Early detection of sexually transmitted infections from skin lesions with deep learning: a systematic review and meta-analysis

Ming Liu, Xin-Yao Yi, Yun-Zhe Chen, Mei-Nuo Li, Yuan-Yuan Zhang, Casper J P Zhang, Jian Huang, Wai-Kit Ming

Summary

Background Sexually transmitted infections (STIs) are a substantial public health concern. We aimed to evaluate the accuracy and applicability of deep learning algorithms in the early detection of STIs from skin lesions.

Methods In this systematic review and meta-analysis, we searched PubMed, Institute of Electrical and Electronics Engineers Xplore, Web of Science, Scopus for studies employing deep learning for classifying clinical skin lesion images of STIs published between Jan 1, 2010, and Dec 31, 2023. Studies that did not include clinical images were excluded. The primary outcome was diagnostic performance, assessed by pooled sensitivity and specificity. We conducted a meta-analysis of the studies providing contingency tables using a unified hierarchical model. We additionally assessed the quality of the studies using modified QUADAS-2 and CheckList for Evaluation of image-based AI Reports in Dermatology (CLEAR Derm) criteria. This study was registered with PROSPERO, CRD42024496966.

Findings Among the 1946 studies identified, we included 101 in our review. The majority of the included studies focused on mpox (91 [88%] of 101 studies), followed by scabies (eight [8%] studies), herpes (four [4%] studies), syphilis (one [1%] study), and molluscum (one [1%] study). A meta-analysis of 55 studies showed that deep learning algorithms had a pooled sensitivity of 0·97 (95% CI 0·95–0·98) and a specificity of 0·99 (0·98–0·99) for mpox, and a sensitivity of 0·95 (0·90–0·98) and specificity of 0·97 (0·86–0·99) for scabies. The majority of studies (86 [85%] of 101 studies) utilised public datasets; traditional convolutional neural networks with backbone architectures such as ResNet and VGGNet were used in all studies. However, notable quality issues related to the data, technical descriptions of labelling methods and diagnostic label references, technical assessment for public evaluation of algorithms, benchmarking and bias assessments, application descriptions of use cases, and target conditions and potential impacts were identified in CLEAR Derm. Potential biases in performance evaluation metrics and applicability concerns in the data, deep learning algorithms, and performance evaluation metrics might impede the generalisability of these models to real-world clinical practice and STI screening across diverse populations.

Interpretation Although deep learning shows potential for early detection of STIs, there are challenges to ensuring the generalisability of such algorithms due to limited heterogeneous data. Standardised, diverse skin lesion image datasets are crucial to ensure fair comparisons and reliable performance.

Funding City University of Hong Kong.

Copyright © 2025 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Lancet Digit Health 2025; 7: 100894

Published Online August 6, 2025
https://doi.org/10.1016/j.landig.2025.100894

Department of Infectious Diseases and Public Health, Jockey Club College of Veterinary Medicine and Life Sciences (M Liu MSc, X-Y Yi MSc, Prof W-K Ming MD) **and Institute of Global Governance and Innovation for a Shared Future** (Prof W-K Ming), **City University of Hong Kong, Hong Kong, Special Administrative Region, China; School of Public Health, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong Special Administrative Region, China** (Y-Z Chen MPH, M-N Li MPH, Y-Y Zhang MPH, C J P Zhang PhD)**; Institute for Human Development and Potential, Agency for Science, Technology and Research (A*STAR), Singapore** (J Huang PhD)**; Bioinformatics Institute, Agency for Science, Technology and Research (A*STAR), Singapore** (J Huang)**; Human Potential Translational Research Programme, Yong Loo Lin School of Medicine, National University of Singapore, Singapore** (J Huang)

Correspondence to:
Prof Wai-kit Ming, Department of Infectious Diseases and Public Health, Jockey Club College of Veterinary Medicine and Life Sciences, City University of Hong Kong, Hong Kong 999077, Special Administrative Region, China
wkming2@cityu.edu.hk

Introduction

Sexually transmitted infections (STIs) pose a substantial public health concern, with an estimated 770 million incident cases and 1·3 million disability-adjusted life-years attributed to STIs in 2019. Incidence has more than doubled from 1990 to 2019, particularly in regions with low social development indices.[1] The early detection of STIs facilitates expeditious treatment, curtails further dissemination, and diminishes both individual protracted health hazards and strains on health-care infrastructure.[2] Here, we define early detection as recognising an STI at the stage when characteristic skin lesions first manifest, before confirmatory laboratory testing or formal clinical diagnosis.

