Kuo *et al.*
BMC Medical Informatics and Decision Making (2023) 23:138
https://doi.org/10.1186/s12911-023-02229-w

BMC Medical Informatics and Decision Making

RESEARCH Open Access

The accuracy of artificial intelligence used for non-melanoma skin cancer diagnoses: a meta-analysis

Kuang Ming Kuo[1], Paul C. Talley[2] and Chao-Sheng Chang[3,4*]

Abstract

Background With rising incidence of skin cancer and relatively increased mortality rates, an improved diagnosis of such a potentially fatal disease is of vital importance. Although frequently curable, it nevertheless places a considerable burden upon healthcare systems. Among the various types of skin cancers, non-melanoma skin cancer is most prevalent. Despite such prevalence and its associated cost, scant proof concerning the diagnostic accuracy via Artificial Intelligence (AI) for non-melanoma skin cancer exists. This study meta-analyzes the diagnostic test accuracy of AI used to diagnose non-melanoma forms of skin cancer, and it identifies potential covariates that account for heterogeneity between extant studies.

Methods Various electronic databases (*Scopus, PubMed, ScienceDirect, SpringerLink*, and *Dimensions*) were examined to discern eligible studies beginning from March 2022. Those AI studies predictive of non-melanoma skin cancer were included. Summary estimates of sensitivity, specificity, and area under receiver operating characteristic curves were used to evaluate diagnostic accuracy. The revised Quality Assessment of Diagnostic Studies served to assess any risk of bias.

Results A literature search produced 39 eligible articles for meta-analysis. The summary sensitivity, specificity, and area under receiver operating characteristic curve of AI for diagnosing non-melanoma skin cancer was 0.78, 0.98, & 0.97, respectively. Skin cancer typology, data sources, cross validation, ensemble models, types of techniques, pre-trained models, and image augmentation became significant covariates accounting for heterogeneity in terms of both sensitivity and/or specificity.

Conclusions Meta-analysis results revealed that AI is predictive of non-melanoma with an acceptable performance, but sensitivity may become improved. Further, ensemble models and pre-trained models are employable to improve true positive rating.

Keywords Artificial intelligence, Cross validation, Diagnostic test accuracy, Ensemble models, Meta-analysis, Non-melanoma skin cancers, Pre-trained models

*Correspondence:
Chao-Sheng Chang
zincfinger522@yahoo.com.tw
Full list of author information is available at the end of the article

Background

Skin cancer accounts for 32.5% of all diagnosed malignancies, and it has a prevalence of 7.96 million cases occurring globally each year among the general population [1]. With respect to etiology, previous studies have demonstrated a deleterious association with chronic exposure to sunlight because the ultraviolet component induces deoxyribonucleic acid damage which later triggers malignant mutations to occur. Other possible contributors to skin cancer incidence may also include viral infection, drug usage and exposure to chemicals [2].

Pathologically, skin cancer is categorized into either melanoma or non-melanoma. Albeit relatively rare, three hundred thousand annual cases of melanoma are determined as highly malignant, with a reported mortality rate of 1.6 per 100,000 worldwide [1]. By contrast, non-melanoma cases, which comprise a number of pathologically-distinct entities such as basal cell carcinoma and intra-epithelial carcinoma (i.e., actinic keratosis and Bowen's disease) [3], are less malignant considering Mohs micrographic surgery and a 5-year cure rate of 98.9% [4]. Sixty-five thousand victims die on average, per annum, worldwide due to non-melanoma incidence when combined with a delayed diagnosis factor [1]. Furthermore, non-melanoma skin cancers such as basal cell carcinoma show a trend of increasing cases [5] and are easily misdiagnosed [6]. The above-mentioned evidence clearly shows the diagnosis of non-melanoma skin cancer is of similar importance to melanoma forms of skin cancer.

Currently, clinical examination and dermoscopic evaluation are major techniques for screening skin cancers [7]. These screening techniques are estimated to achieve 75–84% of diagnosis accuracy, indicating human error may remain accountable via these approaches [8, 9]. When taking into account the high prevalence and life-threatening risk of this disease, it is important to make a timely diagnosis for appropriate treatment to follow.

Artificial intelligence (AI) techniques are being employed to provide diagnostic assistance to dermatologists since most diagnoses rely principally on visual patterning recognition [10], a particular strength of such a technology. Machine learning is a sub-field of AI which refers to an effort to automate intellectual tasks normally performed by humans; and, deep learning is in turn a subset within machine learning [11]. A veritable plethora of attempts to utilize machine learning techniques aimed at supporting the accurate diagnosis of melanoma and non-melanoma types of skin cancer have already taken place [9, 12–34]. As such, a systematic reporting is deemed necessary for reliable interpretation and aggregation of these results. However, the comparison of pre-existing skin lesion classification evidence is difficult because differences may exist in the data types used or in the statistical quantities presented [35].

Until present time, synthetic evidence regarding the performance of AI techniques applied for the diagnosis of non-melanoma skin cancer remains insufficient [7, 10]. Without reliable evidence, the application of AI in the diagnosis of non-melanoma skin cancer is frequently obstructed. Furthermore, what important factors/strategies that may influence the performance of AI in the diagnosis of non-melanoma skin cancer are at times unclear.

In viewing the unfulfilled areas of knowledge, the purposes of this meta-analysis are therefore: 1) to meta-analyze the accuracy of diagnosis for non-melanoma skin cancer via machine learning and deep learning; and, 2) to examine potential covariates that can account for the heterogeneity found among these studies. The main contributions of this study are:

- Summary of the performance of AI for diagnosing non-melanoma skin cancer with quantitative evidence so that AI's utility assessment can be made with greater efficacy and objectivity.
- Identification of potential covariates as they relate to AI performance since it may improve through an adoption of those strategies indicated by these identified covariates whenever building AI models.
- Accumulation of knowledge of diagnostic test accuracy for AI in non-melanoma skin cancer takes place so that earlier and more accurate diagnosis of non-melanoma skin cancer is practical.

The remainder of this paper is structured as follows. Related work section introduces prior reviews on the topic of diagnostic test accuracy, focusing on how these reviews were planned and evaluated. Material and methods section presents the research method adopted in this study. Results section describes the analytical findings based on collected data, Discussion section interprets and describes the significance of the findings, and Conclusions section summarizes the findings of the current study.

Related work

Up until the most recent examples, a number of studies have started to review existing evidence related to AI techniques for skin-lesion classification [7, 10, 23, 35–37]. Several themes may be observed from Table 1. First, much evidence is qualitative in nature [10, 35–37], except for the study of Sharma et al. [7] and Rajpara et al. [23]. Without quantitative evidence, the performance of AI-based predictive models are not easily or objectively assessed. Second, few reviews [7, 10] have focused solely on non-melanoma forms of skin cancer, with such efforts

Table 1 Prior reviews on skin cancer diagnosis based on artificial intelligence

Sources	Type of skin cancer	Method	AI techniques	Studies	Review component(s)	Metrics	Main findings
[23]	Melanoma	Quantitative	ML and DL	30	Comparison between dermoscopy and AI	• DOR • ±LR • Sensitivity • Specificity	• Pooled DOR of AI: 57.80 • Pooled + LR of AI: 6.36 • Pooled -LR of AI: 0.13 • Pooled sensitivity of AI = 0.91 • Pooled specificity of AI = 0.79
[35]	Both	Qualitative	DL	13	Methods of classification	• Accuracy • AUROC • Sensitivity • Specificity	• Accuracy: 0.50–0.93 • AUROC: 0.82–0.98 • Sensitivity: 0.58–0.95 • Specificity: 0.74–0.85 • CNN can be used to extract features or classify lesions • The comparison of different AI procedures is difficult
[10]	Non-melanoma	Qualitative	ML and DL	39	• Skin lesion database • Feature extraction • Methods of classification • Diagnostic accuracy	• Accuracy • AUROC	• Accuracy: 0.72–1 • AUROC: 0.83–1
[7]	Non-melanoma	Quantitative	ML and DL	57	• Skin cancer type • Algorithm type • Diagnostic standard • Data set source and size	• Sensitivity • Specificity	• Pooled sensitivity = 0.89 • Pooled specificity = 0.81 • ML algorithms for diagnosing skin cancer is promising
[36]	Both	Qualitative	ML and DL	102	• Contributions • Methods • Achieved results	• Accuracy • Precision • Sensitivity • Specificity	• Accuracy: 0.64–1 • Sensitivity: 0.35–1 • Specificity: 0.60–1 • Precision: 0.62–0.98 • The need for creation of large public image datasets • Including clinical data may help improve classification accuracy
[37]	Melanoma	Qualitative	ML and DL	53	• Data characteristics • Diagnostic techniques • Evaluation metrics	• Accuracy • AUROC • F1-score • Sensitivity • Specificity	• Accuracy: 0.67–0.99 • AUROC: 0.82–0.92 • F1-score: 0.83 • Sensitivity: 0.77–0.96 • Specificity: 0.70–0.96

AI Artificial intelligence, *AUROC* Area under receiver operating characteristic curve, *CNN* Convolutional neural network, *DL* Deep learning, *DOR* Diagnostic odds ratio, *ML* Machine learning, ± *LR* Positive/negative likelihood ratio

being devoted to the review of evidence concerning melanoma [16, 23] or both [35, 37]. By focusing exclusively on non-melanoma skin cancer, a better understanding may yet be achieved. Third, most reviews include studies that have adopted machine learning and deep learning, with the exception of Brinker et al. [35]. Despite deep learning being widely considered as having better performance than machine learning, studies that adopted machine learning should also be included in order to have a more holistic understanding of AI performance in the diagnosis of melanoma and non-melanoma skin cancers. Finally, review components/metrics for assessing the performance of AI techniques are quite diversified. Classification methods, data source, and diagnostic accuracy are primary components of these reviews. Further, reviews that followed the Preferred Reporting Items for a Systematic Review and Meta-analysis statement (PRISMA) for Diagnostic Test Accuracy (DTA) commonly reported pooled diagnostic odds ratio, pooled positive/negative likelihood ratio, pooled sensitivity, and pooled specificity, while other reviews usually reported separate accuracy, area under receiver characteristic curve, F1-score, precision, sensitivity, or specificity by individual study. This study therefore follows PRISMA-DTA for reporting summary metrics of included studies for global assessment of AI performance for the diagnosis of non-melanoma skin cancer.

