# Artificial intelligence for the classification of pigmented skin lesions in populations with skin of colour: A systematic review

*Yuyang Liu, Clare A Primiero, Vishnu Kulkarni, Peter Soyer, Brigid Betz-Stablein[1]*

*The University of Queensland Diamantina Institute, Brisbane, Australia*

## Introduction

Skin cancer is the most common malignancy worldwide, with melanoma representing the deadliest form. While skin cancers are less prevalent in people with skin of colour, they are more often diagnosed at a later stage and have a poorer prognosis when compared to Caucasian populations. The use of artificial intelligence (AI) models can potentially improve early detection of skin cancers, however, the lack of skin colour diversity in training datasets may widen pre-existing racial discrepancies in dermatology. While multiple reviews have compared AI-based model performances for skin cancer detection, the use of AI in populations with skin of colour has not been evaluated.

## Objective

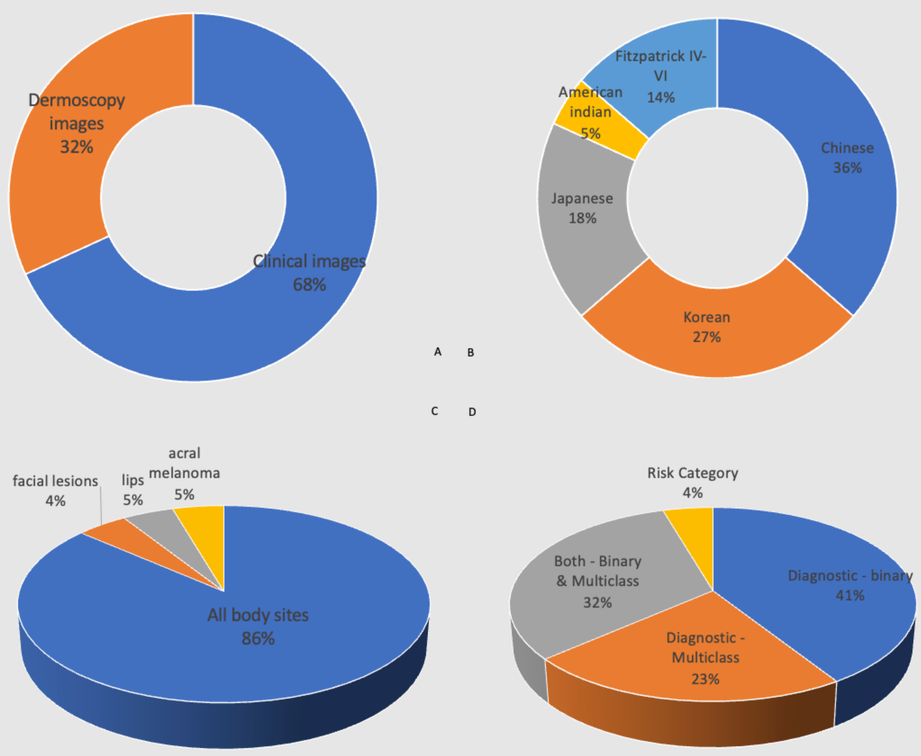
To review the technique, quality, accuracy, and implications of studies using AI models trained in populations with skin of colour, for classification of pigmented skin lesions.

## Methods

Two independent reviewers screened PubMed via key term search to identify AI studies on classification of pigmented skin lesions with at least 10% of images from skin of colour populations.