The majority of STIs are asymptomatic, such as chlamydia, gonorrhea, trichomonas, and genital herpes; however, certain infections such as mpox, scabies, syphilis, and some human papillomavirus subtypes frequently manifest with cutaneous presentations that can serve as pivotal diagnostic clues.[3–5] Diagnosis of STIs via laboratory testing is often constrained by the availability of affordable diagnostic tests, particularly in low-income and middle-income countries (LMICs). Consequently, many countries have adopted syndromic management approaches to manage people with STIs.[6] Moreover, the complexity of the clinical diagnosis of STIs is compounded by the presence of a wide range of skin conditions, overlapping symptoms, frequent co-infection, and the stigma that discourages

Research in context

Evidence before this study

Deep learning is a common method used in machine learning for image classification tasks across a range of fields, including dermatology. In 2022, an outbreak of mpox, a sexually transmitted infection (STI), sparked a growing interest in the application of deep learning for the early detection of STIs. We did a thorough review of several online databases, including PubMed, Institute of Electrical and Electronics Engineers Xplore, Web of Science, Scopus, and Google Scholar, for studies published between Jan 1, 2010, and Dec 31, 2023, which used deep learning to classify clinical skin lesion images related to STIs; we did not use language restrictions. We used search terms related to deep learning, skin lesions, and STIs. The timeframe that we selected aimed to capture the significant advancements in deep learning techniques following the ImageNet Large Scale Visual Recognition Challenge 2010. We identified 1946 studies, of which 101 studies were included in this review. Although numerous primary studies have reported promising diagnostic capabilities in STI classification using deep learning methods, those findings often relied on small datasets.

Added value of this study

This systematic review and meta-analysis presents a comprehensive assessment of deep learning algorithms related to the early detection of STIs. Most included studies included clinical images to classify STI-related skin lesions; some studies integrated additional tabular data such as demographic information and patient medical history. Deep learning algorithms were generally effective at classifying STIs, especially mpox and scabies; traditional convolutional neural network architectures were most commonly used. Our findings emphasise the significance of explainable artificial intelligence techniques, and we present a roadmap for developing new deep learning models for detecting STIs and enhancing the design, evaluation, and implementation of such models. However, this review also identified substantial quality issues pertaining to multiple research facets (eg, data quality, technical methodology, and performance evaluation metrics). These issues constrain the applicability of the models to practical settings.

Implications of all the available evidence

This research emphasises the importance of standardised datasets that encompass a diverse range of skin tones, body locations, and geographical regions, to facilitate equitable comparisons and enhance the reliability of deep learning models in the detection of STIs. Furthermore, the establishment of medical guidelines for artificial intelligence-driven STI screening, which outline specific use cases and target conditions, is essential for successful implementation of these models. To integrate deep learning models into mobile applications, it is imperative to prioritise the optimisation of model size and parameters reduction to facilitate rapid predictions. Additionally, the local implementation of algorithms and privacy preserving techniques are crucial factors to consider for future applications, particularly in scenarios where skin lesions might present in sensitive body locations.

many people, especially sexual and gender minorities, from seeking testing and treatment.[7–11]

Deep learning technologies have shown promise in classifying skin lesions, including skin cancer detection[12] and screening for STIs such as mpox,[13] syphilis,[14] and scabies.[15] The 2022 global mpox outbreak, together with the release of several open mpox skin lesion image datasets, the distinctive skin lesions, and complexities around laboratory diagnosis, has accelerated interest in this approach.[4] The field of computer vision has boosted interest in deep learning following the ImageNet Large Scale Visual Recognition Challenge 2010,[16] leading to the development of convolutional neural networks such as AlexNet,[17] VGGNet,[18] and ResNet,[19] followed more recently by attention-based architectures such as Vision Transformer.[20,21] Transfer learning is commonly used in medical image analysis, where pre-trained models are fine-tuned for specific tasks.[22] Mobile applications utilising deep learning models present a promising avenue for direct-to-consumer STI testing, addressing barriers such as stigma and inconvenience.[23] These models can also support health-care providers, especially in LMICs where access to adequately trained general practitioners is scarce. However, the so-called black box nature of deep learning raises concerns about model transparency, prompting the use of eXplainable Artificial Intelligence (XAI) techniques such as Class Activation Maps (CAM), saliency maps, Local Interpretable Model-agnostic Explanations (LIME), and SHapley Additive exPlanations (SHAP).[12]

In this systematic review, we aimed to evaluate the performance and applicability of deep learning models for early detection of STIs, focusing on their capability to classify skin lesions associated with STIs. We also did a meta-analysis to assess the diagnostic accuracy of deep learning algorithms and explored concerns regarding their generalisability due to data variety and representation.

