scientific reports

Check for updates

OPEN

# Image analysis of cutaneous melanoma histology: a systematic review and meta-analysis

Emily L. Clarke[1,2]✉, Ryckie G. Wade[3], Derek Magee[4], Julia Newton-Bishop[2] & Darren Treanor[1,2,5,6]

**The current subjective histopathological assessment of cutaneous melanoma is challenging. The application of image analysis algorithms to histological images may facilitate improvements in workflow and prognostication. To date, several individual algorithms applied to melanoma histological images have been reported with variations in approach and reported accuracies. Histological digital images can be created using a camera mounted on a light microscope, or through whole slide image (WSI) generation using a whole slide scanner. Before any such tool could be integrated into clinical workflow, the accuracy of the technology should be carefully evaluated and summarised. Therefore, the objective of this review was to evaluate the accuracy of existing image analysis algorithms applied to digital histological images of cutaneous melanoma.Database searching of PubMed and Embase from inception to 11th March 2022 was conducted alongside citation checking and examining reports from organisations. All studies reporting accuracy of any image analysis applied to histological images of cutaneous melanoma, were included. The reference standard was any histological assessment of haematoxylin and eosin-stained slides and/or immunohistochemical staining. Citations were independently deduplicated and screened by two review authors and disagreements were resolved through discussion. The data was extracted concerning study demographics; type of image analysis; type of reference standard; conditions included and test statistics to construct 2 × 2 tables. Data was extracted in accordance with our protocol and the Preferred Reporting Items for Systematic Reviews and Meta-Analyses-Diagnostic Test Accuracy (PRISMA-DTA) Statement. A bivariate random-effects meta-analysis was used to estimate summary sensitivities and specificities with 95% confidence intervals (CI). Assessment of methodological quality was conducted using a tailored version of the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool. The primary outcome was the pooled sensitivity and specificity of image analysis applied to cutaneous melanoma histological images. Sixteen studies were included in the systematic review, representing 4,888 specimens. Six studies were included in the meta-analysis. The mean sensitivity and specificity of automated image analysis algorithms applied to melanoma histological images was 90% (CI 82%, 95%) and 92% (CI 79%, 97%), respectively. Based on limited and heterogeneous data, image analysis appears to offer high accuracy when applied to histological images of cutaneous melanoma. However, given the early exploratory nature of these studies, further development work is necessary to improve their performance.**

Despite advances in therapy, the 5-year survival of patients with metastatic melanoma is still less than 30%[1]. Moreover, the incidence of melanoma is predicted to rise by 7% from 2014 to 2035[2]. Diagnosis of melanoma depends upon a histopathologist's interpretation of the tissue at a cellular level, with subjective thresholds for morphological features. Histopathological interpretation can be challenging, resulting in high levels of interobserver variation[3], which may in part be due to the wide range of histological appearances (see Fig. 1). Consequently,

[1]Department of Histopathology, Leeds Teaching Hospitals NHS Trust, Leeds, UK. [2]Division of Pathology and Data Analytics, Leeds Institute of Cancer and Pathology, University of Leeds, Beckett Street, Leeds LS9 7TF, UK. [3]Leeds Institute for Medical Research, University of Leeds, Leeds, UK. [4]School of Computing, University of Leeds, Leeds, UK. [5]Department of Clinical Pathology, and Department of Clinical and Experimental Medicine, Linköping University, Linköping, Sweden. [6]Center for Medical Image Science and Visualization (CMIV), Linköping University, Linköping, Sweden. ✉email: e.l.clarke@leeds.ac.uk

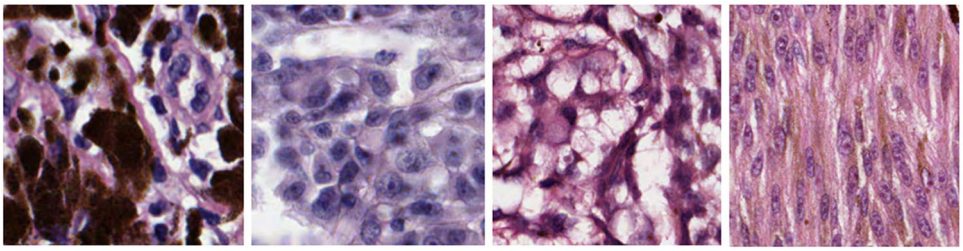

**Figure 1.** An example of the range of histopathological appearances of melanoma. The first image on the far left shows a tumour in which the tumour cells are obscured by large amounts of melanin pigment; the second image from the left shows a more conventional melanoma without pigment; the third image from the left shows a balloon cell variant of melanoma; the image on the right is an example of a spindle cell melanoma. This is an original image created by the authors using Medical Image Manager, HeteroGenius Limited, UK.