Methods

This study was conducted according to the PRISMA statement [38] (see Additional file 1: Appendix A for diagnostic test accuracy checklist and Additional file 2: Appendix B for diagnostic test accuracy abstracts checklist). The Institutional Review Board of E-Da Hospital (EMRP-108–128) approved the study-wide protocol.

Search strategy and selection process

A literature search, carried out 31st March, 2022, of *Scopus, PubMed, ScienceDirect, SpringerLink,* and Dimensions, by means of keyword combinations of the terms "basal cell carcinoma", "intra-epithelial carcinoma", "Bowen's disease", "actinic keratosis", "skin lesion", "non-melanoma skin cancer", "artificial intelligence", "machine learning", and "deep learning".

Inclusion criteria was determined by: 1) studies investigating the accuracy of non-melanoma skin cancer; 2) studies written in English; and, 3) studies adopting machine-learning or deep-learning techniques. Studies were dis-qualified for inclusion, if: 1) they only investigated the incidence of melanoma skin cancer; 2) studies were irrelevant to our research purpose; and, 3) full texts were unavailable for purposes of examination. We located 134 potentially eligible articles, of which 95 were excluded with reason (see Fig. 1), and the remaining 39 articles being included in the eventual quantitative meta-analysis that was made.

Data extraction

From each study, we extracted the following information: Authorship, publication year, sample size, types of non-melanoma skin cancer described, whether data sources were publicly available, whether cross-validation procedures were undertaken, whether ensemble models were employed, and what type of artificial intelligence technique was employed (i.e., deep learning or machine learning). Only studies that adopted a neural network algorithm with more than one hidden layer were categorized as being part of the deep learning group, with others categorized as being part of the machine learning group for purposes of our study. For models based on deep learning, further recorded information including whether pre-trained models were utilized and whether image augmentation was implemented. Further, we extracted the original numbers of *true/false* positives and *true/false* negatives from each study to derive outcome measures, including summary sensitivity, specificity, and area under receiver operating characteristic curve, for purposes of diagnostic accuracy. Finally, if an article had classified more than one non-melanoma skin cancer simultaneously, we considered each of the non-melanoma skin cancers as a different study, with relevant data extracted based upon the above-listed procedures.

Methodological analysis

Regarding the quality of each of the included studies, we evaluated the risk of bias and applicability in accordance with the revised Quality Assessment of Diagnostic Studies (QUADAS-2) including four domains: sample selection, index test, reference standard, flow, and timing [30].

Statistical analysis

Following the suggestion of prior evidence [39], sensitivity and specificity were pooled with a bivariate model. Area under receiver operating characteristic

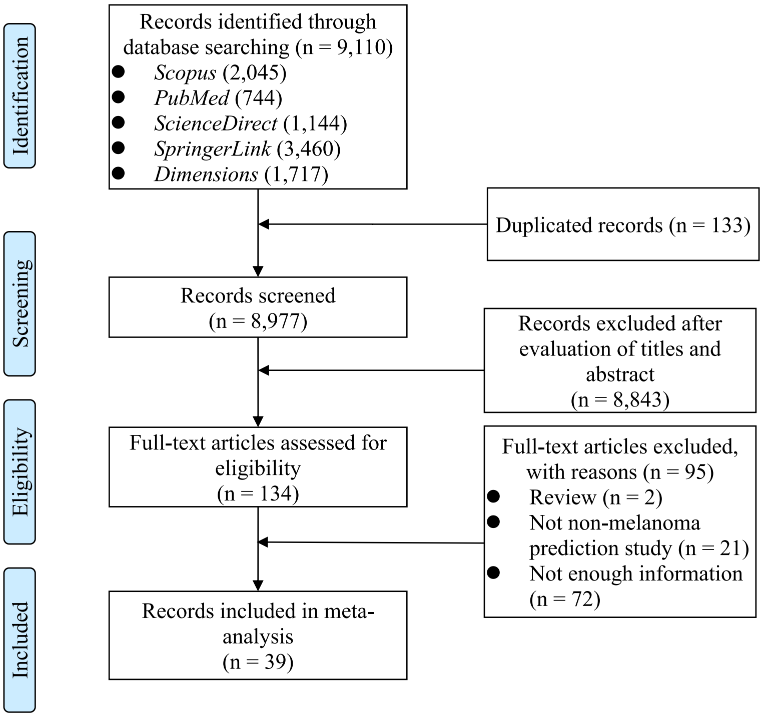
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 9,110)
● *Scopus* (2,045)
● *PubMed* (744)
● *ScienceDirect* (1,144)
● *SpringerLink* (3,460)
● *Dimensions* (1,717)
Duplicated records (n = 133)
Records screened (n = 8,977)
Records excluded after evaluation of titles and abstract (n = 8,843)
Full-text articles assessed for eligibility (n = 134)
Full-text articles excluded, with reasons (n = 95)
● Review (n = 2)
● Not non-melanoma prediction study (n = 21)
● Not enough information (n = 72)
Records included in meta-analysis (n = 39)

Fig. 1 Article selection process

curve, diagnostic odds ratio, positive likelihood ratio, and negative likelihood ratio were also estimated in this study. Forest plots were produced to depict variability amongst the studies up for consideration. Besides, summary receiver operating characteristic curves with 95% confidence intervals (CI) and 95% prediction intervals (PI) were adopted to assess the existence of a threshold effect among the included studies [40]. The R statistics [41] with lme4 [42] and mada [43] packages were used for diagnostic accuracy test meta-analysis.

Several meta-regressions with plausible covariates, including types of non-melanoma skin cancer (i.e., basal cell carcinoma and intra-epithelial carcinoma), whether data sources were publicly available (public or proprietary), whether cross-validation procedures were undertaken, whether ensemble models were adopted, types of AI technique employed (machine learning or deep learning), whether pre-trained deep learning models (e.g., *DenseNet, ResNet*, or *AlexNet*) were used (*Yes* or *No*), and whether image augmentation procedures were used by deep learning models (*Yes* or *No*) were undertaken to check for possible heterogeneity among studies. The significance level is set to 0.05 for present study.

Results

General study characteristics

Among the 39 included articles, 13 articles [6, 19, 20, 24, 26, 29, 32, 44–49] reported the identification of only one type of non-melanoma skin cancer, while other 26 articles simultaneously reported the identification of more than one non-melanoma skin cancers. Totally, 67 studies were included in our meta-analysis of diagnostic test accuracy.

Half of the included studies aimed to identify images of basal cell carcinoma (56.72%) and then intra-epithelial carcinoma (43.28%). About 76% of datasets were publicly available. Cross-validation procedures were adopted by near 22% of included studies, while approaches of ensemble models were employed by about 16%. Deep learning techniques (76%) were adopted more frequently than machine learning techniques (19%). Among 51 studies adopting deep learning, about 73% of these utilized pre-trained models for prediction, while nearly 65% of studies employed image augmentation techniques when building predictive models. Detailed characteristics of the inclusive studies is given in Tables 2 and 3, respectively.

Quality assessment

According to the QUADAS-2 tool [30], 18 out of the 39 included articles were assessed to have a moderate risk of bias, and 21 articles were considered as possessing a low risk of bias regarding patient selection. Three, four, and three articles were assessed as having a moderate risk factor related to index test, reference standard, and flow and timing, respectively (see Fig. 2). Furthermore, 18 out of 39 included articles purporting a moderate risk are regarding the applicability of patient selection processes. Further, 38 articles were categorized as having a low applicability risk about the given index test and reference standard applied.