For the **study protocol** see https://www.crd.york.ac.uk/PROSPERO/view/CRD42024496966

See Online for appendix

Methods

Search strategy and selection criteria

This systematic review and meta-analysis was done in accordance with PRISMA guidelines,[24] and the protocol was registered with PROSPERO.[25] Two authors (ML and X-YY) independently searched PubMed, Institute of Electrical and Electronics Engineers (IEEE) Xplore, Web of Science, and Scopus for primary research articles published in peer-reviewed journals or conference proceedings between Jan 1, 2010, and Dec 31, 2023, using keywords related to deep learning, skin lesions, and STIs (appendix pp 3–6), without language restrictions. ML and X-YY also searched Google Scholar for the same keywords.

We included studies utilising deep learning algorithms for the classification of skin lesions associated with STIs, where datasets used in the studies involved clinical images as the input for the algorithm. We excluded non-primary studies, such as review articles, book chapters, or conference abstracts; studies that did not utilise deep learning algorithms; studies employing deep learning algorithms not specifically for the classification of skin lesions associated with STIs; studies involving tasks not designed for or not suitable for the early detection of STIs; and studies not utilising clinical images as inputs for the algorithm. The first 100 studies ranked by relevance that satisfied the inclusion criteria were reviewed.

After duplicate studies were removed, two authors (ML and X-YY) individually screened the titles and abstracts of the studies to identify eligible investigations; discrepancies were addressed via discussions. The full texts of the articles were then independently evaluated for inclusion, and disagreements were resolved via consensus or via referral to a third reviewer (W-KM).

This study involved the review of publicly available data and previously published studies; therefore ethical approval was not required in accordance with the ethical review regulations of the City University of Hong Kong.

Data analysis

Data were extracted from each paper independently by two reviewers (ML and one of the following coauthors: X-YY, M-NL, Y-ZC or Y-YZ); any discrepancies were addressed via discussions within the research team or by consulting a third reviewer (W-KM). The databases cited in the studies were cross-referenced with their original sources, and the data were obtained whenever possible. One author (ML) then consolidated and summarised the extracted information: the primary tasks in the studies, data (data type, dataset used, and metadata), model (model type and backbone architecture), and performance metrics. The primary tasks were categorised into three categories: detection of STIs, classification of skin conditions, and segmentation of skin lesions. The public or self-collected datasets used in the studies were evaluated, and the details of publicly available datasets were compared. Metadata including dataset (number of participants and images, dataset partitions, and diseases), demographic characteristics (age, sex, country of origin, race and ethnicity, and skin tone), and study design (retrospective or prospective) was recorded. Furthermore, we documented backbone models or self-designed architectures used, and the deep learning models employed in the research, which were cross-referenced with the model names listed in the Papers with Code repository.[26] Model architectures were categorised into traditional convolutional neural networks architectures, attention-based convolutional neural networks architectures, and Vision Transformer. We also examined transfer learning techniques and evaluated whether pre-trained weights were used and the origin dataset of the backbone model's pre-training. Performance assessment metrics, encompassing accuracy, precision, sensitivity, F1 score, specificity, and area under the receiver operating characteristic curve, were additionally analysed. For studies involving XAI, we evaluated XAI methods such as feature maps, attention maps, saliency maps, CAM-based methods, LIME, and SHAP.

We used the CheckList for Evaluation of image-based AI Reports in Dermatology (CLEAR Derm) for quality assessment; this checklist outlines best practices in image-based artificial intelligence (AI) development and assessment in dermatology.[27] The 25 checklist items were further elaborated, and criteria were established to determine the presence or absence of descriptions for each item (appendix pp 44–46). The quality of each study was evaluated based on the percentage of items with descriptions present. Subgroup analysis was performed to monitor quality trends and compare the quality of studies published in journals and conference proceedings, with statistical significance assessed using the Mann-Whitney *U* test, where a p value of less than 0·05 indicated significance.

Risk of bias and applicability concerns were evaluated using modified QUADAS-2 frameworks (appendix pp 52–57). That work was conducted in accordance with CLEAR Derm guidelines to facilitate a thorough AI-centred assessment. The quality assessment was carried out independently by two reviewers (ML and one of the following coauthors: X-YY, M-NL, Y-ZC or Y-YZ); any discrepancies were resolved via discussion within the research team or in consultation with a third reviewer (W-KM).