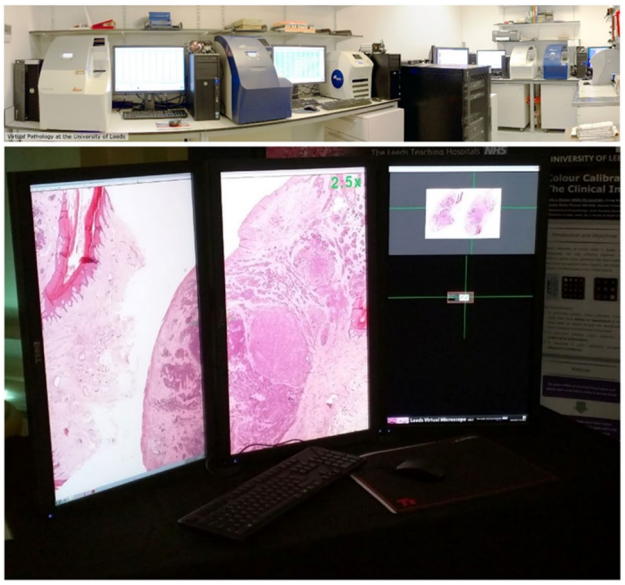


**Figure 2.** Glass slides are scanned using a slide scanner (photo above, credit: Mike Hale, University of Leeds) to create a whole slide image (image below) which can be viewed on a computer display. This digitisation of whole slide images has permitted the application of image analysis algorithms.

up to 17% of diagnoses are reclassified as false positives or false negatives when reviewed by a specialist panel of pathologists[3].

The current prognostic biomarkers based on histological features are contained within the current staging system (American Joint Committee on Cancer, AJCC)[4], with maximum tumour depth (Breslow thickness) remaining the most important predictor of survival for over 50 years[5]. Other prognostic biomarkers contained within the AJCC staging include ulceration, mitoses, lymph node involvement and metastatic disease detected in viscera. However, the staging system explains an insufficient proportion of the variance in survival[6] with some thin tumours unexpectedly causing metastatic disease.

Given that there is currently a staffing shortage in pathology services globally, with only 3% of pathology departments reporting being fully staffed in the United Kingdom (UK)[7], there is a clear need for tools that aid pathologists and improve workflow. It is important that any new prognostic biomarkers not demand additional work of histopathologists.

Digital microscopy has become an essential tool in pathological research over the past few decades. Initially, cameras mounted on microscopes enabled the generation of standard digital images, but the invention of the whole slide scanner over 20 years ago, glass slides can now be scanned to create a whole slide image (WSI) enabling the tissue to be viewed at multiple magnifications (see Fig. 2). The generation of digital images of histological tissue has allowed the creation and application of image analysis (IA) algorithms. More recently still, artificial intelligence (AI) models have been applied to these images and yielded promising results[8–12].

There is a clear need to improve our current subjective histopathological assessment of cutaneous melanoma, which may be achieved by the implementation of image analysis algorithms. There have been several studies of IA applied to melanoma digital slides, all of which have reported variable methodologies and performances.

Prior to any algorithm being adopted into clinical workflows, extensive clinical validation is required, the first step of which would be to provide sufficient evidence to indicate that a model is likely to meet end user requirements. This represents the rationale for this review which summarises the existing evidence and evaluates the performance of these algorithms.

## Materials and methods

This systematic review and meta-analysis was written and performed in accordance with our protocol (PROSPERO ID 336,714) and the Preferred Reporting Items for Systematic Reviews and Meta-Analyses-Diagnostic Test Accuracy (PRISMA-DTA) Statement[13].

**Participants and studies.** We included studies of any design that reported accuracy outcomes of IA applied to histological images of cutaneous melanoma.

**Target condition.** The target condition was cutaneous melanoma.

**Index test.** The index test was any form of automated IA. This includes more conventional IA techniques as well as neural networks. Manual annotation of histological images was not included.

**Reference standard.** The reference standard was any form of histopathological assessment of the haematoxylin and eosin (H&E) histological image and/or immunohistochemical staining.

**Search strategy.** PubMed and Embase were searched from inception to 11th March 2022, restricted to English language (for full search strategy see Appendix 1). Citation checking was also conducted.

**Study selection.** All citations were independently deduplicated and screened by ELC and RGW. Where possible, the full texts of potentially eligible articles were obtained and independently assessed by the same two individuals with disagreements resolved by discussion. We included abstracts as well as full texts.

**Data extraction.** Data were extracted concerning study demographics; type of IA; type of reference standard; conditions included and test statistics to construct $2 \times 2$ tables of the number of true-positives (TP), false-positives (FP), false-negatives (FN) and true-negatives (TN).

**Methodologic quality assessment.** The QUADAS-AI tool was in development at the time of carrying out this work and therefore a tailored version of the Quality Assessment of Diagnostic Accuracy Studies QUADAS-2 was created (per a recent important article in Nature[14]) and used to appraise the risk of bias and applicability of the included studies (Appendix 2).