Diagnostic accuracy of non-melanoma skin cancer predictive models based on machine learning

As shown in Table 4, the overall summary area under receiver operating characteristic curve for machine learning to predict non-melanoma skin cancers is about

Table 2 Characteristics of included studies

Characteristics	Values	Frequency	%
Study type (*n* = 67)	Basal cell carcinoma	38	56.72
	Intra-epithelial carcinoma	29	43.28
Data sources (*n* = 67)	Proprietary	16	23.88
	Public	51	76.12
Cross validation (*n* = 67)	No	52	77.61
	Yes	15	22.39
Ensemble models (*n* = 67)	No	56	83.58
	Yes	11	16.42
Types of technique (*n* = 67)	Deep learning	51	76.12
	Deep learning + Machine learning	3	4.48
	Machine learning	13	19.40
Pre-trained models (*n* = 51)	No	14	27.45
	Yes	37	72.55
Image augmentation (*n* = 51)	No	18	35.29
	Yes	33	64.71

Table 3 Attributes of included articles

Studies	Study type	Data sources	Cross validation	Ensemble models	Type of techniques	Pre-trained models	Image augmentation	Number of lesions	Reference standards	Techniques
[44]	BCC	Public	Y	N	DL	Y	Y	395	Pathological examination	Dermoscopic images
[50]	BCC + IEC	Public	N	N	DL	Y	N	200	Pathological examination	Dermoscopic images
[51]	BCC + IEC	Public	Y	N	DL	N	Y	4766	Pathological examination	Dermoscopic images
[52]	BCC + IEC	Public	N	N	DL	Y	Y	1981	Pathological examination	Dermoscopic images
[53]	BCC + IEC	Public	N	N	DL	Y	Y	2003	Pathological examination	Dermoscopic images
[13]	BCC + IEC	Proprietary	Y	N	ML	N	N	960	Clinical and pathological examinations	Dermoscopic images
[54]	BCC + IEC	Public	N	N	DL	N	Y	5066	Pathological examination	Dermoscopic images
[55]	BCC + IEC	Public	N	N	DL	Y	Y	1502	Pathological examination	Dermoscopic images
[56]	BCC + IEC	Public	N	N	DL	Y	N	589	Pathological examination	Dermoscopic images
[45]	IEC	Public	N	N	DL	N	Y	1200	Pathological examination	Dermoscopic images
[46]	BCC	Proprietary	N	N	DL	Y	Y	180	Unclear	Dermoscopic images
[57]	BCC + IEC	Public	N	N	DL	Y	N	1536	Unclear	Unclear
[47]	BCC	Proprietary	N	N	DL	Y	N	825	Pathological examination	Digital camera
[58]	BCC + IEC	Public	N	N	DL	Y	Y	604	Pathological examination	Dermoscopic images
[59]	BCC + IEC	Proprietary	Y	N	ML	N	N	78	Histopathology	Optical coherence tomography
[20]	BCC	Proprietary	N	N	DL	N	N	450	Histopathology	Dermoscopic images
[19]	BCC	Proprietary	Y	Y	ML	N	N	664	Expert opinions	Dermoscopic images
[60]	BCC + IEC	Public	N	Y	DL	Y	Y	1006	Pathological examination	Dermoscopic images
[61]	BCC + IEC	Public	Y	Y	DL + ML	Y	N	1002	Pathological examination	Dermoscopic images
[48]	BCC	Proprietary	Y	N	ML	N	N	520	Histology	Optical coherence tomography
[62]	BCC + IEC	Public	Y	N	DL	Y	Y	2003	Pathological examination	Dermoscopic images
[63]	BCC + IEC	Public	N	N	ML	N	N	480	Pathological examination	Dermoscopic images
[64]	BCC + IEC	Public	N	Y	DL	Y	Y	25,331	Pathological examination	Dermoscopic images
[65]	BCC + IEC	Public	N	N	DL	Y	Y	552	Pathological examination	Dermoscopic images
[66]	BCC + IEC	Public	N	N	DL	Y	N	1002	Pathological examination	Dermoscopic images
[67]	BCC + IEC	Public	N	N	DL	Y	Y	2003	Pathological examination	Dermoscopic images
[68]	BCC + IEC	Public	Y	Y	DL	Y	Y	1642	Pathological examination	Dermoscopic images
[69]	BCC + IEC	Public	N	N	DL	N	N	938	Pathological examination	Dermoscopic images
[49]	BCC	Public	N	Y	DL + ML	N	Y	4817	Pathological examination	Dermoscopic images
[70]	BCC + IEC	Public	N	N	DL	N	Y	802	Pathological examination	Dermoscopic images
[24]	BCC	Proprietary	Y	Y	ML	N	N	964	Unclear	Dermoscopic images
[26]	IEC	Proprietary	N	N	ML	N	N	373	Expert opinions	Photographic images
[27]	BCC + IEC	Public	N	N	ML	N	N	3004	Pathological examination	Dermoscopic images

Table 3 (continued)

Studies	Study type	Data sources	Cross validation	Ensemble models	Type of techniques	Pre-trained models	Image augmentation	Number of lesions	Reference standards	Techniques
[71]	BCC + IEC	Public	N	N	DL	Y	Y	2019	Pathological examination	Dermoscopic images
[29]	BCC	Public	N	N	ML	N	N	56	Pathological examination	Dermoscopic images
[72]	BCC + IEC	Public	N	N	DL	N	Y	2153	Pathological examination	Dermoscopic images
[31]	BCC + IEC	Proprietary	N	N	DL	Y	N	2909	Histopathology	Dermoscopic images
[32]	BCC	Proprietary	Y	N	DL	Y	N	507	Dermatologists	Dermoscopic images
[6]	BCC	Proprietary	N	N	DL	Y	N	1150	Dermatologists	Dermoscopic images

BCC Basal cell carcinoma, *IEC* Intra-epithelial carcinoma, *DL* Deep learning, *ML* Machine learning

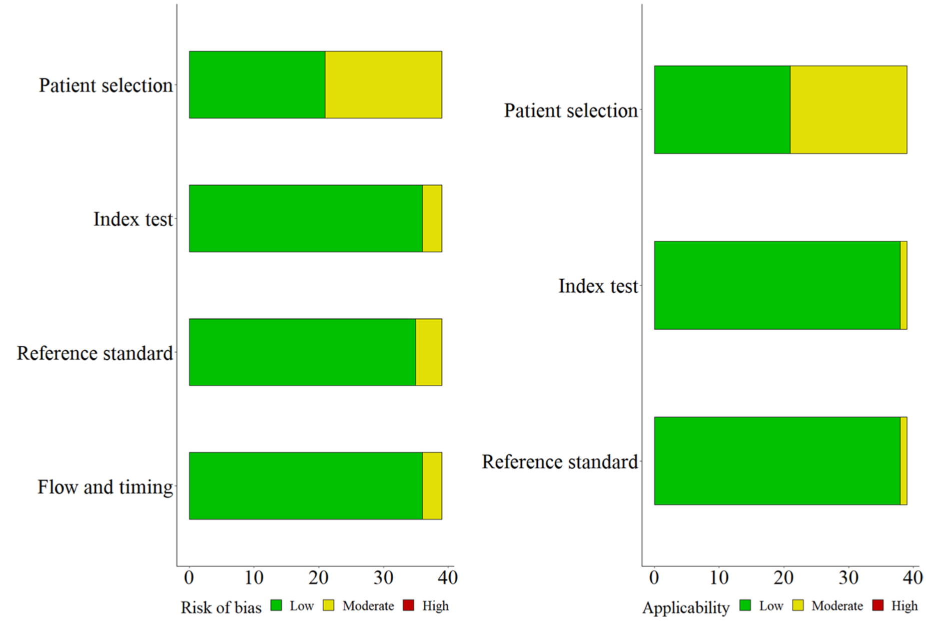

Fig. 2 Methodological assessment by QUADAS-2

0.97. Moreover, summary sensitivity, specificity, diagnostic odds ratio, positive likelihood ratio, and negative likelihood ratio were 0.78, 0.98, 224.31, 49.75, and 0.22 respectively. Since our study focused on non-melanoma skin cancers, Figs. 3 and 4 illustrate the sensitivity and specificity of each of the included studies for basal cell carcinoma and intra-epithelial carcinoma, respectively.

Pooling sensitivity and specificity based on univariate meta-analysis of diagnostic test accuracy can sometimes be misleading [40]. The bivariate meta-analysis of summary sensitivity, specificity, and their respective confidence intervals is depicted in Table 5. Figure 5 illustrates the summary ROC curve with 95% CI and 95% PI for included studies. We conducted χ^2 tests to evaluate the heterogeneity of sensitivity and specificity. The test for equality of sensitivity and of specificity were both significant, χ^2 (66) = 2702.33, $p < 0.001$ and χ^2 (66) = 12,373.02, $p < 0.001$, indicating significant between-study heterogeneity existed for both sensitivities and specificities.

As shown in Table 5 and Fig. 6(a), the sensitivity was higher for the studies classifying basal cell carcinoma (0.83; 95% CI, 0.77–0.88) than for the studies classifying intra-epithelial carcinoma (0.70; 95% CI, 0.53–0.82), and a statistically significance was evident ($p = 0.046$). The corresponding specificity was close between basal cell carcinoma and intra-epithelial carcinoma (0.98; 95% CI, 0.96–0.98 vs. 0.99; 95% CI, 0.98–0.99), but a significant difference was still detected ($p = 0.006$).

Studies utilized public data sources, as displayed in Table 5 and Fig. 6(b) achieved a higher sensitivity (0.79, 95% CI, 0.69–0.86 vs. 0.78, 95% CI, 0.69–0.86) than those used proprietary data sources, but didn't reach statistical significance ($p = 0.686$). Specificity was however significantly higher for studies which adopted public data sources (0.99; 95% CI, 0.98–0.99) than proprietary data sources (0.93; 95% CI, 0.88–0.97, $p < 0.001$).

As presented in Table 5 and Fig. 6c), studies that implemented cross-validation procedures during the training phase had a higher sensitivity (0.79; 95% CI, 0.64–0.89 vs. 0.78; 95% CI, 0.69–0.85) than studies that did not, but no statistical significance was observed ($p = 0.380$). However, studies that implemented cross validation showed a significant lower specificity than studies that did not (0.97; 95% CI, 0.92–0.98 vs. 0.99; 95% CI, 0.98–0.99, $p = 0.008$).

