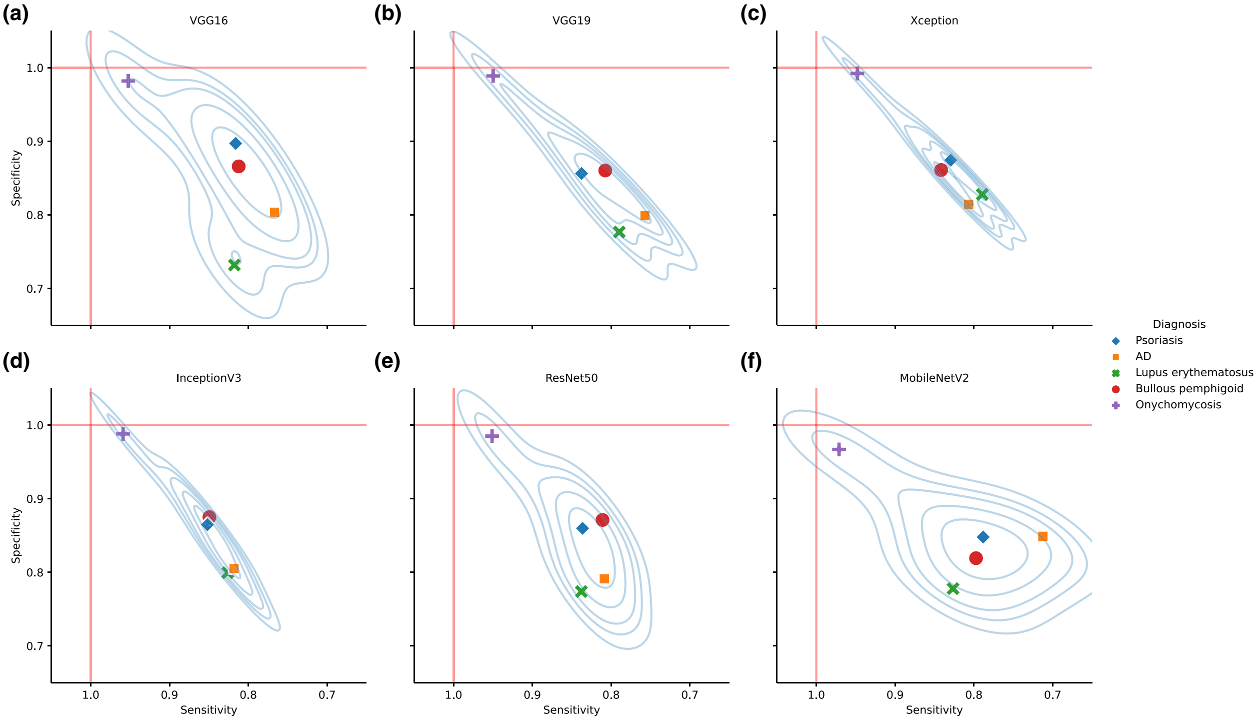

FIGURE 2 Average performances (sensitivity and specificity) and distribution of performance per convolutional neural networks (CNN) indicated as the respective centre of the dark blue borders and the course of the further borders. Additionally, the average performance for each disease and CNN are represented as means with no indication of gradient. Every datapoint shows the achieved sensitivity (*X*-axis) and specificity (*Y*-axis) values for a given pathology. (a) VGG-16, (b) VGG-19, (c) Xception, (d) InceptionV3, (e) ResNet50 and (f) MobileNetV2.

Overall performance

The average accuracy value of 0.910 in the two best performing data sets is substantial and exceeds previous results achieved for binary classifications of dermatoscopic images of melanoma (0.819).[31] These results resemble accuracy values found for the differentiation of psoriasis, atopic dermatitis and healthy skin (0.895–0.926).[32] However, the herein presented procedure resulted in CNNs trained on non-dermatoscopic images and a binary detection of outliers. This in turn results in more diverse tasks and a direct comparison needs to be made with caution. As about 91% of all classifications for outliers were correct, usage of such a system can be highly beneficial, especially for training students, dermatological residents and other physicians in dermatological diagnostics. By providing direct feedback on previous diagnostic decisions, these tools may be highly valuable for rethinking pathophysiology and clinical expression.