Assessment of risk of bias for patient selection included whether there was one WSI per patient and if they were contained within one set, since studies that involved multiple WSIs per patient with cases from the same patient spread across the training and test sets result in an overestimation of the model's performance. Risk of bias with regards to the index test included details of the presence of a separate (ideally external) test set and whether all the cases were included in the analysis. Studies that do not use a separate test set are also at risk of overestimating a model's performance. Bias of the reference standard included whether the reference standard results were interpreted without knowledge of the index test, alongside the ability of the reference standard to correctly identify melanoma. If the reference standard is interpreted with knowledge of the index test, then this may bias the reference standard to mirror the index test results, again overestimating performance. Finally, studies including cases with a time interval of more than 10 years between the diagnosis of the reference standard and the digital image creation indicated a high risk of bias for flow and timing, since diagnostic criteria and terminology changes with time and glass slides fade introducing risk that they may not clearly depict the pathology and underestimate the model's performance.

Applicability assessment involved whether the case selection, index test or reference standard matched the review question.

This data was summarised using the Risk-of-bias VISualisation (robvis) tool[15].

**Statistical analysis.** The MetaDTA: Diagnostic Test Accuracy Meta-Analysis v2.01 shinnyapp[16,17] was used to generate summary sensitivities, specificities, forest plots and summary receiver operating characteristic (SROC) plots using a bivariate random-effects model. A sensitivity analysis was performed including only those studies generating IA models concerned with melanoma tumour detection. A flow-diagram was generated using the PRISMA2020 tool[18]. Publication bias was not assessed because the determinants are not well understood for diagnostic accuracy reviews[19] and the Deeks test has low power in the presence of substantial heterogeneity[20]. The significance level was set at 5%. Confidence intervals were generated to the 95% level.

## Results

**Study selection.** Ultimately, sixteen studies were included (Fig. 3).

**study characteristics.** Study characteristics are presented in Supplementary Table 1. Studies originated from the UK[21], Germany[11,22], France[23], Italy[12], Sweden[24], USA[10], Canada[25–28], Japan[29] and China[27,30–33(p202)] and

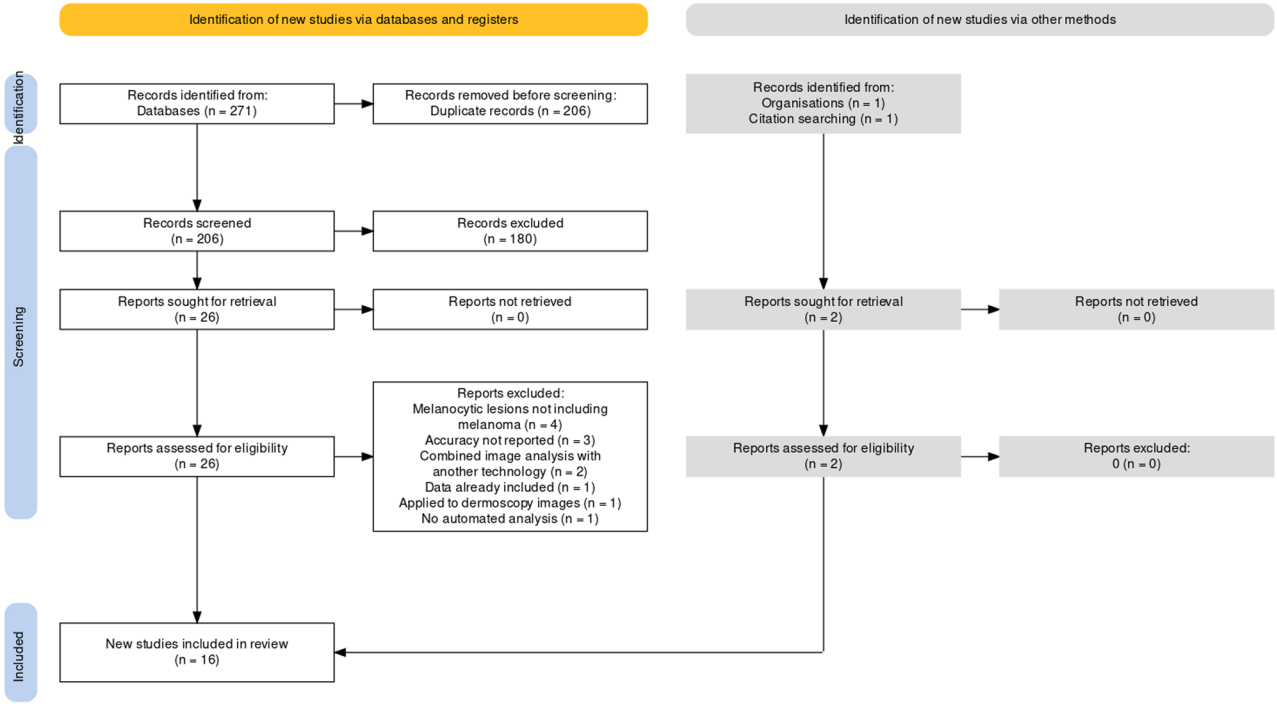


**Figure 3.** Study flow diagram generated using PRISMA2020 available at: https://estech.shinyapps.io/prisma.

were performed between 2012 and 2021. Studies varied in size with a median sample size of 100 specimens or slides (interquartile range [IQR] 66–583.5; range 1–1079). Overall, 4,888 specimens were included, of which at least 2,715 were melanoma specimens. The diagnostic entities within the datasets varied between studies, with some only containing melanoma deposits[12,21,24–26,33] and others containing more than one pathology[10,11,22,27–32].