Studies that adopted ensemble models, as depicted in Table 5 and Fig. 6(d), had a significant higher sensitivity (0.91; 95% CI, 0.84–0.95 vs. 0.75; 95% CI, 0.66–0.82, $p = 0.014$), and a tied specificity (0.98; 95% CI, 0.94–0.99 vs. 0.98; 95% CI, 0.98–0.99).

The studies that used deep-learning techniques, as showed in Table 5 and Fig. 6(e), achieved a lower sensitivity (0.75; 95% CI, 0.66–0.83 vs. 0.83; 95% CI, 0.69–0.91) and a higher specificity (0.99; 95% CI, 0.98–0.99 vs. 0.95; 95% CI, 0.88–0.98) than the studies that used machine-learning techniques. Significant differences were however only found in specificity ($p = 0.001$) but not sensitivity ($p = 0.344$).

As depicted in Table 5 and Fig. 6(f), deep learning studies that adopted pre-trained models achieved a higher sensitivity (0.81; 95% CI, 0.73–0.87 vs. 0.58; 95% CI,

Table 4 Performance of predicting non-melanoma by artificial intelligence

Metrics	Performance (95% CI)
Area under receiver operating characteristic curve	0.97
Sensitivity	0.78 (0.71, 0.84)
Specificity	0.98 (0.98, 0.99)
Diagnostic odds ratio	224.31 (123.15, 325.46)
Positive likelihood ratio	49.75 (33.94, 65.56)
Negative likelihood ratio	0.22 (0.15, 0.29)

CI Confidence interval

0.33–0.79) and a higher specificity (0.99; 95% CI, 0.98–0.99 vs. 0.98; 95% CI, 0.97–0.99) than studies that did not. Statistical significant difference was only established for sensitivity ($p=0.048$) rather than specificity ($p=0.263$).

Table 5 and Fig. 6(g) showed that the sensitivity of the deep learning studies that adopted image-augmentation procedures to deal with over-fitting was lower than the studies without adopting image-augmentation procedures (0.73; 95% CI, 0.62–0.82 vs. 0.80; 95% CI, 0.61–0.91), but no statistical difference was observed ($p=0.413$). There was however a statistically significant difference between studies adopted image-augmentation procedures and studies that did not (0.99; 95% CI, 0.99–0.99 vs. 0.98; 95% CI, 0.96–0.99, $p=0.031$).

Discussion

Considering the potential health threat posed by non-melanoma skin cancer and the advances of AI techniques made in recent years, it is timely to both synthesize and report the results in diagnosis of non-melanoma skin cancers based on AI. According to the included studies, our meta-analysis shows there to be diagnostic performance with a moderate sensitivity, a strong specificity and AUC. Sensitivity was significantly dependent on the types of non-melanoma skin cancer encountered and whether ensemble models were adopted, while specificity was significantly dependent on included covariates except for whether ensemble models were adopted and whether pre-trained models were used.

Prior studies [8, 9] show that the accuracy for diagnosing non-melanoma skin cancers based on AI is closer to human diagnosis based on sensitivity. It is also notable that the performance of AI is better than that based on human diagnosis in terms of specificity and AUC [8, 9]. However, the pooled sensitivity in this study is lower than the pooled sensitivity in diagnosing melanoma skin

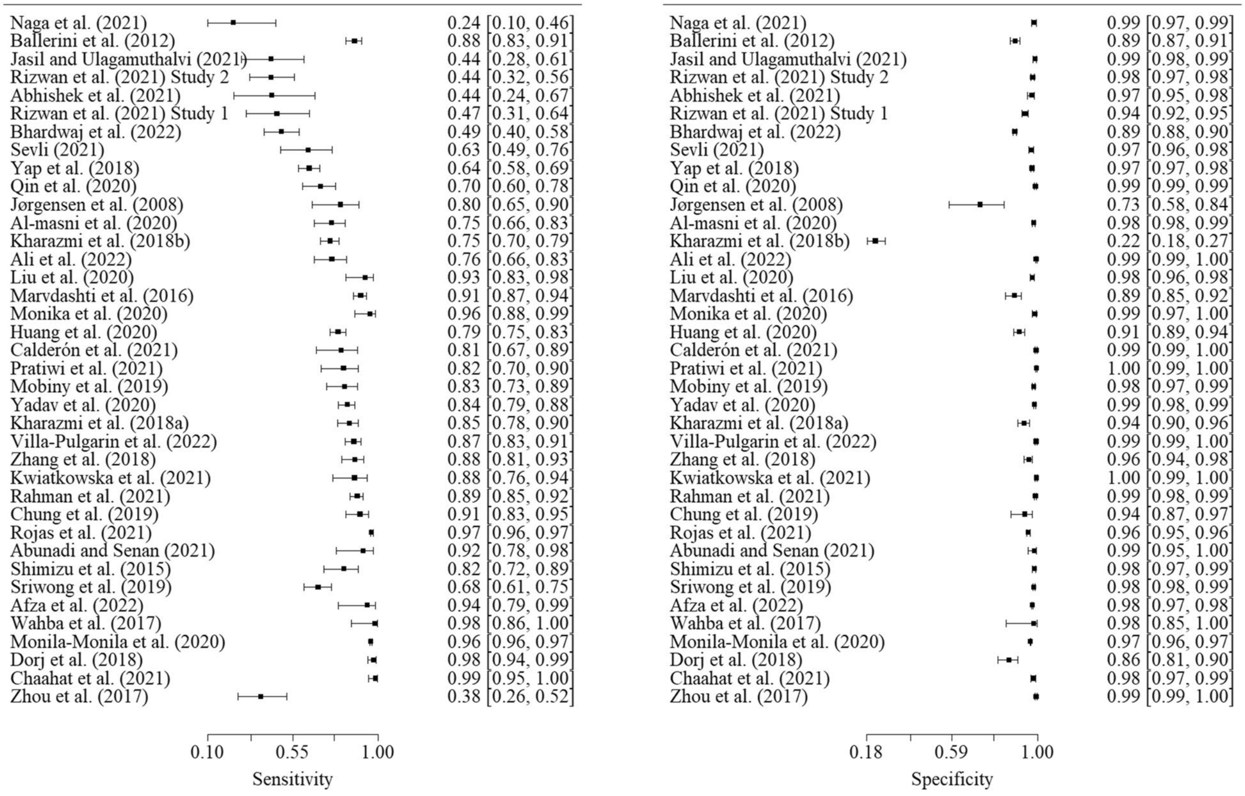

Fig. 3 Sensitivity and specificity of basal cell carcinoma studies ($n=38$)

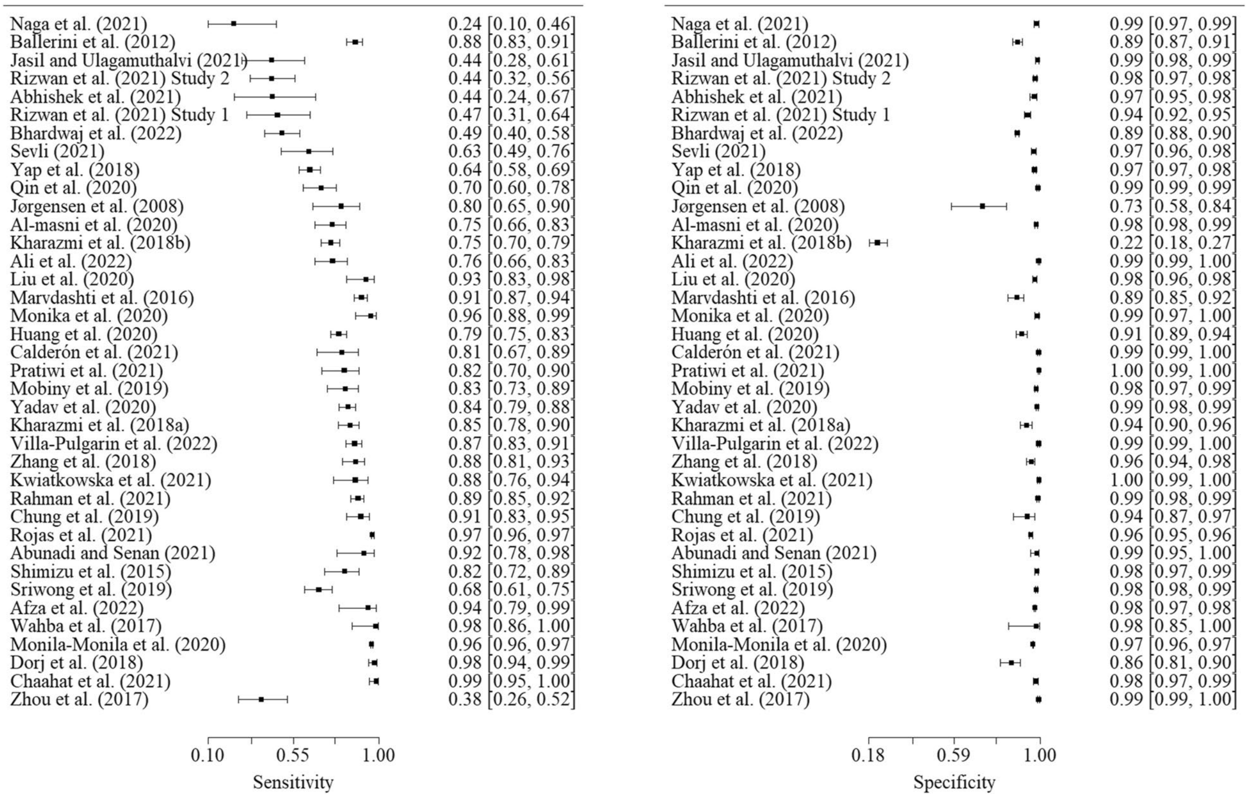

Fig. 4 Sensitivity and specificity of intra-epithelial carcinoma studies (*n* = 29) Plausible covariates to explain between-study heterogeneity

cancer based on AI. More studies are required to improve the performance of AI used to diagnose non-melanoma skin cancer, especially in terms of sensitivity. Significant covariates identified in this study may be included for purposes of this improvement.