To assess the diagnostic accuracy of deep learning algorithms for STIs, we did a post-hoc meta-analysis of studies presenting contingency tables for deep learning algorithms with the highest documented accuracy in each study. Whenever feasible, binary diagnostic accuracy data for individual STI classes were extracted, and class-specific contingency tables were constructed based on the specified threshold values. These contingency tables included true positive, false positive, false negative, and true negative results, thereby facilitating the computation of sensitivity and specificity. When studies employed deep learning algorithms on diverse datasets or with varying levels of coverage within a single dataset, we extracted contingency tables for the algorithm with the highest accuracy.

To explore the sources of heterogeneity, we performed subgroup analyses by STI, dataset, number of images and diseases, study design, and publication type (conference proceedings or journals). Meta-regression was used to explore the impact of these factors on the overall heterogeneity. Heterogeneity was assessed using the I^2 statistic. A random-effects model was used because studies varied in task, data, and model. Next, we applied a unified hierarchical model to produce hierarchical summary receiver operating characteristic curves and we plotted the calculated sensitivity and specificity values.[28] Publication bias was assessed using the funnel plot asymmetry test. We also performed a sensitivity analysis by excluding studies with low quality assessment. The statistical analysis was conducted using the midas package in STATA (version 17.0).

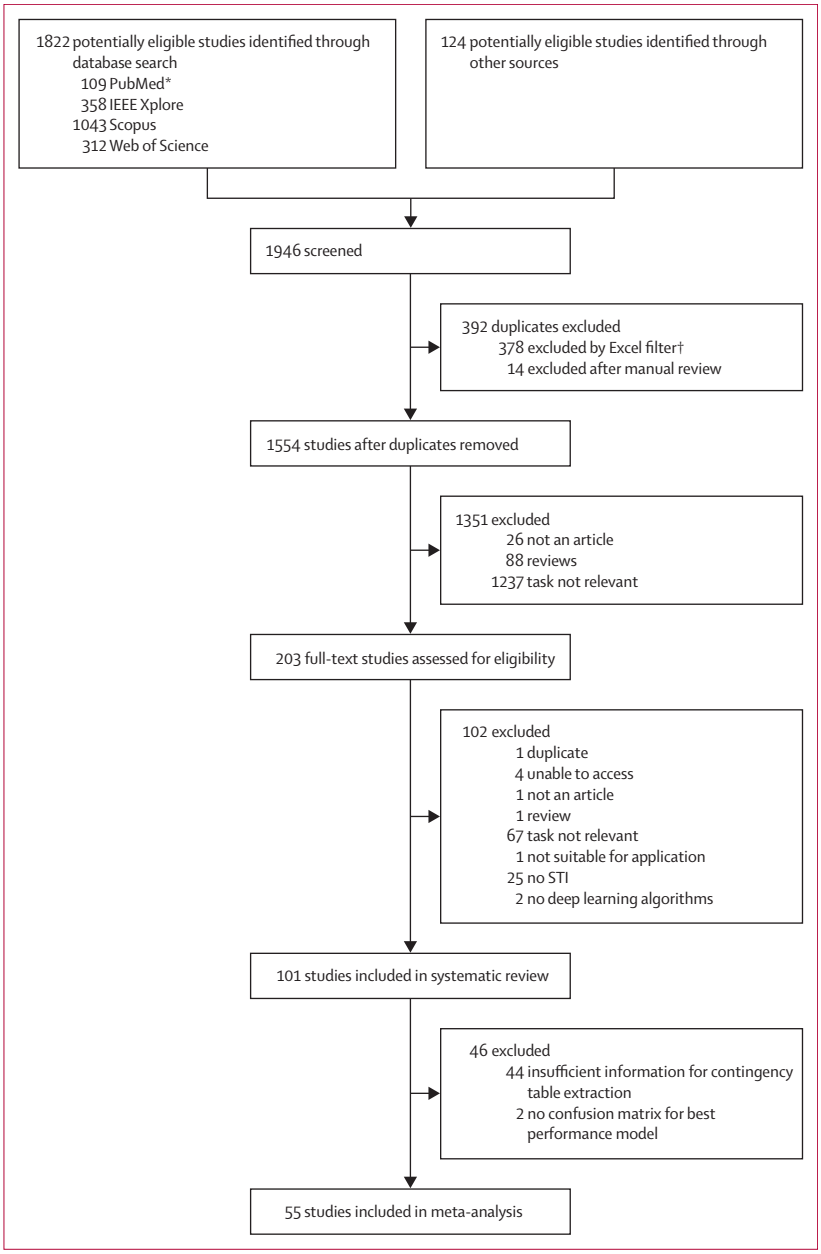


Figure 1: **PRISMA flow diagram for studies included in the systematic review and meta-analysis**
IEEE=Institute of Electrical and Electronics Engineers. STI=sexually transmitted infection.*PubMed citations were obtained from MEDLINE indexed journals and journals or manuscripts deposited in PubMed Central and the National Center for Biotechnology Information Bookshelf. †Screened on title, first author, publication year, and journal or conference.