Model performance

InceptionV3, Xception and ResNet50 were selected as best performing most frequently when applied to the selected data sets. While InceptionV3 yielded the best average performances for accuracy and sensitivity, Xception yielded the best average performance for specificity. This is, however, partly in contrast to the respective CNN's Top-1 accuracy values achieved in prior studies. Xception previously outperformed both InceptionV3 and ResNet50 regarding the CNNs' accuracy (0.790 vs. 0.779 vs. 0.749).[26–28] Even though, these differences are small, all three on average perform better than the remaining three networks, VGG-16, VGG-19 and MobileNetV2, comparable to the herein presented performance on dermatological data. Therefore, data presented in this study represents a resource for researchers about which CNN to use depending on the dermatoses and the pre-selection of included images.

Considering the accuracy values, most of the deeper architectures (architectures that contain more layers; InceptionV3, Xception, ResNet50 and not MobileNetV2) performed better in binary classification tasks among dermatological non-dermatoscopic images than models with less layers (VGG-16, VGG-19). VGG-16 and VGG-19 only constituted the best architectural choice for 4% of each of the data sets. While Xception outperformed InceptionV3 in the ImageNet Large Scale Visual Recognition Challenge 2017 (ILSVRC2017),[23,26] the contrary applies to data presented in this study. Herein, InceptionV3 slightly outperformed Xception (36.0% vs. 32.0%). The presented data therefore can help researchers in choosing which CNN to use depending on the group of dermatoses and the pre-processing of included dermatological non-dermatoscopic images. This groundwork, in turn, can minimize repetitive, exploratory and time-consuming trainings of multiple CNN architectures for future research.