There was between-study variation in terms of intended use of the IA. Most studies focused on a binary classification task, with some focussing on detection and localisation of melanoma deposits in WSIs containing melanoma (melanoma versus not melanoma)[12,25,26,32] and others performing diagnostic classifications including melanomas versus naevi[11,22,31] and primary melanoma versus metastatic melanoma[33]. Five studies addressed more complex classifications into three or more diagnostics entities[10,23,28–30]. One study[24] did not focus on a classification task, but instead studied automation of the proliferation index in melanoma.

There was some degree of variation in the IT used. Most employed the use of a convolutional neural networks (CNN), with the architecture differing considerably[10–12,22,23,26,29–31,33]. Studies using CNNs were more recently conducted. Two of the earliest studies employed the use of a support vector machine (SVM)[24,27]. Two studies used a combination of a CNN and SVM as their index test[11,32]. A further study used more basic image processing and adaptive thresholding method[29].

There was heterogeneity in the reference standard. In most studies pathologists provided diagnostic labels[10–12,22,31], categorised specimens by histological features[23], carried out manual annotation[21,25,28,30] or interpreted immunohistochemical staining[23,24,26]. Some studies used a combination of these approaches[23,32]. Two studies did not detail the reference standard used[29,33].

There was variation in the reported units for performance analysis. Some studies reported pixel-based outcomes[25,26] or cell-level outcomes[24,27,28], whereas others focussed on patch-level[12,22,31,32] or slide-level classifications[10,11,30,32]. Three studies did not report on their unit for analysis[23,29,33], whilst one appeared to be at the WSI-level[23].

**Risk of bias and applicability concerns.** The risk of bias and applicability assessment are summarised in Fig. 4. Twelve studies were at risk of selection bias[10,11,22–24,26–31,33], of which, four studies were at high risk since more than one histological image was included per patient and were spread across the training/ test sets[10,26,27,33]. The remaining studies were at unclear risk of patient selection bias[12,21,23,27–29,31,33]. Thirteen studies were at risk of bias from the index test[11,12,21–23,25–31,33]; seven studies were at high risk either due to the index test not being tested on an external test set (i.e., a source separate to those used for training/ validation)[21,22,25,27,30], or not reporting results from a separate test set[28], or the test set being derived from the same histological slide as the training and validation sets[26]. Six studies were at unclear risk of bias from the index test[11,12,23,29,31,33]. Eleven studies were at unclear risk of bias from the reference standard[11,23–29,31,31–33] because it was not clear if the reference standard results were interpreted without knowledge of the IA or if the reference standard was likely to correctly classify the target condition. No studies were at high risk. Fifteen studies were at risk of bias due to the flow and timing[10–12,21,23,25–33]; one study was at high risk of bias since the reference standard was determined over 10 years prior to the index test being conducted[24]. The remaining fourteen studies were at unclear risk of bias since the timings for the determination of the reference standard and index test were not reported[10–12,21,23,25–33].


**Figure 4.** QUADAS-2 summary diagram assessing risk of bias and applicability in the included studies. For more information on how the judgements were made, see Appendix S2.

Twelve studies were of unclear concern regarding the applicability of patient selection[12,21,23–29,31,33], due to it not being stated if the cases were purposively selected. There were applicability concerns for seven studies regarding the index test; three studies were of high concern[22,26,28] and four studies[23,29,31,33] were of unclear concern that the index test[22,23,26,28,29,31,33] or its conduct or interpretation differed from the review question. Eight studies[23–29,33] had unclear concerns for applicability of the reference standard because it wasn't clear if the individual determining the reference standard was suitably qualified or there were unclear criteria for diagnosis.

**Synthesis of results.** Of the sixteen studies included in this systematic review[10–12,21–33], six studies had data that could be meta-analysed[11,12,21,22,32,33]. The extracted data from five of these studies were from published work[12,21,22,32,33] and additional data from one study was provided by the authors[11]. Over these 5 studies, 1,935 specimens were included, of which at least 1,088 were melanoma specimens. The true-positive, false-positive, false-negative and true-negative rates can be seen in Supplementary Table 2.