Role of the funding source
The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

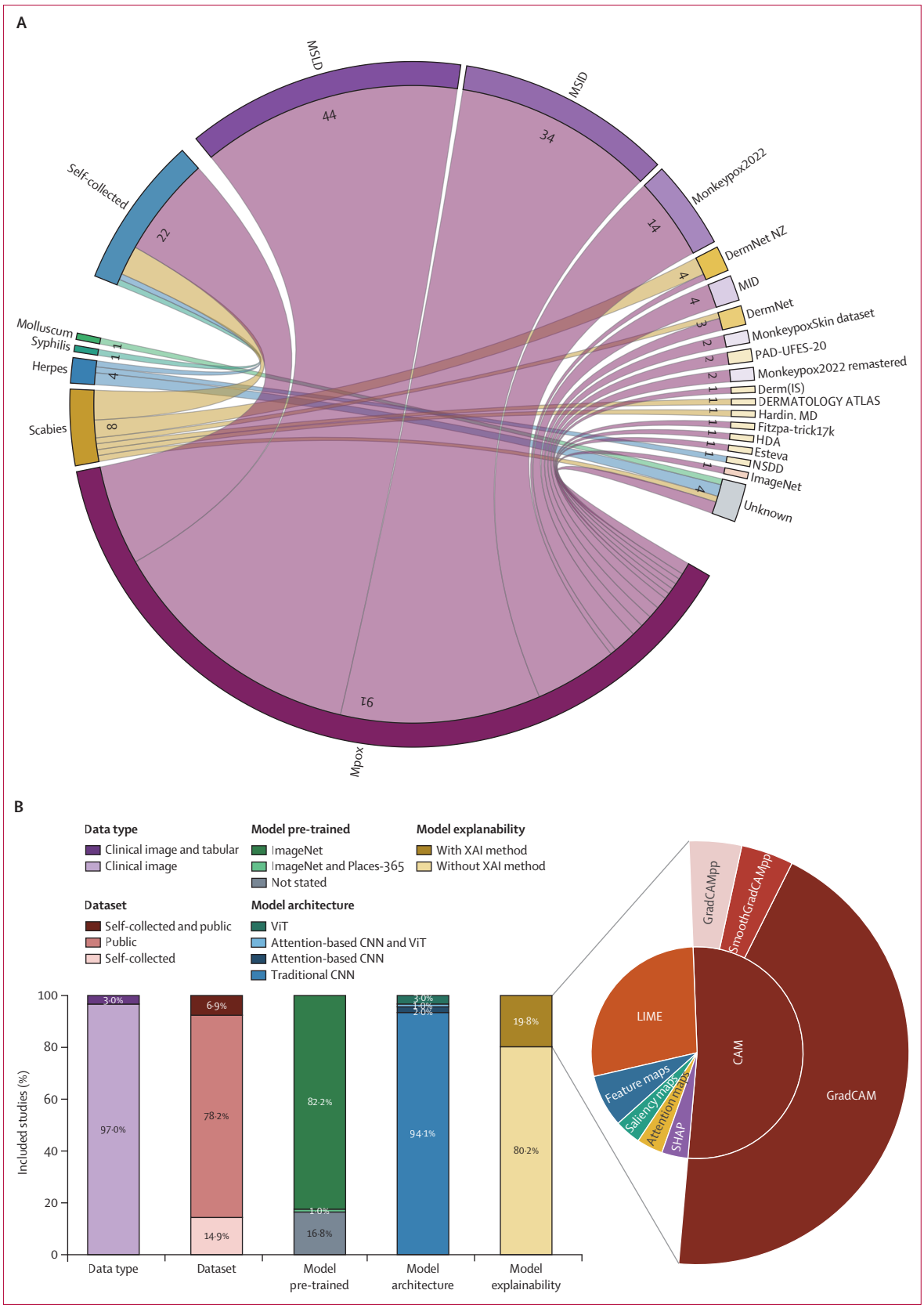
Results
Our search identified 1946 studies, of which 101 were included in this review. Of these, 57 studies contained sufficient data for contingency table extraction. However, two studies did not provide the confusion matrix for their best performing model.[29,30] As a result, 55 studies were included in meta-analysis (figure 1). Most studies focused on mpox (n=91; [88%]) and scabies was addressed in eight (8%) studies, herpes in four (4%) studies, syphilis in one (1%) study, and molluscum in one (1%) study (figure 2A). The primary task in the majority of studies (91 [90%] of 101 studies) was the detection of STIs, with nine studies (9%) on the classification of skin conditions and one study[31] (1%) on skin lesion segmentation.