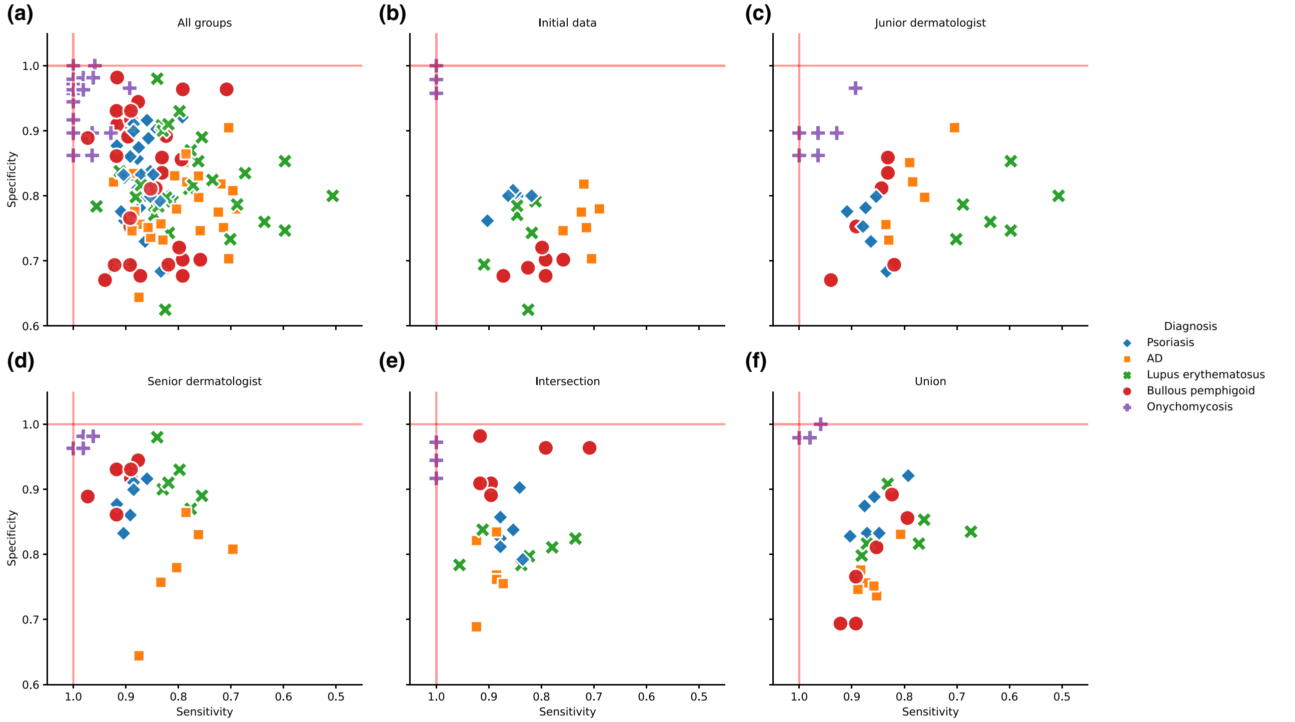

FIGURE 3 All performances (sensitivity and specificity) of all convolutional neural networks (CNN) overall as well as stratified according to data set. (a) Overall, (b) initial (c) junior dermatologist (5 years of experience), (d) senior dermatologist (11 years of experience), (e) intersection and (f) union.

Not being explored in this study is the possibility of using ensemble learning as a way to further increase the achieved sensitivity and specificity values. Additionally, while this study worked with non-dermatoscopic dermatological imagery data, utilization of dermatoscopic images presents another possibility for even better results in the classification tasks. With the achieved accuracy of the networks depending on the quality of the ground truth data (i.e., the image labels), the usage of biopsy results or panel consensus is another option for improving results.

Sample performance

Data sets reviewed by a senior dermatologist and the data set which included images consistently included by both a junior and a senior dermatologist (intersection) performed best among all data sets. This strongly indicates a positive effect of experienced selection as well as appraisal by several dermatologists regarding AI approaches for dermatological non-dermatoscopic images. While a previous study suggested the benefit of the combination of human and artificial intelligence to optimize predictions after the training of a CNN,[7] the current study shows the beneficial effects of the inclusion of expert knowledge in the selection process of included clinical images. It is important to note that by reviewing the clinical images, data sets become more homogenous. However, the effects are also based on the quality of the selection process and not merely on the number of images that are excluded. This is particularly evident from the selection by the experienced dermatologist, as CNNs trained on this data obtained substantially better results despite a similar exclusion rate. Additionally, it stands out that the appraisal by two dermatologists also resulted in a better specificity. Overall, this indicates that adequate, experienced and prudent pre-processing of clinical dermatological images is crucial for the quality and performance of the respective AI approaches.

Strengths and limitations

One of the main limitations of this study is the underlying database. The images taken during routine care in a dermatological university hospital might not necessarily be representative of the respective diseases. As photos are more likely to be taken of rare dermatoses, for example psoriasis pustulosa, these less common manifestations of a disease might be overrepresented compared to their actual epidemiological prevalence. Additionally, no histological confirmation of disease was present, and no external validation was conducted, which both could improve the ground-truth labels of the data. Another limitation of the study is the inhomogeneous distribution of data set sizes among the respective pathologies. The initial psoriasis data set is 4.3 times larger than the onychomycosis data set. As onychomycosis only appears on feet and toenails, while the other included diseases typically do not, the feature selection of this pathology's CNNs may be based more on the detection of feet and toenails than the clinical presentation of the disease. Despite these limitations, the strength of this study is that it presents the first direct comparison of six state-of-the-art CNNs among clinical dermatological images. Furthermore, the total number of underlying images was relatively high (>4000).

CONCLUSION

Overall, the herein presented AI approach for the detection of outliers in dermatological diagnoses represents one of the first studies to evaluate the performance of different CNNs for binary decisions in non-dermatoscopic images of a variety of dermatological diseases other than melanoma. The first conclusion to draw is that InceptionV3, Xception and ResNet50 were the three most promising CNN architectures. Another conclusion is that the clinical expertise of a senior physician is crucial for the development of better performing networks and is directly reflected in our data with higher sensitivity and specificity values. This underlines the importance of the clinical expertise of practicing dermatologists, especially when developing AI-based approaches to diagnostics. With more than 9 out of 10 images being classified correctly, the evaluated CNNs represent a promising and useful tool for minimizing human error in daily clinical practice and potentially helping guide physicians.

AUTHOR CONTRIBUTIONS

MCS and JC jointly designed and performed experiments, analysed data and wrote the first draft of the manuscript; SS substantially contributed to the study design and the interpretation of data; CP substantially contributed to the curation and preparation of images; RK made substantial contributions to the preparation and interpretation of data as well as to the graphical presentation of the study results; TB made substantial contributions to the interpretation of data; TL and AZ jointly supervised the project and substantially contributed to the study design. All authors have approved the submitted version of the manuscript and have agreed both to be personally accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated and resolved and that the resolution be documented in the literature.