In terms of differing types of non-melanoma skin cancer, the sensitivity of basal cell carcinoma was higher than those found in intra-epithelial carcinoma. Significant difference between basal cell carcinoma and intraepithelial carcinoma was observed. The potential reason that basal cell carcinoma reached a significant higher sensitivity may be due to more images of this type of skin cancer are included and thus more characteristics are learned than the other type of skin cancer (10,363 vs. 3,829). The specificity of basal cell carcinoma was lower than, but close to, that of intra-epithelial carcinoma. In words, there are few false positive results when using artificial intelligence to detect non-melanoma skin cancers.

Data for building machine-learning models that are collected from public data sources, in general, are more frequently leveraged than proprietary data sources in the present study. It is therefore reasonable that predictive models based on public data sources often perform better than those based on proprietary data sources since data from public sources are better understood than data from proprietary sources. Our study confirmed studies that data used from public sources performs better in terms of both sensitivity and specificity than those of data used from proprietary sources. This finding was also reported by Dick et al. [16], but their study focused on meta-analyzing the computer-aided diagnoses of melanoma skin cancers. The finding reported here suggests that the performance of predictive models using proprietary data sources may still have room for improvement by leveraging differing machine-learning or deep-learning techniques. Further, echoing the point of Tschandl et al. [3], more skin cancer images of all pertinent kinds should be collected and, most importantly, be made publicly available.

In order to prevent over-fitting [73], it is widely acknowledged to split collected data into a training dataset and a test one. The training dataset is used for training the predictive model while the test dataset is used to validate the established model [74]. To better estimate the model performance, cross- validation methods focused on the training dataset are usually adopted [73]. In this research, studies that adopted cross validation showed a non-significant, higher sensitivity but a significant, lower specificity than those that did not. One possible explanation for studies adopting cross-validation

Table 5 Summary estimates for sensitivity and specificity

Covariates	Values	Metrics	Summary estimates	95% CI	*p* value
Overall (*n* = 67)		Sensitivity	0.78	[0.71, 0.84]	
		Specificity	0.98	[0.98, 0.99]	
Types of non-melanoma skin cancer	Basal cell carcinoma (*n* = 38)	Sensitivity	0.83	[0.77, 0.88]	[Reference]
		Specificity	0.98	[0.96, 0.98]	[Reference]
	Intra-epithelial carcinoma (*n* = 29)	Sensitivity	0.70	[0.53, 0.82]	0.046*
		Specificity	0.99	[0.98, 0.99]	0.006**
Data sources	Public (*n* = 51)	Sensitivity	0.79	[0.69, 0.86]	[Reference]
		Specificity	0.99	[0.98, 0.99]	[Reference]
	Proprietary (*n* = 16)	Sensitivity	0.78	[0.69, 0.86]	0.686
		Specificity	0.93	[0.88, 0.97]	0.000***
Cross validation	Yes (*n* = 15)	Sensitivity	0.79	[0.64, 0.89]	[Reference]
		Specificity	0.97	[0.92, 0.98]	[Reference]
	No (*n* = 52)	Sensitivity	0.78	[0.69, 0.85]	0.380
		Specificity	0.99	[0.98, 0.99]	0.008**
	Yes (*n* = 11)	Sensitivity	0.91	[0.84, 0.95]	[Reference]
		Specificity	0.98	[0.94, 0.99]	[Reference]
Ensemble models	No (*n* = 56)	Sensitivity	0.75	[0.66, 0.82]	0.014**
		Specificity	0.98	[0.98, 0.99]	0.510
Types of technique	Deep learning (*n* = 51)	Sensitivity	0.75	[0.66, 0.83]	[Reference]
		Specificity	0.99	[0.98, 0.99]	[Reference]
	Machine learning (*n* = 13)	Sensitivity	0.83	[0.69, 0.91]	0.344
		Specificity	0.95	[0.88, 0.98]	0.001**
Pre-trained models	Yes (*n* = 36)	Sensitivity	0.81	[0.73, 0.87]	[Reference]
		Specificity	0.99	[0.98, 0.99]	[Reference]
	No (*n* = 15)	Sensitivity	0.58	[0.33, 0.79]	0.048*
		Specificity	0.98	[0.97, 0.99]	0.263
Image augmentation	Yes (*n* = 34)	Sensitivity	0.73	[0.62, 0.82]	[Reference]
		Specificity	0.99	[0.99, 0.99]	[Reference]
	No (*n* = 17)	Sensitivity	0.80	[0.61, 0.91]	0.413
		Specificity	0.98	[0.96, 0.99]	0.031*

* denotes $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, and *CI* Confidence interval

methods showed a significant, lower specificity that may contribute to different types of cross-validation methods. This explanation included threefold [44], fivefold [62, 68], tenfold [19, 24, 32, 51], or leave-one-out [13, 48, 59] that were adopted as part of these studies. As such, they may have caused a larger variation in specificity, as illustrated in Fig. 6c.

Ensemble models are predictive models that combine predictions from more than one model into one model [75]. They tend to have much better predictive performance than singular models [74]. It is therefore reasonable to assume predictive models based on ensemble models perform better than single models. Our study confirmed that studies based on ensemble models performed better in sensitivity (0.91 vs. 0.75) and a tied specificity (0.98 vs. 0.98) between two categories, a statistically significant difference was found only in sensitivity. This finding may indicate that ensemble models can be considered when utilizing AI for predicting non-melanoma skin cancers.

The summary sensitivity and specificity is (0.75, 0.99) and (0.83, 0.95) when deep-learning and machine-learning techniques were used, respectively; however, statistical significance difference was evident only in specificity. As illustrated in Table 4 and Fig. 6 (e), the 95% CIs of sensitivity of deep learning and machine learning are quite wide, indicating that where the real sensitivity lies is unsure, while the 95% CIs of specificity is quite narrow, suggesting the summary specificity is a more precise estimation. Regarding why the sensitivity of machine learning is higher than that of deep learning, we suspect the plausible reason may be due to most studies being based on a machine learning adopted support vector machine [26, 27, 29, 48, 59, 63] or random forest [19, 61]. The machine learning

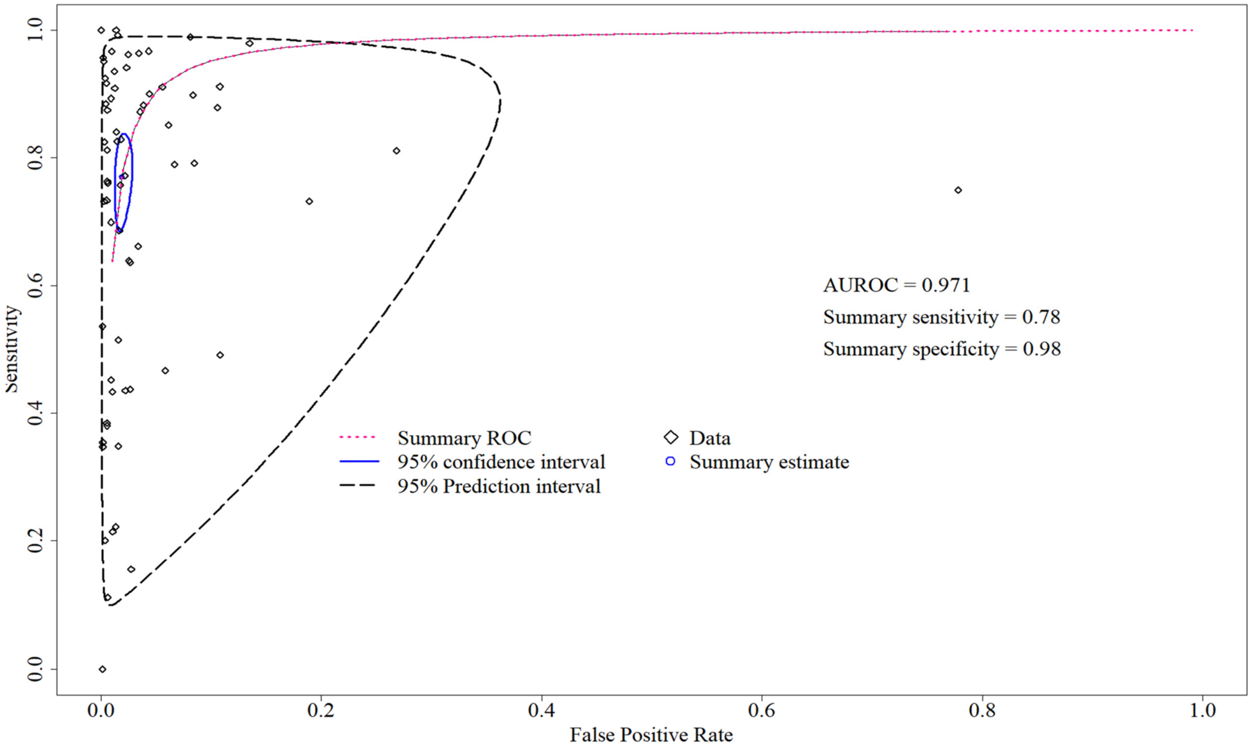

Fig. 5 Summary receiver operating characteristic curve for overall studies

performances are generally considered as quite stable, while studies based on deep-learning techniques adopted a variety of differing techniques which may result in more unstable performance. Prior meta-analyses of diagnostic test accuracy of AI used for diagnosing melanoma skin cancers, by Dick et al. [16] and Rajpara et al. [23], showed summary sensitivities and specificities of (0.44, 0.92) and (0.91, 0.79), respectively. We however found that both deep-learning and machine-learning techniques performed moderately well with sensitivity and quite well with specificity. These inconsistent results are explainable by the fact that these meta-analyzed studies were not conducted in similar clinical settings. Further, both prior meta-analyses [16, 23] focused on melanoma rather than non-melanoma skin cancer types respective this study.