Most studies (98 [97%] of 101 studies) exclusively used clinical images; three studies[14,32,33] (3%) incorporated both clinical images and tabular data such as demographic information and patient medical history. Public datasets were the primary source of data in 86 studies (85%); 22 studies (22%) incorporated self-collected datasets (figure 2B). Commonly used public datasets included Monkeypox Skin Lesion Dataset (MSLD), the Monkeypox Skin Images Dataset (MSID), and Monkeypox2022; in total, these datasets were used in 44 (44%) of 101 studies, 34 (34%) studies, and 14 (14%) studies, respectively (figure 2A; appendix pp 19–23). The publicly available datasets used in these studies only included clinical images; the DermNet NZ dataset also incorporated dermoscopic images. Class imbalance was observed in most datasets with the exception of the MonkeypoxSkin dataset (appendix pp 28–23).

Most studies (71 [70%] of 101 studies) used datasets with fewer than 1000 images, and only 39 (39%) studies reported a train-validation-test split. Most studies (84 [83%] of 101) included images with fewer than five diseases. Information on the number of participants was available in only 35 (35%) of 101 studies, and most (32 [32%] studies) involved fewer than 500 participants. Demographic characteristics, such as age (97 [96%] of 101 studies), sex (98 [97%] studies), country of origin (97 [96%] studies), race and ethnicity (99 [98%] studies), and skin tone (99 [98%] studies) were not reported in most studies. Prospective data collection was used in only two (2%) of 101 studies[34,35] (figure 2C; appendix pp 31–36).

Traditional convolutional neural networks architectures were used in all studies: ResNet (58 [57%] of 101 studies); VGGNet (48 [48%]), Inception (39 [39%]), and MobileNet (32 [32%]) were the most common. Vision Transformer architectures were employed in four (4%) of 101 studies, and attention-based convolutional neural networks architectures were used in three (3%) studies. Four studies[36–39] reported enhanced performance when utilising attention-based architectures (appendix pp 7, 37–42). Most studies (83 [82%] of 101 studies) utilised pretraining on the ImageNet dataset; only one (1%) study used both the ImageNet and Places-365 datasets (figure 2B).

20 (20%) of 101 studies incorporated XAI techniques, with CAM (12 [12%]), LIME (seven [7%]), and feature maps (two [2%]) being the common methods (figure 2B; appendix p 43).



(Figure 2 continues on next page)

