Collective human intelligence outperforms artificial intelligence in a skin lesion classification task

Summary

Background and objectives: Convolutional neural networks (CNN) enable accurate diagnosis of medical images and perform on or above the level of individual physicians. Recently, collective human intelligence (CoHI) was shown to exceed the diagnostic accuracy of individuals. Thus, diagnostic performance of CoHI (120 dermatologists) versus individual dermatologists versus two state-of-the-art CNN was investigated.

Patients and Methods: Cross-sectional reader study with presentation of 30 clinical cases to 120 dermatologists. Six diagnoses were offered and votes collected via remote voting devices (quizzbox®, Quizzbox Solutions GmbH, Stuttgart, Germany). Dermatoscopic images were classified by a binary and multiclass CNN (FotoFinder Systems GmbH, Bad Birnbach, Germany). Three sets of diagnostic classifications were scored against ground truth: (1) CoHI, (2) individual dermatologists, and (3) CNN.

Results: CoHI attained a significantly higher accuracy [95% confidence interval] (80.0% [62.7%–90.5%]) than individual dermatologists (75.7% [73.8%–77.5%]) and CNN (70.0% [52.1%–83.3%]; all P < 0.001) in binary classifications. Moreover, CoHI achieved a higher sensitivity (82.4% [59.0%–93.8%]) and specificity (76.9% [49.7%–91.8%]) than individual dermatologists (sensitivity 77.8% [75.3%–80.2%], specificity 73.0% [70.6%–75.4%]) and CNN (sensitivity 70.6% [46.9%–86.7%], specificity 69.2% [42.4%–87.3%]). The diagnostic accuracy of CoHI was superior to that of individual dermatologists (P < 0.001) in multiclass evaluation, with the accuracy of the latter comparable to multiclass CNN.

Conclusions: Our analysis revealed that the majority vote of an interconnected group of dermatologists (CoHI) outperformed individuals and CNN in a demanding skin lesion classification task.
using a CNN to pre-screen high volumes of imaging data and sort out high-confidence benign lesions, but using trained physicians to validate all other classifications for management decisions [14, 15].

So far, most CNN have been evaluated in competition or in cooperation with individual dermatologists. However, solving difficult tasks may benefit from collective decision making by a group of humans. There is first evidence that so-called collective human intelligence (CoHI) may outperform the diagnostic accuracy of individuals and also of deep-learning technology [16–18]. With the study presented herein, we investigated the collective and the individual diagnostic performance of humans in comparison with two state-of-the-art CNN in a skin lesion classification task.

Methods

In this cross-sectional study, 30 cases of difficult-to-diagnose skin lesions were presented to an audience of 120 dermatologists during a live conference (Thirty years of dermatoscopy conference, 22–23 November 2019, Munich, Germany). The images were an educative convenience sample including tumorous, infectious and inflammatory lesions. Accompanying case data included age, sex, brief case history, close-up images and dermatoscopic images. For each case the audience was offered six differential diagnoses in a multiple-choice format. Individual diagnostic decisions were collected via remote voting devices (quizbox®, Quizzbbox Solutions GmbH, Stuttgart, Germany). Binary classifications of dermatologists were extracted from their multiple-choice answers, for example the vote of a basal cell carcinoma was counted as “true-positive” when a melanoma case was presented. The multiple-choice response with most valid votes was defined as CoHI.

Dermatoscopic images of all cases were additionally submitted to classification by a binary (benign/malignant) and a multiclass CNN (both FotoFinder Systems GmbH, Bad Birnbach, Germany). For lesions with two dermatoscopic images available, both were sent for CNN analyses, for nail lesions only images with an upfront view of the nail plate were included.

The binary CNN (MoleAnalyzer-Pro®, FotoFinder Systems GmbH, Bad Birnbach, Germany) is market-approved and based on a modified version of a pretrained GoogleNet Inception v4 [19] architecture further trained with > 150,000 labelled dermatoscopic images [13, 20].

The multiclass CNN allowed categorization of cases into the following nine diagnostic classes: (1) melanoma, (2) nevus, (3) benign keratosis, (4) actinic keratosis, (5) squamous cell carcinoma, (6) basal cell carcinoma, (7) vascular lesion, (8) angioma and (9) dermatofibroma. Whenever the reference diagnosis (ground truth validated by histopathology) did not fit any of these diagnostic classes (for example case no. 25 depicting cutaneous sarcoidosis) the case was excluded from all multiclass evaluation (6/30 cases excluded).