Figure 5 shows forest plots of the sensitivity and specificity of any form of IA applied to cutaneous melanoma histological images. The mean sensitivity was 90% (CI 82%, 95%) and mean specificity was 92% (CI 79%, 97%), as shown in Fig. 6. For the studies which could not be included in the meta-analysis due to deficiencies in reporting, the performance metrics are summarised in Supplementary Table 3.

*Sensitivity analysis.* A sensitivity analysis was performed using the 5 studies concerned with tumour detection[11,12,22,32,33]. In total there were 1,853 specimens, of which at least 1,088 were melanoma specimens. The mean sensitivity of IA for cutaneous melanoma tumour detection was 88% (CI 79%, 93%) and a mean specificity of 90% (CI 71%, 97%).

## Discussion

For all tasks, IA applied to cutaneous melanoma histological images has a high sensitivity and specificity (Fig. 6). When including only those studies concerned with tumour detection, the results were similar. The performance of the models not included in the meta-analysis were also favourable (Supplementary Table 3).

As shown in Fig. 5, three[12,21,33] of the six meta-analysed studies reported very high sensitivities and specificities, whereas the other three[11,22,32] were more modest. These three studies[11,22,32] applied the IA to a reasonably sized separate test dataset containing more than one diagnostic entity. By contrast, the other three studies[12,21,33]