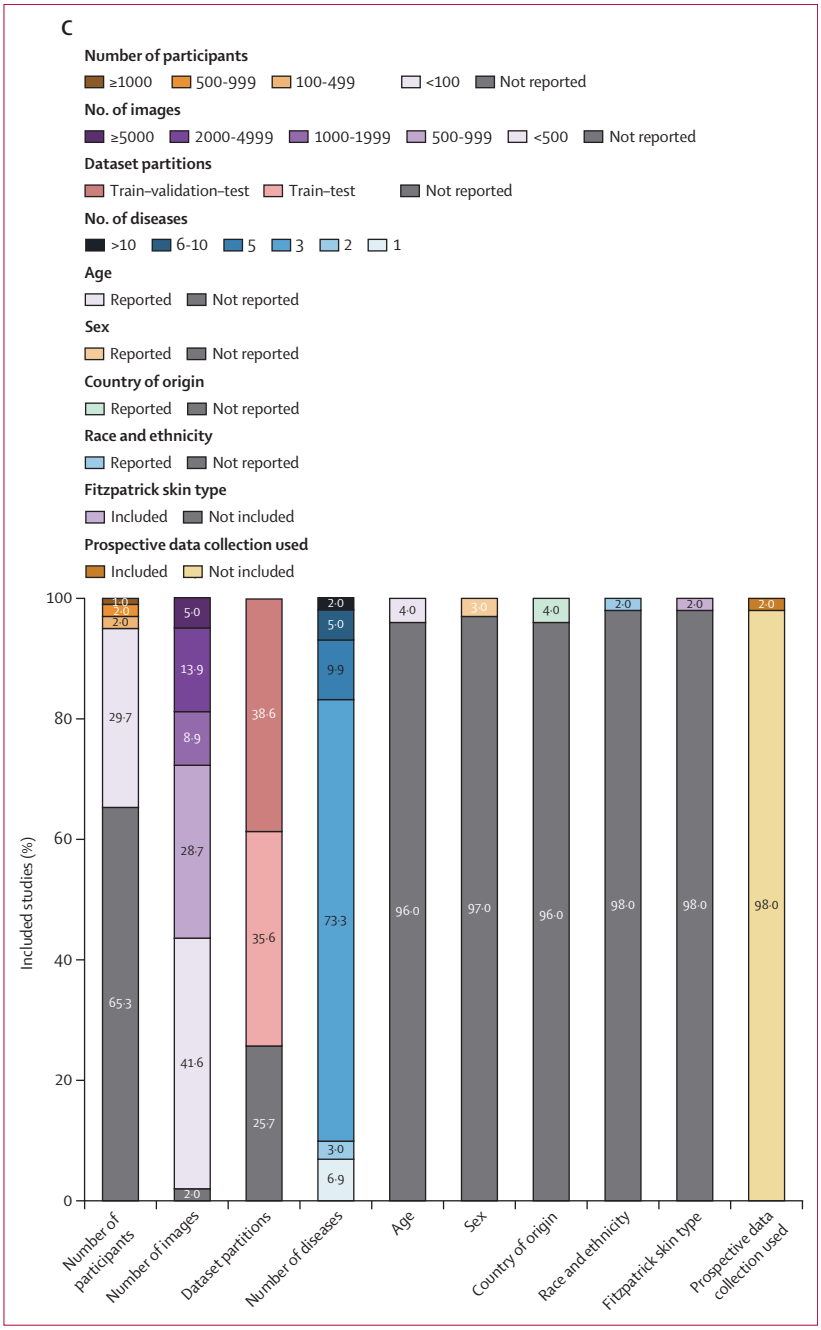


Figure 2: **Characteristics of included studies**
(A) Chord diagram showing the STIs and datasets used. (B) Data and model characteristics of the included studies and XAI methods employed in the included studies. (C) Characteristics of the included studies including dataset, demographic characteristics, and study design. CAM=Class Activation Maps. CNN=convolutional neural network. HDA=Hellenic Dermatological Atlas. LIME=Local Interpretable Model-agnostic Explanations. MID=Monkeypox Images Dataset. MSID=Monkeypox Skin Images Dataset. MSLD=Monkeypox Skin Lesion Dataset. NSDD=National Skin Disease Database. SHAP=SHapley Additive exPlanations. ViT=Vision Transformer. XAI=eXplainable Artificial Intelligence. YOLO=You Only Look Once object-detection network.

We evaluated the studies across four dimensions: data, technique, technique assessment, and application using the CLEAR Derm checklist. First, quality concerns were identified pertaining to the data; we observed the absence of multivendor or out-of-distribution images for testing in all included studies; only three (3%) of 101 studies used an external test set. Furthermore, no studies included descriptions of image artifacts, and most studies did not collect data on skin tone or procedures for assessing skin tone (100 [99%] of 101 studies), patient-level metadata (97 [96%] studies), or omitted technical acquisition details (93 [92%] studies). Dataset partitions were present in 69 (68%) of 101 studies; however, the sample sizes of the training, validation, and test sets were often unlisted (76 [75%] studies). Information regarding class distribution and balance (64 [63%] of 101 studies), and adequate elaboration on the preprocessing procedures (62 [61%] studies), was also often omitted. 72 (83%) of 87 of the studies that used public images adequately referenced the sources (appendix p 8).

In terms of technique, the labelling methods were not described in 94 (93%) of 101 studies; references to common or accepted diagnostic labels were absent in 99 (98%) studies. Detailed information about the algorithm development was only provided in 30 (30%) studies. In terms of technical assessments, only six (6%) of 101 studies provided access to their source code for public evaluation of the algorithm; 12 studies (12%) inaccurately reported their results (eg, did not specify use of the performance measures method, did not present corresponding performance measure results, presented results that contradicted the confusion matrix, or claimed the model had the best performance without reporting the results). Additionally, 57 (56%) of 101 studies did not compare their results with other similar studies, and bias assessments were not reported in 96 studies (95%). In terms of applications, most studies did not provide detailed descriptions of use cases and target conditions (80 [79%] of 101 studies), and potential impacts on the health-care team and patients (80 [79%]; appendix p 8).