ACKNOWLEDGEMENTS

Open Access funding enabled and organized by Projekt DEAL.

FUNDING INFORMATION

None.

CONFLICT OF INTEREST

There are no conflicts of interests.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request. However, access to images, which can be used to identify individuals, cannot be granted.

ORCID

Maximilian C. Schielein https://orcid.org/0000-0003-3767-1337
Joshua Christl https://orcid.org/0000-0001-5364-9418
Sebastian Sitaru https://orcid.org/0000-0001-7324-9139
Anna Caroline Pilz https://orcid.org/0000-0002-3028-4556
Robert Kaczmarczyk https://orcid.org/0000-0002-8570-1601
Tilo Biedermann https://orcid.org/0000-0002-5352-5105
Tobias Lasser https://orcid.org/0000-0001-5669-920X
Alexander Zink https://orcid.org/0000-0001-9313-6588

REFERENCES

1. Russell SJ, Norvig P. Artificial intelligence. A modern approach. Boston, MA: Pearson; 2022.
2. Esteva A, Chou K, Yeung S, Naik N, Madani A, Mottaghi A, et al. Deep learning-enabled medical computer vision. NPJ Digit Med. 2021;4:5. https://doi.org/10.1038/s41746-020-00376-2
3. Pouly M, Koller T, Gottfrois P, Lionetti S. Künstliche intelligenz in der bildanalyse – grundlagen und neue entwicklungen. Hautarzt. 2020;71:660–8. https://doi.org/10.1007/s00105-020-04663-7
4. Du-Harpur X, Watt FM, Luscombe NM, Lynch MD. What is AI? Applications of artificial intelligence to dermatology. Br J Dermatol. 2020;183:423–30. https://doi.org/10.1111/bjd.18880
5. Liu Y, Jain A, Eng C, Way DH, Lee K, Bui P, et al. A deep learning system for differential diagnosis of skin diseases. Nat Med. 2020;26:900–8. https://doi.org/10.1038/s41591-020-0842-3
6. Mitchell TM. Machine learning. New York, NY: McGraw-Hill; 1997.
7. Hekler A, Utikal JS, Enk AH, Hauschild A, Weichenthal M, Maron RC, et al. Superior skin cancer classification by the combination of human and artificial intelligence. Eur J Cancer. 2019;120:114–21. https://doi.org/10.1016/j.ejca.2019.07.019
8. Jutzi TB, Krieghoff-Henning EI, Holland-Letz T, Utikal JS, Hauschild A, Schadendorf D, et al. Artificial intelligence in skin cancer diagnostics: the patients' perspective. Front Med. 2020;7:233. https://doi.org/10.3389/fmed.2020.00233
9. Brinker TJ, Schlager G, French LE, Jutzi T, Kittler H. Computerassistierte hautkrebsdiagnose: wann kommt künstliche intelligenz in der praxis an? Hautarzt. 2020;71:669–76. https://doi.org/10.1007/s00105-020-04662-8
10. Maul LV, Meienberger N, Kaufmann L. Stellenwert der künstlichen intelligenz zur ausbreitungsdiagnostik und verlaufsbeurteilung von dermatosen. Hautarzt. 2020;71:677–85. https://doi.org/10.1007/s00105-020-04657-5
11. Augustin M, Herberger K, Hintzen S, Heigel H, Franzke N, Schäfer I. Prevalence of skin lesions and need for treatment in a cohort of 90 880 workers. Br J Dermatol. 2011;165:865–73. https://doi.org/10.1111/j.1365-2133.2011.10436.x
12. Hay RJ, Johns NE, Williams HC, Bolliger IW, Dellavalle RP, Margolis DJ, et al. The global burden of skin disease in 2010: an analysis of the prevalence and impact of skin conditions. J Invest Dermatol. 2014;134:1527–34. https://doi.org/10.1038/jid.2013.446
13. Parisi R, Iskandar IYK, Kontopantelis E, Augustin M, Griffiths CEM, Ashcroft DM, et al. National, regional, and worldwide epidemiology of psoriasis: systematic analysis and modelling study. BMJ. 2020;369:m1590. https://doi.org/10.1136/bmj.m1590