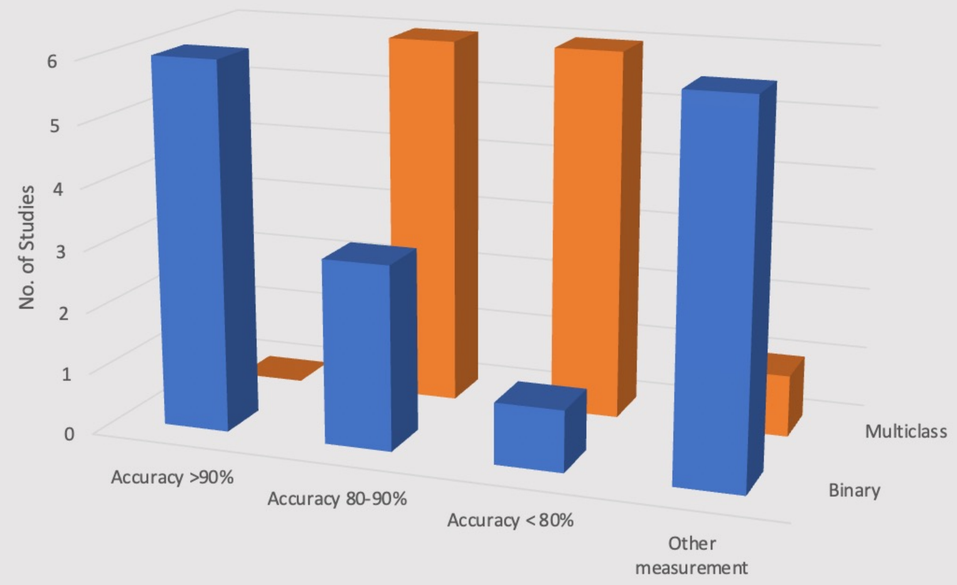
## Results

### Characteristic and metadata




A. **Imaging modality** B. **Ethnicity/Ancestry/Race/Location** C. **Lesion location** D. **Classification system.** Two algorithm classification systems were used, diagnostic and risk-categorical. The diagnostic model included binary classification of the lesion into either benign or malignant and multiclass classification of the lesion into a specific lesion diagnosis. Risk categorical model classed lesions into low, medium, or high risk.

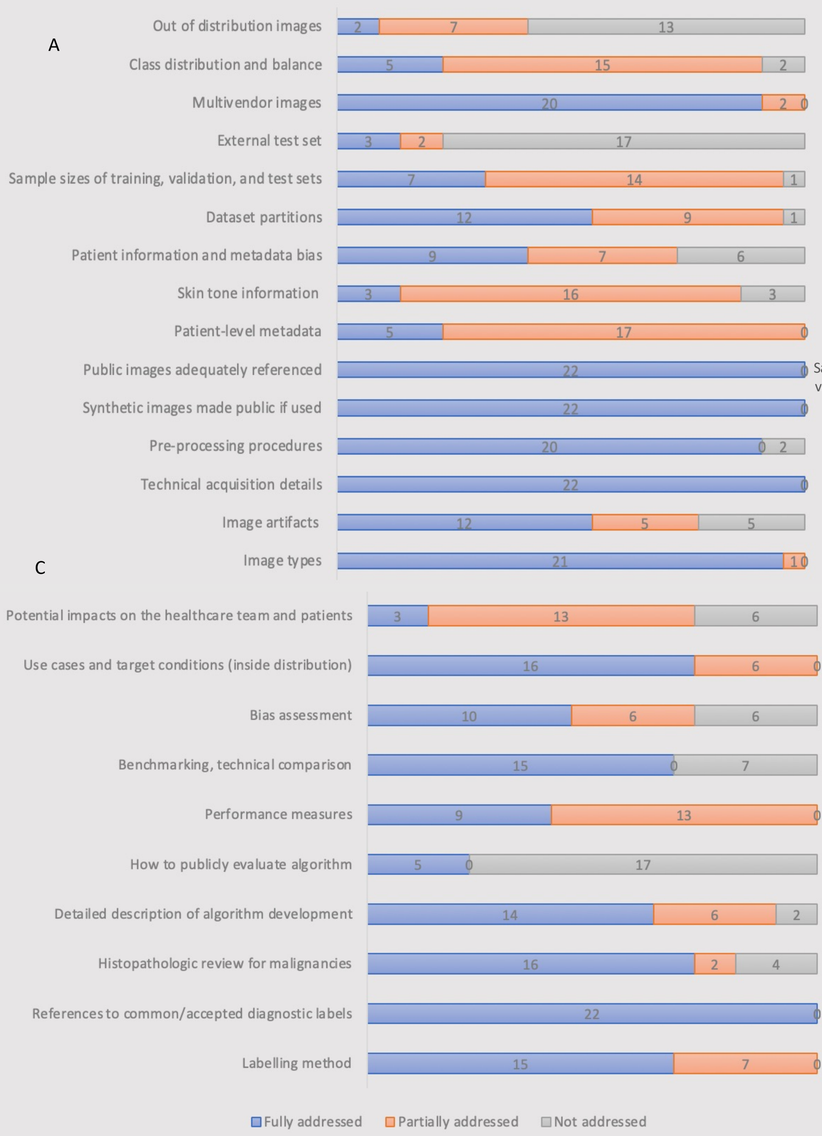
### Outcome of AI models




| | Accuracy >90% | Accuracy 80-90% | Accuracy < 80% | Other measurement |
|---|---|---|---|---|
| Binary | 6 | 3 | 1 | 6 |
| Multiclass | 0 | 6 | 6 | 1 |