The multiclass model was trained with dermatoscopic images used for the binary model but now grouped into the aforementioned nine diagnoses. Due to numerical imbalances between diagnostic classes we introduced additional data augmentation by oversampling (for example image rotation and mirroring). For multiclass predictions an additional softmax layer was applied, which distributed the weights between classes to achieve a sum of the output probabilities equal to 1. The class with the highest probability score was used as the predicted diagnostic category.

The primary outcome measure was the diagnostic accuracy (percentage of correct diagnoses) of CoHI versus individual dermatologists versus CNN in both binary and multiclass classifications. In binary classifications we additionally calculated and compared the sensitivity and specificity for each method. For individual dermatologists only those who had voted on all presented lesions could be included. Individuals with incomplete votes had to be excluded, because missing votes on presented cases could not be counted as “false” or “correct” classification without introducing bias. The overall performance of individual dermatologists was calculated as the arithmetic mean of accuracy, sensitivity, and specificity. CoHI was defined as the majority vote of dermatologists, that is, the binary or multiclass classification with most votes was regarded as the collective vote per case.

For binary classifications, actinic keratosis, Bowen’s disease and Bowenoid papulosis were regarded as malignant lesions due to their potential for malignant transformation.

Ethics: Reviewed and approved by the ethic committee of the medical faculty of the University of Heidelberg (Approval number S-629/2017).

Statistical analysis

Statistical analyses were performed separately for binary and multiclass classifications. The binary CNN produced a malignancy score ranging from 0 to 1 and a validated a priori cut-off of > 0.5 was applied for diagnosing malignancy. Pairwise comparisons of the accuracy of CoHI versus individual dermatologists and individual dermatologist versus CNN were performed by two-sided one sample t-test. The sensitivity and specificity of binary classifications of CoHI versus CNN were compared by pairwise analysis using the McNemar test [21]. In addition, binary CNN classifications were used to calculate the area under the curve (AUC) of receiver operating characteristics (ROC).

For multiclass classifications the absolute numbers and frequency of correct specific diagnoses were assessed.
Multiclass performance of CoHI versus individual dermatologists versus CNN was compared by using the two-sided one sample t-test. A two-sided P-value < 0.05 was considered statistically significant. Statistical calculations were performed with SPSS Version 25 (IBM, SPSS; Chicago, Illinois, USA).

Results

Characteristics of imaged lesions

Thirty skin lesions were collected from 18 men and twelve women with a mean (± standard deviation [SD]) age of 59 (± 22.1) years. Presented lesions (43.3 % benign, 56.7 % malignant) were localized on the face and neck (n = 8), trunk (n = 7), extremities (n = 9), soles (n = 2), nails (n = 3) and genital region (n = 1). Diagnostic categories are summarized in Table 1. Representative dermatoscopic images of four lesions from the test-set are depicted in Figure 1.

Binary classifications: Results of CoHI, individual dermatologists, and CNN

The accuracy [95 % confidence interval] of CoHI was 80.0 % [62.7 %–90.5 %], with a sensitivity of 82.4 % [59.0 %–93.8 %] and specificity of 76.9 % [49.7 %–91.8 %] (Figure 2).

The performance of individual dermatologists was assessed by including those who classified all 30 lesions (66 of 120 participants, 55 %). Exclusion of participants with incomplete votes was necessary because missing votes could not be clearly counted as false or true classifications without introducing bias. The performance of each individual of the 66 dermatologists voting on all lesions is depicted in Figure 2 by giving his or her sensitivity and specificity coordinates. Participants with full votes showed a mean sensitivity, specificity, and accuracy of 77.7 % [75.3 %–80.2 %], 73.0 % [70.6 %–75.4 %], and 75.7 % [73.8 %–77.5 %], respectively. In an alternative approach for the assessment of individual dermatologists all votes per lesion were included in the analysis. Mean results in this approach were slightly worse than for dermatologists who completed all classifications (sensitivity 74.7 %, specificity 70.7 %, and accuracy 73.0 %).

The CNN used for binary classifications attained an accuracy of 70.0 % [52.1 %–83.3 %], a sensitivity of 70.6 % [46.9 %–86.7 %], and specificity of 69.2 % [42.4 %–87.3 %]. The CNN’s ROC AUC was 0.765 [0.595–0.935]. To visualize the most important pixels for each of the CNN’s classifications we created heat maps by vanilla gradient descent backpropagation (Figure 3). For malignant lesions the CNN’s mean malignancy score was 0.69 [0.46–0.92], hence significantly higher than for benign lesions 0.34 [0.07–0.61] \((P = 0.014)\).