Pre-trained models are generalizable deep neural networks which are trained on large-scale data and can be transferred to many tasks [76]. Simply fine-tuning these pre-trained models with a small amount of task-related data can influence performance quite well [77]. Prior review evidence [36] also found pre-trained models outperformed experienced dermatologists. As such, deep-learning studies based on pre-trained models are expected to perform better than those that did not, as confirmed in this study. Both sensitivity and specificity of pre-trained models performed higher than those that did not, but only sensitivity reached statistical significance. This finding may imply that future studies may consider building predictive models based on pre-trained models to enhance predictive accuracy.

In order to prevent over-fitting, deep learning usually required large amounts of data for purposes of training [78]. Without sufficient training data, the generalization of predictive models may be limited [79]. Based on this notion, studies that adopted image-augmentation techniques are expected to perform better than those did not. Our study however showed that sensitivity is lower for studies that adopted image augmentation than those did not, while specificity is higher for studies adopted image augmentation than those that did not. Only specificity reached a statistically significant difference. One plausible reason for why studies adopted image augmentation achieved lower sensitivity than studies that did not is that the implemented-augmentation techniques were different among the included studies. Not every included studies utilized the same augmentation techniques, such as flipping, rotation, shearing, cropping, and translation.

Finally, our study meta-analyzed the diagnostic test accuracy based on PRISMA statement [38]. Common metrics for diagnostic test accuracy including area under receiver operating characteristic curve, sensitivity, specificity, diagnostic odds ratio, positive likelihood ratio and

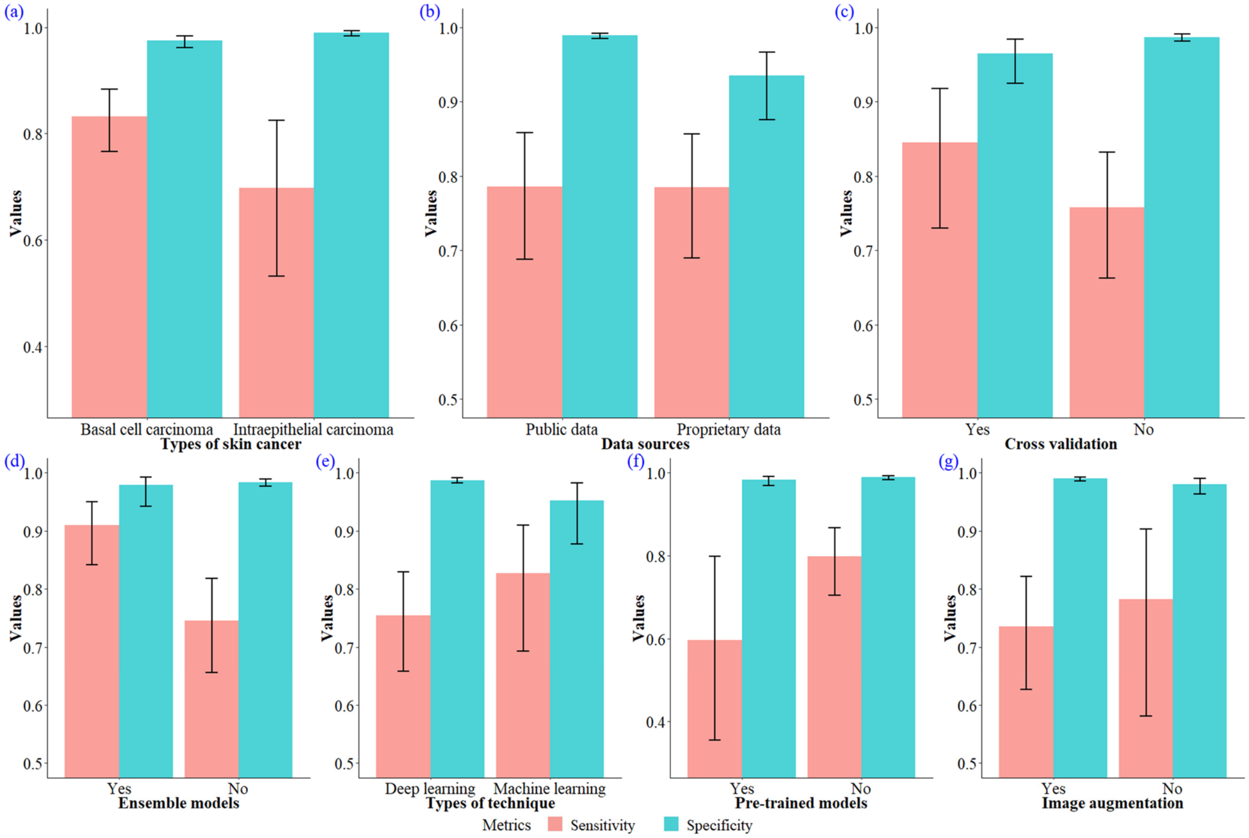

Fig. 6 Summary sensitivity and specificity with 95% confidence interval for different covariates

negative likelihood ratio were included. Furthermore, to account for the threshold effect, the pooled sensitivity and specificity was estimated based on a bivariate model [39]. Other metrics such as mean accuracy were not assessed in this study since prior evidence suggests that sensitivity and specificity are more sensible parameters to be analyzed in a meta-analysis, and they are clinically well known [80].

Just like most meta-analyses, our study has its limitations. First, the interpretation of summary sensitivity and specificity should be approached cautiously since heterogeneity among studies exists. Further, 72 studies were excluded due to insufficient quantitative information. Future diagnostic studies aimed at predicting non-melanoma skin cancers are suggested to include sufficient quantitative information for subsequent meta-analysis to better characterize and profile these studies. The covariates identified in this study are purely based from a statistical viewpoint [81], future research could consider the different design ideas of deep learning-based approaches or machine learning-based approaches to identify the incidence of other potential covariates. Finally, future meta-analysis may adopt emerging techniques [82–85] to cluster or classify models into different groups or categories, so that different insights are obtainable.

Conclusions

Our study aims to meta-analyze the diagnostic test accuracy of applying AI techniques to the diagnosis of non-melanoma type skin cancer which is already considered insufficient in review evidence. Without a better understanding of the performance of AI for the diagnosis of non-melanoma skin cancer, the potential of AI may not be fully realized. Furthermore, the results of this quantitative meta-analysis can provide a more objective synthesis of the AI performance for diagnosing non-melanoma skin cancer. Based on the findings of this study, the usefulness of AI can be assessed with greater facility and objectivity. Moreover, strategies for improving the performance of AI used for screening non-melanoma skin cancer are identifiable. A quick, safe, and non-invasive screening of non-melanoma skin cancers can thus be expected. By searching multiple online databases, 39 articles (67 studies) were included for purposes of meta-analysis. A bivariate meta-analysis of diagnostic test accuracy was undertaken to obtain summary sensitivity, specificity, and AUC. A moderate diagnostic performance of summary sensitivity, a strong summary specificity, and a strong AUC were all observed based according to a bivariate meta-analysis of diagnostic accuracy test. Types of non-melanoma skin cancer, whether data sources were publicly available, whether cross-validation procedures were undertaken, whether ensemble models were adopted, the types of AI technique employed, whether pre-trained deep-learning models were used, and whether image-augmentation procedures were all determined to partially explain some of the heterogeneity found among primary studies. Future studies may consider adopting the suggested techniques to have better predictive performance of AI for the effective diagnosis of non-melanoma skin cancer.

Abbreviations

AI	Artificial intelligence
AUC	Area under receiver operating characteristic curve
BCC	Basal cell carcinoma
CI	Confidence interval
CNN	Convolutional neural network
DOR	Diagnostic odds ratio
DL	Deep learning
DTA	Diagnostic test accuracy
IEC	Intra-epithelial carcinoma
±LR	Positive/negative likelihood ratio
ML	Machine learning
ROC	Receiver operating characteristic curve
PRISMA	Preferred reporting items for a systematic review and meta-analysis statement
PI	Prediction interval
QUADAS-2	Revised quality assessment of diagnostic studies

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-023-02229-w.

Additional file 1.

Additional file 2.

Acknowledgements

We want to show gratitude to Mr. Yao-Kun Cheng for helping data application and collection.

Authors' contributions

KMK and CSC conceived of this study and participated in the design and administration of the study. KMK and CSC drafted the manuscript and performed the statistical analysis. PCT reviewed and substantively revised the manuscript. All authors read and approved the final manuscript.

Funding

This study has been supported by E-Da Hospital, Kaohsiung, Taiwan under grant number EDAHP110022 and ISU-108-IUC-04. The funder had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. This article does not contain any studies with animals performed by any of the authors. The Institutional Review Board of E-Da Hospital waived their mandate for informed consent regarding this study (IRB No. EMRP-108–128).