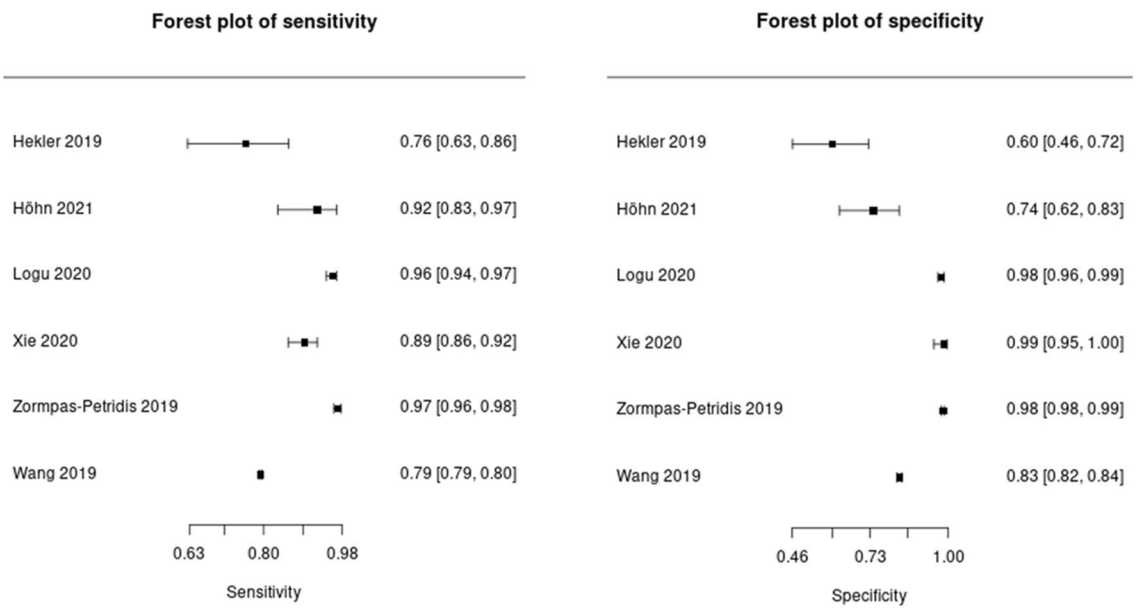


**Figure 5.** Forest plots of the sensitivity and specificity of image analysis applied to melanoma whole slide images.

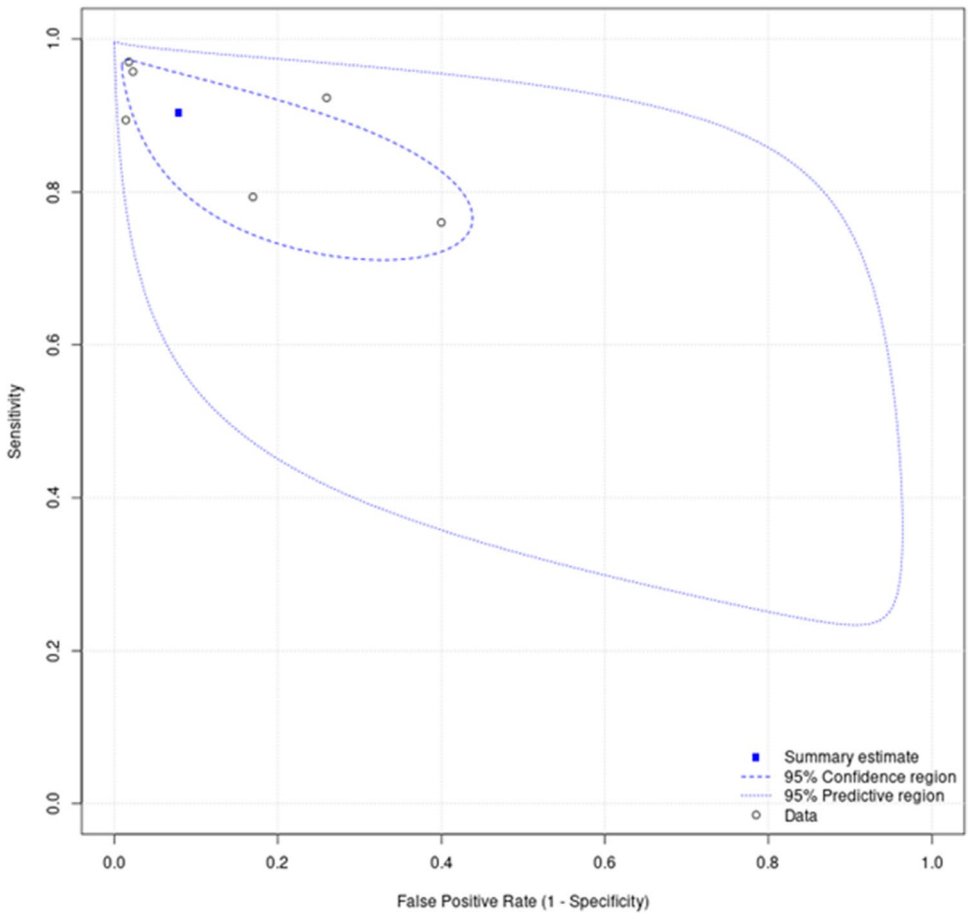


**Figure 6.** Summary receiver operating characteristic plot of image analysis applied to melanoma whole slide images. The confidence region are the 95% confidence intervals around the summary estimate. The predictive region also captures between study statistical heterogeneity, so depicts the region in which we have 95% confidence that the true sensitivity and specificity of a future study should lie. The predictive region encompasses the possibility that the index test may be worse than chance.

contained less varied datasets containing only melanoma specimens, which may explain the more accurate results.

Across the 16 studies included in this review[10–12,21–33], there was no clear association between the type of index test or reference standard and the reported performance. Surprisingly, an increase in the number of data sources did not appear to temper performance; two studies[10,12] contained data from three sources but reported highly accurate results.

Given the exploratory nature of this work, there is currently no consensus regarding whether a more sensitive or specific test is preferable. A more sensitive test would result in fewer false negatives, which in the context of melanoma detection is likely to be of greater utility as missing a melanoma would not be acceptable, particularly as these tests are likely to be used as a screening or triage tool prior to pathologist assessment.

The studies that could not be included in the meta-analysis due to the lack of raw data or data appropriate for back calculation, reported alternative performance metrics including accuracy[25,26,30,31], dice co-efficient[26], area under the curve (AUC)[10,24,30], F-score[24,29,31], precision[24], recall[24], percentage correctly classified[23], positive prediction rate[28], under-segmentation rate[28]. The unit of analysis was also variable for the same reasons and included classifications at a pixel-level[25,26], cell-level[24,27,28], patch-level[12,22,31,32] and slide-level[10,11,30,32]. This variety in reported unit of analysis and performance metrics presents challenges for interpretation and data amalgamation, but it is expected given the wide range of model tasks included in this review. It is essential that the unit of analysis is appropriate for the task to prevent inaccurate performance estimation, as detailed in a seminal paper on the subject[34]. However, regardless of the unit of analysis or performance metrics presented, we urge authors to report their raw data in a confusion matrix (containing the TP, FP, TN, FN counts) for classification-based tasks as per existing guidelines[35].

The clinical utility of the studies presenting results at the slide-level was clear; to assist with specimen triage[10,11,22,23,30,31,33]. However, many studies which detected melanoma at a cell, pixel or patch-level did not address the clinical utility of their models[21,25–29], when these models are suited to prognostic biomarker generation. This may be due to difficulties acquiring the necessary data, but we would recommend that future studies detecting melanoma at a cell, pixel or patch-level, focus on how these models could be applied to predict patient outcome.

Our review had limitations. While 16 studies were included in the review, data extraction was only possible for six of the studies owing to deficient reporting[11,12,21,22,32,33]. There was concern for risk of bias and applicability in all included studies, although reporting standards and methodological rigor did appear to improve with time. This variation in methodological rigor and reporting standards is likely due to a lack of reporting guidelines, although these are currently under development[14]. Additionally, our risk of bias and applicability assessments may be suboptimal since the QUADAS-AI tool was still in development at the time of completion of this work. Future reviewers should deploy this AI-specific tool.