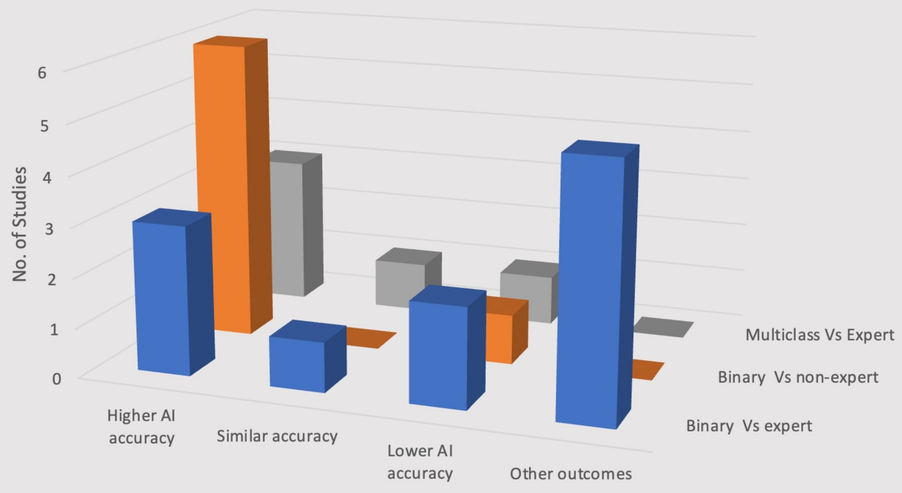
**Study outcomes.** 16 studies used binary classification, out of these, 10 reported an accuracy ranging from 70% to 99.7%, 6 used specificity and sensitivity as a measure of performance. 13 studies used multiclass classification (i.e. specific diagnosis), out of these, 12 reported accuracy ranging from 43% to 93%, 1 used specificity and sensitivity as a measure of performance

### Quality assessment




**Quality assessment.** Studies included in the review were evaluated against the 25-point CLEAR Derm Consensus Guidelines, covering four domains (data, technique, technical assessment, and application). For each checklist item, each study was assessed whether it fully, partially or did not address the criteria and scored either 1, 0.5 or 0, respectively, using a scoring rubric. A&C Number of studies in each assessment category. B&D Average score of included studies for each checklist item. The most poorly addressed criteria were patient-level metadata (e.g., sex, gender, ethnicity), skin colour information, using an external test dataset, class distribution and balance of the images, out-of-distribution images, public evaluation of the model, and discussion around potential impacts on healthcare teams.

### Comparison of AI models to clinician classification




| | Higher AI accuracy | Similar accuracy | Lower AI accuracy | Other outcomes |
|---|---|---|---|---|
| Binary Vs expert | 3 | 1 | 2 | 5 |
| Binary Vs non-expert | 6 | 0 | 1 | 0 |
| Multiclass Vs Expert | 3 | 1 | 1 | 0 |

**Reader study outcomes.** The performance of AI models and clinician classification is compared. Six studies compared AI outcomes to classification by experts, e.g., dermatologists. Eight studies compared outcomes to both experts and non-experts, e.g., dermatology residents and general practitioners

## Conclusion

1. We present, to our knowledge, the first systematic review summarizing existing AI image-based algorithms for the classification of skin lesions in people with skin of colour.
2. Within the identified AI studies involving skin of colour populations, there is further under-representation of darker skin types.
3. AI models in this review showed reasonably high-performance classification using both dermoscopic and clinical images in populations with skin of colour.
4. However, several critical items from the CLEAR Checklist were insufficiently addressed by the majority of AI imaging studies in populations with skin of colour.
5. Active inclusion of skin-of-colour populations and standardized, high-quality reporting in AI model development is essential.