Consent for publication

The manuscript does not contain any individual's data in any form.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Business Management, National United University, No.1, Miaoli 360301 Lienda, Taiwan, Republic of China. [2]Department of Applied English, I-Shou University, No. 1, Sec. 1, Syuecheng Rd., Dashu District, 84001 Kaohsiung City, Taiwan, Republic of China. [3]Department of Occupational Therapy, I-Shou University, No. 1, Yida Rd., Yanchao District, 82445 Kaohsiung City, Taiwan, Republic of China. [4]Department of Emergency Medicine, E-Da Hospital, I-Shou University, Kaohsiung, Taiwan, Republic of China.

Received: 15 December 2022 Accepted: 7 July 2023
Published online: 28 July 2023

References

1. Global Burden of Disease Cancer Collaboration. Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-Years for 29 Cancer Groups, 1990 to 2017: A Systematic Analysis for the Global Burden of Disease Study. Jama Oncol. 2019;5(12):1749–68.
2. Koh HK, Geller AC, Miller DR, Grossbart TA, Lew RA. Prevention and Early Detection Strategies for Melanoma and Skin Cancer: Current Status. Arch Derm. 1996;132(4):436–43.
3. Tschandl P, Codella N, Akay BN, Argenziano G, Braun RP, Cabo H, Gutman D, Halpern A, Helba B, Hofmann-Wellenhof R, 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.
4. Madan V, Lear JT, Szeimies R-M. Non-melanoma skin cancer. Lancet. 2010;375(9715):673–85.
5. Rubin AI, Chen EH, Ratner D. Basal-Cell Carcinoma. N Engl J Med. 2005;353(21):2262–9.
6. Zhou H, Xie F, Jiang Z, Liu J, Wang S, Zhu C. Multi-classification of skin diseases for dermoscopy images using deep learning 2017. Beijing: IEEE International Conference on Imaging Systems and Techniques (IST); 2017. https://doi.org/10.1109/IST.2017.8261543.
7. Sharma AN, Shwe S, Mesinkovska NA. Current state of machine learning for non-melanoma skin cancer. Arch Dermatol Res. 2022;314:325–7.
8. Argenziano G, Soyer HP, Chimenti S, Talamini R, Corona R, Sera F, Binder M, Cerroni L, De Rosa G, Ferrara G, et al. Dermoscopy of pigmented skin lesions: Results of a consensus meeting via the Internet. J Am Acad Dermatol. 2003;48(5):679–93.
9. Wahba MA, Ashour AS, Guo Y, Napoleon SA, Elnaby MMA. A novel cumulative level difference mean based GLDM and modified ABCD features ranked using eigenvector centrality approach for four skin lesion types classification. Comput Meth Prog Bio. 2018;165:163–74.
10. Marka A, Carter JB, Toto E, Hassanpour S. Automated detection of non-melanoma skin cancer using digital images: a systematic review. BMC Med Imaging. 2019;19(1):21–21.
11. Chollet F. Deep Learning with Python (1st ed.). Manning Publications Co. 2018.
12. Abbas Q. Computer-aided decision support system for classification of pigmented skin lesions. Int J Comput Sci Net Sec. 2016;16(4):9–15.
13. Ballerini L, Fisher RB, Aldridge B, Rees J. Non-melanoma skin lesion classification using colour image data in a hierarchical K-NN classifier. Barcelona: 2012 9th IEEE International Symposium on Biomedical Imaging; 2012. https://doi.org/10.1109/ISBI.2012.6235558.
14. Cheng B, Stanley RJ, Stoecker WV, Hinton K. Automatic telangiectasia analysis in dermoscopy images using adaptive critic design. Skin Res Technol. 2012;18(4):389–96.

15. Chuang S-H, Sun X, Chang W-Y, Chen G-S, Huang A, Li J, McKenzie FD. BCC skin cancer diagnosis based on texture analysis techniques. Lake Buena Vista (Orlando): Medical Imaging 2011: Computer-Aided Diagnosis; 2011. https://doi.org/10.1117/12.878124.
16. Dick V, Sinz C, Mittlböck M, Kittler H, Tschandl P. Accuracy of Computer-Aided Diagnosis of Melanoma: A Meta-analysis. Jama Dermatol. 2019;155(11):1291–9.
17. Ferris LK, Harkes JA, Gilbert B, Winger DG, Golubets K, Akilov O, Satyanarayanan M. Computer-aided classification of melanocytic lesions using dermoscopic images. J Am Acad Dermatol. 2015;73(5):769–76.
18. Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE. Classification of the Clinical Images for Benign and Malignant Cutaneous Tumors Using a Deep Learning Algorithm. J Invest Dermatol. 2018;138(7):1529–38.
19. Kharazmi P, Kalia S, Lui H, Wang ZJ, Lee T. Computer-aided detection of basal cell carcinoma through blood content analysis in dermoscopy images Medical Imaging 2019. San Diego: Computer-Aided Diagnosis; 2018b. https://doi.org/10.1117/12.2293353.
20. Kharazmi P, Kalia S, Lui H, Wang ZJ, Lee TK. A feature fusion system for basal cell carcinoma detection through data-driven feature learning and patient profile. Skin Res Technol. 2018;24(2):256–64.
21. Kharazmi P, Lui H, Wang ZJ, Lee TK. Automatic detection of basal cell carcinoma using vascular-extracted features from dermoscopy images. Vancouver: 2016 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE); 2016. https://doi.org/10.1109/CCECE.2016.7726666.
22. Olsen TG, Jackson BH, Feeser TA, Kent MN, Moad JC, Krishnamurthy S, Lunsford DD, Soans RE. Diagnostic Performance of Deep Learning Algorithms Applied to Three Common Diagnoses in Dermatopathology. J Pathol Inform. 2018;9:32–32.
23. Rajpara SM, Botello AP, Townend J, Ormerod AD. Systematic review of dermoscopy and digital dermoscopy/ artificial intelligence for the diagnosis of melanoma. Br J Dermatol. 2009;161(3):591–604.
24. Shimizu K, Iyatomi H, Celebi ME, Norton K-A, Tanaka M. Four-class classification of skin lesions with task decomposition strategy. IEEE T Bio-Med Eng. 2015;62(1):274–83.
25. Shoieb DA, Youssef SM, Aly WM. Computer-aided model for skin diagnosis using deep learning. J Image Graphics. 2016;4(2):122–9.
26. Spyridonos P, Gaitanis G, Likas A, Bassukas ID. Automatic discrimination of actinic keratoses from clinical photographs. Comput Biol Med. 2017;88:50–9.
27. Sriwong K, Bunrit S, Kerdprasop K, Kerdprasop N. Dermatological Classification Using Deep Learning of Skin Image and Patient Background Knowledge. Int J Mach Learn Comput. 2019;9(6):862–7.
28. Upadhyay PK, Chandra S. Construction of adaptive pulse coupled neural network for abnormality detection in medical images. Appl Artif Intell. 2018;32(5):477–95.
29. Wahba MA, Ashour AS, Napoleon SA, Abd Elnaby MM, Guo Y. Combined empirical mode decomposition and texture features for skin lesion classification using quadratic support vector machine. Health Info Sci Syst. 2017;5(1):10.
30. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MMG, Sterne JAC, Bossuyt PMM. the Q-G: QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. Ann Intern Med. 2011;155(8):529–36.
31. Yap J, Yolland W, Tschandl P. Multimodal skin lesion classification using deep learning. Exp Dermatol. 2018;27(11):1261–7.
32. Zhang X, Wang S, Liu J, Tao C. Towards improving diagnosis of skin diseases by combining deep neural network and human knowledge. BMC Med Inform Decis. 2018;18(Suppl 2):59–59.
33. Nindl I, Gottschling M, Stockfleth E. Human Papillomaviruses and Non-Melanoma Skin Cancer: Basic Virology and Clinical Manifestations. Dis Markers. 2007;23:942650.
34. Srinivasu PN, SivaSai JG, Ijaz MF, Bhoi AK, Kim W, Kang JJ. Classification of Skin Disease Using Deep Learning Neural Networks with MobileNet V2 and LSTM. Sensors. 2021;21(8). https://doi.org/10.3390/s21082852.
35. Brinker TJ, Hekler A, Utikal JS, Grabe N, Schadendorf D, Klode J, Berking C, Steeb T, Enk AH, von Kalle C. Skin Cancer Classification Using Convolutional Neural Networks: Systematic Review. J Med Internet Res. 2018;20(10):e11936.
36. Kassem MA, Hosny KM, Damaševičius R, Eltoukhy MM. Machine Learning and Deep Learning Methods for Skin Lesion Classification and Diagnosis: A Systematic Review. Diagnostics. 2021;11(8):1390.
37. Takiddin A, Schneider J, Yang Y, Abd-Alrazaq A, Househ M. Artificial Intelligence for Skin Cancer Detection: Scoping Review. J Med Internet Res. 2021;23(11):e22934.
38. McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM, and the P-DTAG: Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement. JAMA 2018, 319(4):388–396.
39. Takwoingi Y, Riley RD, Deeks JJ. Meta-analysis of diagnostic accuracy studies in mental health. Evidence Based Mental Health. 2015;18(4):103.
40. Gatsonis C, Paliwal P. Meta-Analysis of Diagnostic and Screening Test Accuracy Evaluations: Methodologic Primer. Am J Roentgenol. 2006;187(2):271–81.
41. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2022. https://www.R-project.org/.
42. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. J Stat Softw. 2015;67(1). https://doi.org/10.18637/jss.v067.i01.
43. Doebler P. mada: Meta-Analysis of Diagnostic Accuracy. 2019. https://CRAN.R-project.org/package=mada.
44. Abhishek K, Kawahara J, Hamarneh G. Predicting the clinical management of skin lesions using deep learning. Sci Rep. 2021;11(1):7769.
45. Chin CK, Mat DAbA, Saleh AY. Hybrid of convolutional neural network algorithm and autoregressive integrated moving average model for skin cancer classification among Malaysian. IAES International J Artificial Intelligence (IJ-AI). 2021;10(3):707–16.
46. Chung HJ, Kim YJ, Song H, Ahn SK, Kim H, Hwang H. Deep Learning-Based Classification of Korean Basal Cell Carcinoma Using Convolutional Neural Network. Journal of Medical Imaging and Health Informatics. 2019;9(1):195–201.
47. Huang K, He X, Jin Z, Wu L, Zhao X, Wu Z, . . . Chen X. Assistant Diagnosis of Basal Cell Carcinoma and Seborrheic Keratosis in Chinese Population Using Convolutional Neural Network. J Healthc Eng. 2020:Article 1713904. https://doi.org/10.1155/2020/1713904.
48. Marvdashti T, Duan L, Aasi SZ, Tang JY, Bowden AKE. Classification of basal cell carcinoma in human skin using machine learning and quantitative features captured by polarization sensitive optical coherence tomography. Biomed Opt Express. 2016;7(9):3721–35.
49. Rojas JAÁ, Calderón Vilca HD, Tumi Figueroa EN, Ramos KJC, Matos Manguinuri SS, Calderón Vilca EF. Hybrid model of convolutional neural network and support vector machine to classify basal cell carcinoma. Computacion y Sistemas. 2021;25(1):83–95.
50. Abunadi I, Senan EM. Deep Learning and Machine Learning Techniques of Diagnosis Dermoscopy Images for Early Detection of Skin Diseases. Electronics. 2021;10(24):3158.
51. Afza F, Sharif M, Khan MA, Tariq U, Yong H-S, Cha J. Multiclass Skin Lesion Classification Using Hybrid Deep Features Selection and Extreme Learning Machine. Sensors. 2022;22(3):799.
52. Ali K, Shaikh ZA, Khan AA, Laghari AA. Multiclass skin cancer classification using EfficientNets – a first step towards preventing skin cancer. Neuroscience Informatics. 2022;2(4):100034.
53. Al-masni MA, Kim D-H, Kim T-S. Multiple skin lesions diagnostics via integrated deep convolutional networks for segmentation and classification. Comput Methods Programs Biomed. 2020;190:105351.
54. Bhardwaj S, Somani A, Gupta K. Detection of Skin Lesion Disease Using Deep Learning Algorithm. Delhi: 3rd International Conference on Artificial Intelligence and Speech Technology; 2022. https://doi.org/10.1007/978-3-030-95711-7_32.
55. Calderón C, Sanchez K, Castillo S, Arguello H. BILSK: A bilinear convolutional neural network approach for skin lesion classification. Comput Methods Programs Biomed Update. 2021;1:100036.
56. Chaahat G. NK, Lehana PK: An Evolutionary Approach for the Enhancement of Dermatological Images and Their Classification Using Deep Learning Models. J Healthc Eng. 2021;2021:8113403.
57. Dorj UO, Lee KK, Choi JY, Lee M. The skin cancer classification using deep convolutional neural network. Multimed Tools Appl. 2018;77(8):9909–24.
58. Jasil SPG, Ulagamuthalvi V. Deep learning architecture using transfer learning for classification of skin lesions. J Ambient Intell Humanized Comput. 2021. https://doi.org/10.1007/s12652-021-03062-7.
59. Jørgensen TM, Tycho A, Mogensen M, Bjerring P, Jemec GBE. Machine-learning classification of non-melanoma skin cancers from image features obtained by optical coherence tomography. Skin Res Technol. 2008;14(3):364–9.