Binary classifications: Statistical comparison of CoHI, individual dermatologists, and CNN

The primary outcome measure was the diagnostic accuracy (percentage of correct diagnoses) of CoHI versus individual dermatologists versus CNN. CoHI showed a significantly higher accuracy (80.0 % [62.7 %–90.5 %]) than individual dermatologists (75.7 % [73.8 %–77.5 %]) and individual dermatologists showed a significantly higher accuracy than the CNN (70.0 % [52.1 %–83.3 %]; all \(P < 0.001\)).

Pairwise comparisons of the sensitivity and specificity of CoHI versus CNN revealed no significant differences (all \(P ≥ 0.687\), most probably because of the low number
Figure 2  ROC curve for the binary CNN at the a priori operation point (cut off > 0.5 for diagnosing malignancy). The sensitivity and specificity coordinates of individual dermatologists voting on all lesions (n = 66) are depicted as black dots. Mean dermatologists’ performance with corresponding standard deviations as error bars is depicted. Moreover, the sensitivity and specificity coordinates of CoHI are shown.

Figure 1  Representative lesions of the test-set. Two melanomas (a, b) and two seborrheic keratoses (c, d) were correctly diagnosed in binary classifications (benign/malignant) by CoHI.

Figure 3  Heat maps created by vanilla gradient descent backpropagation of two melanomas (a, b) and two seborrheic keratoses (c, d) with accompanying malignancy scores of the binary CNN; scores > 0.5 indicated malignancy.
of cases. Moreover, the performance of individual dermatologists was compared with the CNN by fixing the CNN’s specificity at the mean specificity of dermatologists (73.0 %). When using the corresponding cut-off as the operating point on the CNN ROC curve, the CNN’s sensitivity was 58.8 %, which was significantly lower compared with the dermatologists’ mean sensitivity of 77.7 % (P < 0.001).

Multiclass classifications

For multiclass classifications six of 30 cases had to be excluded because diagnoses did not fit any of the offered diagnostic classes. CoHI showed an accuracy of 79.2 % (19/24 correct diagnoses), individual dermatologists 64.6 % [61.6 %–67.6 %] and the CNN 62.5 % (15/24 correct diagnoses). Accuracy achieved by CoHI was significantly higher than for individual dermatologists (P < 0.001), who achieved an accuracy comparable with the multiclass CNN (P = 0.146).

Discussion

In several fields of medicine CoHI has proven an excellent diagnostic and problem-solving performance [17, 18]. Previous data also suggested that CoHI might be a promising approach for the diagnosis of skin lesions [17, 18]. Yet, to our knowledge the study presented here is the first to compare CoHI with state-of-the-art deep learning artificial intelligence (AI) algorithms in a skin lesion classification task.

Our analyses revealed a superior diagnostic accuracy for CoHI compared with individual dermatologists and two deep-learning CNN. The binary CNN achieved a sensitivity and specificity of approximately 70 %, which was markedly lower compared with previously reported sensitivities of roughly 95 % and specificities of 75 % for the same classifier [13, 20]. These differences were most probably due to differences in the level of diagnostic difficulty of the included quiz cases. In particular, the test-set of our study comprised cases of inflammatory skin diseases (for example cutaneous sarcoidosis and psoriasis), infectious diseases (molluscum contagiosum, leishmaniasis), lesions in special localizations (for example Bowen’s disease of the nail unit, melanoma of genital skin), and rare diseases such as onychopapilloma. Therefore, we had to exclude six cases for multiclass assessments to guarantee a fair comparison of the multiclass CNN with dermatologists. Excluded cases did not fit into any of the predefined diagnostic classes that were used for CNN training. This points to one important limitation of multiclass classifiers when applied in clinical routine. Rare skin diseases may not fit predefined diagnostic classes that had been used for multiclass training of a CNN, which may result in frequent misclassifications. Hence, expanding training images by number and (rare) diagnoses is of utmost relevance to improve CNN performance.

With regard to the participating dermatologists only 55 % voted on all lesions of the test-set and only those could be included for the analysis of individual dermatologists. Dermatologists with incomplete votes were excluded to neither overestimate nor underestimate human performance. Otherwise, individual dermatologists voting only on a few cases (whenever feeling highly confident) would possibly attain much better results leading to an overestimation of individual human performance. Performance of individual dermatologists varied greatly, depicted by their widely scattered operation points (sensitivity/specificity coordinates) in the figure of the CNN ROC curve. Despite the observed variations in human performance, CoHI achieved a favorable diagnostic accuracy in both binary and multiclass evaluations. This could most probably be attributed to the large size of the collective comprising 120 participating dermatologists.