14. Bylund S, Von Kobyletzki LB, Svalstedt M, Svensson Å. Prevalence and incidence of atopic dermatitis: a systematic review. Acta Derm Venereol. 2020;100:adv00160. https://doi.org/10.2340/00015555-3510
15. Noels EC, Lugtenberg M, Egmond S, Droger SM, Buis PAJ, Nijsten T, et al. Insight into the management of actinic keratosis: a qualitative interview study among general practitioners and dermatologists. Br J Dermatol. 2019;181:96–104. https://doi.org/10.1111/bjd.17818
16. Harkemanne E, Baeck M, Tromme I. Training general practitioners in melanoma diagnosis: a scoping review of the literature. BMJ Open. 2021;11:e043926. https://doi.org/10.1136/bmjopen-2020-043926
17. Young AT, Fernandez K, Pfau J, Reddy R, Cao NA, von Franque MY, et al. Stress testing reveals gaps in clinic readiness of image-based diagnostic artificial intelligence models. NPJ Digit Med. 2021;4:10. https://doi.org/10.1038/s41746-020-00380-6
18. Brinker TJ, Kiehl L, Schmitt M, Jutzi TB, Krieghoff-Henning EI, Krahl D, et al. Deep learning approach to predict sentinel lymph node status directly from routine histology of primary melanoma tumours. Eur J Cancer. 2021;154:227–34. https://doi.org/10.1016/j.ejca.2021.05.026
19. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115–8. https://doi.org/10.1038/nature21056
20. Romero-Lopez A, Giro-i-Nieto X, Burdick J, Marques O. Skin lesion classification from dermoscopic images using deep learning techniques. In: Kiss R, Thurner PJ, editors. Proceeding of the International Conference on Biomedical Engineering. February 20–21, 2017, Innsbruck, Austria. Piscataway, NJ: IEEE; 2017.
21. Li Y, Shen L. Skin lesion analysis towards melanoma detection using deep learning network. Sensors (Basel). 2018;18:556. https://doi.org/10.3390/s18020556
22. Brinker TJ, Hekler A, Utikal JS, Grabe N, Schadendorf D, Klode J, et al. Skin cancer classification using convolutional neural networks: systematic review. J Med Internet Res. 2018;20:e11936. https://doi.org/10.2196/11936
23. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis. 2015;115:211–52. https://doi.org/10.1007/s11263-015-0816-y
24. Irwin S. Qualitative secondary data analysis: ethics, epistemology and context. Prog Dev Stud. 2013;13:295–306. https://doi.org/10.1177/1464993413490479
25. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [2014 Sep 04].
26. Chollet F. Xception: deep learning with depthwise separable convolutions [2016 Oct 07].
27. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision [2015 Dec 02].
28. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition [2015 Dec 10].
29. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C. MobileNetV2: inverted residuals and linear bottlenecks. *The IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*).
30. Masud M, Eldin Rashed AE, Hossain MS. Convolutional neural network-based models for diagnosis of breast cancer. Neural Comput Appl. 2020;34:11383–94. https://doi.org/10.1007/s00521-020-05394-5
31. Maron RC, Utikal JS, Hekler A, Hauschild A, Sattler E, Sondermann W, et al. Artificial intelligence and its effect on dermatologists' accuracy in dermoscopic melanoma image classification: web-based survey study. J Med Internet Res. 2020;22:e18091. https://doi.org/10.2196/18091
32. Wu H, Yin H, Chen H, Sun M, Liu X, Yu Y, et al. A deep learning, image based approach for automated diagnosis for inflammatory skin diseases. Ann Transl Med. 2020;8:581. https://doi.org/10.21037/atm.2020.04.39

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Schielein MC, Christl J, Sitaru S, Pilz AC, Kaczmarczyk R, Biedermann T, et al. Outlier detection in dermatology: Performance of different convolutional neural networks for binary classification of inflammatory skin diseases. J Eur Acad Dermatol Venereol. 2023;37:1071–1079. https://doi.org/10.1111/jdv.18853