## Conclusion

Based on limited and heterogenous data, IA offers high accuracy when applied to melanoma histological images. The focus of work to date has been on developing the technology in this field, which has accelerated over the past decade. Going forwards, future work should address the clinical application of such models and evaluate their use as a screening/ triage tool or for prognostic/ predictive biomarker generation. The quality of existing studies is variable but is improving with time—it is important that authors report their data according to AI-specific guidelines[14] once they are published.

## Data availability

Data used to derive the results presented in this paper are available in the supplementary material.

Received: 19 October 2022; Accepted: 14 March 2023
Published online: 23 March 2023

## References

1. Cancer Research UK. Melanoma Skin Cancer Survival. https://www.cancerresearchuk.org/about-cancer/melanoma/survival
2. Cancer Research UK. Melanoma Skin Cancer Statistics. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/melanoma-skin-cancer#heading-Zero
3. Elmore, J. G. *et al.* Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: Observer accuracy and reproducibility study. *BMJ Online* https://doi.org/10.1136/bmj.j2813 (2017).
4. Keung EZ, Gershenwald JE. (2018) The eighth edition American joint committee on cnacer (AJCC) melanoma staging system: Implications for melanoma treatment and care.*Exp. Rev. Anticancer Ther*. 18(8):775–784. https://doi.org/10.4049/jimmunol.180147 3.
5. Breslow, A. Thickness, cross-sectional areas and depth of invasion in the prognosis of cutaneous melanoma. *Ann. Surg.* **172**(5), 902–908. https://doi.org/10.1097/00000658-197011000-00017 (1970).
6. Nsengimana, J. *et al.* Independent replication of a melanoma subtype gene signature and evaluation of its prognostic value and biological correlates in a population cohort. *Oncotarget* **6**(13), 11683 (2015).
7. The Royal College of Pathologists. Meeting pathology demand Histopathology workforce census. Published online 2018. https://www.rcpath.org/uploads/assets/952a934d-2ec3-48c9-a8e6e00fcdca700f/meeting-pathology-demand-histopathology-workforce-census-2018.pdf
8. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**(7639), 115–118. https://doi.org/10.1038/nature21056 (2017).
9. Steiner, D. F. *et al.* Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *Am. J. Surg. Pathol.* **42**(12), 1636–1646. https://doi.org/10.1097/PAS.0000000000001151 (2018).
10. Sankarapandian, S. *et al.* A pathology deep learning system capable of triage of melanoma specimens utilizing dermatopathologist consensus as ground truth*. *Proc. IEEE Int. Conf. Comput. Vis.* https://doi.org/10.1109/ICCVW54120.2021.00076 (2021).