60. Kwiatkowska D, Kluska P, Reich A. Convolutional neural networks for the detection of malignant melanoma in dermoscopy images. Postepy Dermatol Alergol. 2021;38(3):412–20.
61. Liu J, Wang W, Chen J, Sun G, Yang A. Classification and Research of Skin Lesions Based on Machine Learning. Comput Mater Cont. 2020;62(3):1187–200.
62. Mobiny A, Singh A, Van Nguyen H. Risk-Aware Machine Learning Classifier for Skin Lesion Diagnosis. J Clin Med. 2019;8(8):1241.
63. Monika MK, Vignesh NA, Kumari CU. Kumar MNVSS, Lydia EL: Skin cancer detection and classification using machine learning. Mater Today Proc. 2020;33:4266–70.
64. Molina-Molina EO, Solorza-Calderón S, Álvarez-Borrego J. Classification of Dermoscopy Skin Lesion Color-Images Using Fractal-Deep Learning Features. Appl Sci. 2020;10(17):5954.
65. Swetha RN, Shrivastava VK, Parvathi K. Multiclass skin lesion classification using image augmentation technique and transfer learning models. Int J Intell Unmanned Syst. 2021. https://doi.org/10.1108/IJIUS-02-2021-0010. ahead-of-print(ahead-of-print).
66. Pratiwi RA, Nurmaini S, Rini DP, Rachmatullah MN, Darmawahyuni A. Deep ensemble learning for skin lesions classification with convolutional neural network. IAES International J Artificial Intelligence (IJ-AI). 2021;10(3):563–70.
67. Qin Z, Liu Z, Zhu P, Xue Y. A GAN-based image synthesis method for skin lesion classification. Comput Methods Programs Biomed. 2020;195:105568.
68. Rahman Z, Hossain S, Islam R, Hasan M, Hridhee RA. An approach for multiclass skin lesion classification based on ensemble learning. Inform Med Unlocked. 2021;25:100659.
69. Rizwan W, Adnan SM, Ahmed W, Faizi MI. Skin Lesions Detection and Classification Using Deep Learning. International J Advanced Trends Comput Sci Engineering. 2021;10(3):1720–8.
70. Sevli O. A deep convolutional neural network-based pigmented skin lesion classification application and experts evaluation. Neural Comput Appl. 2021;33(18):12039–50.
71. Villa-Pulgarin JP, Ruales-Torres AA, Arias-Garzón D, Bravo-Ortiz MA, Arteaga-Arteaga HB, Mora-Rubio A, Alzate-Grisales JA, Mercado-Ruiz E, Hassaballah M, Orozco-Arias S, et al. Optimized convolutional neural network models for skin lesion classification. Comput Mater Continua. 2022;70(2):2131–48.
72. Yadav U, Kumar A, A T, Mukherjee S. Deep learning in Dermatology for Skin Diseases Detection. International J Recent Technol Eng. 2020;8(6):3929–33.
73. Provost F, Fawcett T. Data Science for Business: What you need to know about data mining and data-analytic thinking. 2nd ed. Sebastopol, CA: O'Reilly Media, Inc.; 2013.
74. Kuhn M, Johnson K. Applied Predictive Modeling, vol. New. York. New York: Springer; 2013.
75. Brownlee J. Ensemble Learning Algorithms with Python. Machine Learning Matery. 2020. https://machinelearningmastery.com/ensemble-learning-algorithms-with-python/.
76. You K, Liu Y, Zhang Z, Wang J, Jordan MI, Long M. Ranking and Tuning Pre-trained Models: A New Paradigm for Exploiting Model Hubs. J Mach Learn Res. 2022;23:1–47.
77. Han X, Zhang Z, Ding N, Gu Y, Liu X, Huo Y, Qiu J, Yao Y, Zhang A, Zhang L, et al. Pre-trained models: Past, present and future. AI Open. 2021;2:225–50.
78. Chlap P, Min H, Vandenberg N, Dowling J, Holloway L, Haworth A. A review of medical image data augmentation techniques for deep learning applications. J Med Imaging Radiat Oncol. 2021;65(5):545–63.
79. Khalifa NE, Loey M, Mirjalili S. A comprehensive survey of recent trends in deep learning for digital images augmentation. Artif Intell Rev. 2022;55(3):2351–77.
80. Zwinderman AH, Bossuyt PM. We should not pool diagnostic likelihood ratios in systematic reviews. Stat Med. 2008;27(5):687–97.
81. Hareer M, Cuijpers P, Furukawa TA, Ebert DD. Doing Meta-Analysis with R: A Hands-On Guide. Boca Raton, FL: CRC Press; 2021.
82. Hu L, Pan X, Tang Z, Luo X. A Fast Fuzzy Clustering Algorithm for Complex Networks via a Generalized Momentum Method. IEEE Trans Fuzzy Syst. 2022;30(9):3473–85.
83. Hu L, Wang X, Huang Y-A, Hu P, You Z-H. A survey on computational models for predicting protein–protein interactions. Brief Bioinform. 2021;22(5):bbab036.
84. Hu L, Zhang J, Pan X, Yan H, You Z-H. HiSCF: leveraging higher-order structures for clustering analysis in biological networks. Bioinformatics. 2021;37(4):542–50.
85. Zhao B-W, Hu L, You Z-H, Wang L, Su X-R. HINGRL: predicting drug–disease associations with graph representation learning on heterogeneous information networks. Brief Bioinform. 2022;23(1):bbab515.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.