To determine CoHI various approaches such as majority or quorum rules, follow-the-most-confident or most-senior rules have been defined [26, 27]. For our analyses the majority rule was used considering responses with most valid votes as collective responses. It has previously been reported that even small collectives of eight non-experts may achieve accuracies comparable with single experts [18]. Including a higher number of non-experts instead of single experts may be more realistic to implement and even more cost-effective, for instance in a teledermatology setting.

In addition, few previous studies have also quantitatively considered the physicians’ confidence for their votes. Rinner et al. reported that ratings given with low confidence, indicated by longer time intervals until sending the vote, marked reduced the overall performance of the group [18]. The authors of this study suggested that tracking time intervals until arriving at a diagnosis and excluding the slowest votes could further improve the diagnostic performance of CoHI. In line with these results, we found that individual dermatologists who did not vote on all test cases showed a lower performance in comparison with dermatologists completing all votes. It must be assumed that dermatologists with incomplete votes were less confident and thus skipped unclear cases. Hence, our analyses of individual dermatologists imply a certain bias, which probably slightly overestimated the performance of individual dermatologists.

The majority rule of CoHI as used in our study is mostly applied in crowd-based assessments, where individuals are not interconnected. Yet, there are more advanced interpretations of CoHI requiring additional software tools which imply more intense interactions of participating humans.
and which have been compared to the “swarm-intelligence” of birds and other animals [15, 23]. In this setting, swarm software-platforms allow networked individuals to interact as one intelligent group [15]. Swarms are defined by real-time decision-making with mutual interference of individuals, whereas crowds mostly allow for retrospectively collecting votes based on the aforementioned rules (such as the majority rule, most-senior rule). Thus, studies including swarm platforms for skin cancer classification are of future interest.

Currently, real-life applications of collective intelligence are still scarce. However, online exchange and interconnections are continuously becoming more important and facilitate collective decision making approaches [18]. Online platforms have been shown to attract participants for voting on classification tasks [18]. Teledermatology is increasingly being used and would allow to incorporate tools based on CoHI.

The initial enthusiasm about the high-level diagnostic performance of deep learning networks for skin lesion classification has prompted discussions concerning the future significance of human expertise. Patel et al. have previously reported that swarm-based technology and deep-learning technology in combination are capable of outperforming either method alone [15]. Thus, the future challenge is to address how humans and machines may cooperate for an unprecedented performance [28]. Tschandl et al. reported a comparable performance of the collective with CNN top predictions, while the best performance was achieved by combining AI-based multiclass probabilities and human collectives [28]. Prospectively, a CNN might provide a preselection of cases classified at high confidence without need for further validation by humans, while difficult-to-classify cases (low confidence) would require assessments for management decisions by CoHI [15].

There are several limitations to our analyses. Due to the setting of a live conference the number of test cases was limited. As a result, statistically significant differences could not be found for a number of assessments. Moreover, the amount of information provided for humans and CNN differed, since CNN analyzed only dermatoscopic images, whereas humans were allowed to review additional clinical images as well as patients’ characteristics and a brief case history. Moreover, our approach did not allow real-time interaction of physicians which is of interest for further studies.

In conclusion, we demonstrate that in a demanding skin lesion classification task CoHI outperformed the binary and multiclass performance of individual dermatologists and two state-of-the-art CNN. It seems worthwhile to further develop the framework of human and AI cooperation in consideration of CoHI.

Acknowledgement

We are thankful to all dermatologists for their voting during the conference.

Conflict of interest

HA Haenssle received honoraria and/or travel expenses from companies involved in the development of devices for skin cancer screening: Scibase AB, FotoFinder Systems GmbH, Heine Optotechnik GmbH, Magnasco GmbH. C Fink received travel expenses from Magnasco GmbH. A Blum received honoraria and/or travel expenses from Heine Optotechnik GmbH and FotoFinder Systems GmbH. All other authors state “no conflict of interest”.