11. Hohn, J. *et al.* Combining CNN-based histologic whole slide image analysis and patient data to improve skin cancer classification. *Eur. J. Cancer Oxf. Engl.* **2021**(149), 94–101. https://doi.org/10.1016/j.ejca.2021.02.032 (1990).
12. De Logu, F. *et al.* Recognition of cutaneous melanoma on digitized histopathological slides via artificial intelligence algorithm. *Front. Oncol.* **10**, 1–8. https://doi.org/10.3389/fonc.2020.01559 (2020).
13. McInnes, M. D. F. *et al.* Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies the PRISMA-DTA statement. *JAMA J. Am. Med. Assoc.* **319**(4), 388–396. https://doi.org/10.1001/jama.2017.19163 (2018).
14. Sounderajah, V. *et al.* Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: The STARD-AI protocol. *BMJ Open* **11**(6), 1–7. https://doi.org/10.1136/bmjopen-2020-047709 (2021).
15. McGuinness, L. A. & Higgins, J. P. T. Risk-of-bias visualization (robvis): An R package and Shiny web app for visualizing risk-of-bias assessments. *Res. Synth. Methods* https://doi.org/10.1002/jrsm.1411 (2020).
16. Patel, A., Cooper, N., Freeman, S. & Sutton, A. Graphical enhancements to summary receiver operating characteristic plots to facilitate the analysis and reporting of meta-analysis of diagnostic test accuracy data. *Res. Synth. Methods* **12**(1), 34–44. https://doi.org/10.1002/jrsm.1439 (2021).
17. Freeman, S. C. *et al.* Development of an interactive web-based tool to conduct and interrogate meta-analysis of diagnostic test accuracy studies: MetaDTA. *BMC Med. Res. Methodol* **19**(1), 1–11. https://doi.org/10.1186/s12874-019-0724-x (2019).
18. Haddaway, N. R., Page, M. J., Pritchard, C. C. & McGuinness, L. A. PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. *Campbell Syst. Rev.* **18**(2), 1–12. https://doi.org/10.1002/cl2.1230 (2022).
19. Macaskill P, Gatsonis C, Deeks J, Harbord R, Takwoingi Y. *Chapter 10 Analysing and Presenting Results*. Version 1. The Cochrane Collaboration; 2010:79. http://srdta.cochrane.org/.
20. Deeks, J., Macaskill, P. & Irwig, L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J. Clin. Epidemiol.* **58**(9), 882–893 (2005).
21. Zormpas-Petridis, K., Noguera, R. D. K. I., Roxanis, I., Jamin, Y. & Yuan, Y. SuperHistopath: A deep learning pipeline for mapping tumor heterogeneity on low-resolution whole-slide digital histopathology images. *Front. Oncol.* **10**, 586292–586292. https://doi.org/10.3389/fonc.2020.586292 (2020).
22. Hekler, A. *et al.* Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. *Eur. J. Cancer* **118**, 91–96. https://doi.org/10.1016/j.ejca.2019.06.012 (2019).
23. Henriet, J., Monnin, C., Clerc, J., Morello, B. & Zehrouni, N. 508 diagnosis of spitz tumor using artificial neural networks. *Lab. Invest.* **97**, 130A-130A. https://doi.org/10.1038/labinvest.2016.165 (2017).
24. Rexhepaj, E. *et al.* A texture based pattern recognition approach to distinguish melanoma from non-melanoma cells in histopathological tissue microarray sections. *PLoS ONE* https://doi.org/10.1371/journal.pone.0062070 (2013).
25. Alheejaw, S., Xu, H., Berendt, R., Jha, N. & Mandal, M. Novel lymph node segmentation and proliferation index measurement for skin melanoma biopsy images. *Comput. Med. Imaging Graph Off. J. Comput. Med. Imaging Soc.* **73**, 19–29. https://doi.org/10.1016/j.compmedimag.2019.01.006 (2019).
26. Alheejawi, S., Berendt, R., Jha, N., SP, M. & Mandal, M. Detection of malignant melanoma in H&E-stained images using deep learning techniques. *Tissue Cell.* **73**, 101659. https://doi.org/10.1016/j.tice.2021.101659 (2021).
27. Lu, C. & Mandal, M. Automated analysis and diagnosis of skin melanoma on whole slide histopathological images. *Pattern Recognit.* **48**(8), 2738–2750. https://doi.org/10.1016/j.patcog.2015.02.023 (2015).
28. Lu, C., Mahmood, M., Jha, N. & Mandal, M. A robust automatic nuclei segmentation technique for quantitative histopathological image analysis. *Anal. Quant. Cytopathol. Histopathol.* **34**, 1–13 (2012).
29. Ota, Y. *et al.* Deep ackerman a novel deep learning method to develop dermatopathology diagnosis by artificial intelligence. *J. Invest. Dermatol.* **138**, 5 (2018).
30. Li, T. *et al.* Automated diagnosis and localization of melanoma from skin histopathology slides using deep learning: A multicenter study. *J. Healthc. Eng.* https://doi.org/10.1155/2021/5972962 (2021).
31. Li, F. *et al.* 828 Dermatopathologist-level classification of skin cancer with deep neural networks at multi-magnification. *J. Invest. Dermatol.* https://doi.org/10.1016/j.jid.2019.03.904 (2019).
32. Wang, L. *et al.* Automated identification of malignancy in whole-slide pathological images: Identification of eyelid malignant melanoma in gigapixel pathological slides using deep learning. *Br J. Ophthalmol.* https://doi.org/10.1136/bjophthalmol-2018-313706 (2019).
33. Xie, P. *et al.* Predicting metaststic melanoma from melanoma pathologica images using a convolutional neural network: A multicentre study. *J. Invest. Dermatol.* https://doi.org/10.1016/j.jid.2020.05.047 (2020).
34. Reinke A, Eisenmann M, Tizabi MD, et al. (2021)Common limitations of image processing metrics: A picture story. Pub. Online 1–11.
35. The Cochrane Collaboration. *The Cochrane Handbook for DTA Reviews*.; (2016).


## Author contributions

E.L.C., D.T., J.N.B. and D.M. planned the study and conducted the searches. Publications were screened and data extracted by E.L.C. and R.G.W. E.L.C. analysed the data and wrote the manuscript, which was revised by R.G.W., D.M., J.N.B. and D.T.

## Funding

This work was supported by the Medical Research Council (MR/S001530/1), National Institute for Health Research (NIHR, DRF-2018–11-ST2-028) and NPIC (Project no. 104687) which is supported by a £50 m investment from the Data to Early Diagnosis and Precision Medicine strand of the government's Industrial Strategy Challenge Fund, managed and delivered by UK Research and Innovation (UKRI). The views expressed are those of the author(s) and not necessarily those of the United Kingdom's National Health Service, NIHR or Department of Health.

## Competing interests

We declare the following interests: Author DT is Director of National Pathology Imaging Co-operative (NPIC). Author DM is a director of HeteroGenius Limited. Author RW is a Doctoral Research Fellow funded by the National Institute for Health Research. Author EC is a Clinical Research Training Fellow funded by the Medical Research Council. Author JNB has no conflicts of interest to declare.

## Additional information

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-31526-7.

**Correspondence** and requests for materials should be addressed to E.L.C.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.