Correspondence to

Holger A. Haenssle, MD
Department of Dermatology
University of Heidelberg
Im Neuenheimer Feld 440
69120 Heidelberg, Germany
E-mail: holger.haenssle@med.uni-heidelberg.de

References

1. Apalla Z, Lallas A, Sotiriou E et al. Epidemiological trends in skin cancer. Dermatol Pract Concept 2017; 7: 1.
2. Apalla Z, Nashan D, Weller RB, Castellsague X. Skin cancer: epidemiology, disease burden, pathophysiology, diagnosis, and therapeutic approaches. Dermatol ther 2017; 7: 5–19.
3. Brunssen A, Waldmann A, Eisemann N, Katalinic A. Impact of skin cancer screening and secondary prevention campaigns on skin cancer incidence and mortality: A systematic review. J Am Acad of Dermatol 2017; 76: 129–39. e10.
4. Kittler H, Pehamberger H, Wolff K, Binder M. Diagnostic accuracy of dermoscopy. Lancet Oncol 2002; 3: 159–65.
5. Zalaudek I, Argenziano G, Di Stefani A et al. Dermoscopy in general dermatology. Dermatology 2006; 212: 7–18.
6. Schweizer A, Fink C, Bertič I et al. Differentiation of combined nevi and melanomas: Case-control study with comparative analysis of dermoscopic features. J Dtsch Dermatol Ges 2020; 18: 111–8.
7. Vestergaard M, Macaskill P, Holt P, Menzies S. Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting. Br J Dermatol 2008; 159: 669–76.
8. Ahlgrim-Siess V, Weitzer F, Arzberger E et al. Diagnostic impact of reflectance confocal microscopy as a second-level examination for facial skin lesions. J Dtsch Dermatol Ges 2019; 17: 266–73.
9. Esteva A, Kuprel B, Novoa RA et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017; 542: 115–8.
10. Tschandl P, Kodell 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 Oncol 2019; 20(7): 938–47.

11 Okur E, Turkan M. A survey on automated melanoma detection. Eng Appl Artif Intell 2018; 73: 50–67.

12 Tschandl P, Rosendahl C, Akay BN et al. Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks. JAMA Dermatol 2019; 155: 58–65.

13 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; 29: 1836–42.

14 Verghese A, Shah NH, Harrington RA. What this computer needs is a physician: humanism and artificial intelligence. JAMA 2018; 319: 19–20.

15 Patel BN, Rosenberg L, Willcox G et al. Human-machine partnership with artificial intelligence for chest radiograph diagnosis. NPJ Digit Med 2019; 2: 1–10.

16 Barnett ML, Boddupalli D, Nundy S, Bates DW. Comparative accuracy of diagnosis by collective intelligence of multiple physicians vs individual physicians. JAMA Netw Open 2019; 2: e190096.

17 Kurvers RH, Krause J, Argenziano G et al. Detection accuracy of collective intelligence assessments for skin cancer diagnosis. JAMA dermatol 2015; 151: 1346–53.

18 Rinner C, Kittler H, Rosendahl C, Tschandl P. Analysis of collective human intelligence for diagnosis of pigmented skin lesions harnessed by gamification via a web-based training platform: simulation reader study. J Med Internet Res 2020; 22: e155597.

19 Szegedy C, Liu W, Jia Y, Sermanet P et al. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015: 1–9.

20 Haenssle HA, Fink C, Toberer F et al. Man against machine reloaded: performance of a market-approved convolutional neural network in classifying a broad spectrum of skin lesions in comparison with 96 dermatologists working under less artificial conditions. Ann Oncol 2020; 31: 137–43.

21 Xiang JX. On two-sample McNemar test. J Biopharm Stat 2016; 26: 217–26.

22 Kammer JE, Hautz WE, Herzog SM et al. The potential of collective intelligence in emergency medicine: pooling medical students’ independent decisions improves diagnostic performance. Med Decis Making 2017; 37: 715–24.

23 Rosenberg L, Lungren M, Halabi S et al. Artificial swarm Intelligence employed to amplify diagnostic accuracy in radiology. IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) 2018: 1186–91.

24 Wolf M, Krause J, Carney PA et al. Collective intelligence meets medical decision-making: the collective outperforms the best radiologist. PloS one 2015; 10: e0134269.

25 Kurvers RH, Herzog SM, Hertwig R et al. Boosting medical diagnostics by pooling independent judgments. Proc Natl Acad Sci USA 2016; 113: 8777–82.

26 Radcliffe K, Lyson HC, Barr-Walker J, Sarkar U. Collective intelligence in medical decision-making: a systematic scoping review. BMC Med Inform Decis Mak 2019; 19: 158.

27 Kämmer JE, Hautz WE, Herzog SM et al. The potential of collective intelligence in emergency medicine: pooling medical students’ independent decisions improves diagnostic performance. Med Decis Making 2017; 37: 715–24.

28 Tschandl P, Rinner C, Apalla Z et al. Human-computer collaboration for skin cancer recognition. Nat Med 2020; 26(8): 1229–34.