Our quality assessments showed that studies only satisfied an average of 22·5% (SD 11·4%) of the checklist items. Subgroup analysis identified no significant difference in quality between studies published in 2022 (n=22) and 2023 (n=77; Mann-Whitney *U* test p=0·44), but studies published in journals (n=49) had better quality assessment results than those in conference proceedings (n=52; Mann-Whitney U test p=0·037).

We assessed the risk of bias and applicability concerns using the modified QUADAS-2 framework. A high risk of bias was prevalent in outcomes (100 [99%] of 101 studies), index test (60 [59%]), and participant domains (49 [49%]). Conversely, the reference standard domain displayed a relatively low risk of bias: 86 (85%) of 101 studies were categorised as being low risk, and 15 studies (15%) were categorised as being of unclear risk. In terms of applicability, most studies demonstrated high levels of concern across all three domains. Notably, the index test and outcomes domains were associated with the highest

proportion of studies with applicability concerns: 95 (94%) of 101 studies and 92 (91%) studies were categorised as high risk in those domains, respectively (appendix pp 9, 58–60).

We extracted 65 contingency tables from 55 studies, with 50 studies (58 tables) on mpox diagnosis, and five studies (six tables) on scabies (appendix pp 17–21). A low likelihood of publication bias was observed using a Deeks' funnel plot asymmetry test (p=0·68), suggesting symmetry in data and a low likelihood of publication bias (appendix p 10). The analysis identified a high level of heterogeneity, with I^2 values of 99·18% and 99·90% for sensitivity and specificity, respectively (p<0·0001; appendix p 10). The subgroup analyses identified persistently high levels of heterogeneity, and high sensitivity and specificity estimates across all subgroups. The meta-regression analysis showed the number of images used and study design were likely contributors to the observed heterogeneity (appendix pp 61–62).

The pooled sensitivity for mpox diagnosis was 0·97 (95% CI 0·95–0·98); the specificity was 0·99 (0·98–0·99), and the area under the curve (AUC) was 1·00 (0·99–1·00). The pooled sensitivity for scabies diagnosis was 0·95 (0·90–0·98); the specificity was 0·97 (0·86–0·99), and the AUC was 0·98 (0·97–0·99; appendix p 10).

A sensitivity analysis was conducted, excluding studies that had less than 20% of the CLEAR Derm checklist items where the description is present or did not show any low risk of bias and low applicability concerns in the modified QUADAS-2 domains. 43 contingency tables from 37 studies were included. A high likelihood of publication bias was observed (p<0·0001; appendix p 10). High levels of heterogeneity and high sensitivity and specificity estimates across all subgroups were also observed. Number of images used remained a likely source of heterogeneity (appendix p 10, appendix pp 12, 63–64).

Discussion

The early detection of STIs is crucial for mitigating disease outbreaks and planning effective interventions. Since 2019 there have been advancements in deep learning models, particularly those focused on the classification of skin lesions in images, which holds promise for the timely diagnosis of STIs. This investigation is the first systematic review, to the best of our knowledge, to evaluate deep learning algorithms for the early detection of STIs. Our analysis of 101 studies identified the diverse applications in classifying STIs such as mpox, scabies, and syphilis. Although most studies had satisfactory performance, particularly for mpox and scabies, limitations in generalisability were noted. Substantial quality concerns were observed across the data, technique, technique assessment, and application domains. Both the risk of bias and applicability concerns frequently manifested, particularly in the outcomes and index test domains. To improve study quality and generalisability, we propose a practical roadmap.

Challenges remain in the real-world application of these algorithms. Public datasets such as MSLD and MSID contain a small number of images and few disease categories, and all images were collected from internet-based sources with low quality, which could hinder the ability of deep learning models to learn robust and generalisable features. Furthermore, the paucity of patient-level metadata, including demographics such as age, sex, and skin tone restricts model performance evaluation across diverse populations. Class imbalance, underrepresentation of darker skin tones, and absence of inclusive samples, particularly for skin-of-colour and genital lesions, further constrain applicability. In real-world clinical settings, practical factors such as variability in cameras, light sources, and factors affecting skin lesions (eg, body hair, tattoos, piercings, and wounds) can affect algorithm accuracy.[40] Rather than relying on clean, curated datasets, it is preferable to validate deep learning models with diverse, noisy data from real-life patients and to continuously update the models until they are validated to assist clinicians in both hospital and remote settings.[41]

XAI methods can improve the transparency and reliability of deep learning models by indicating image regions related to classification.[42] However, rigorous evaluation of the accuracy and reliability of XAI is necessary before these methods are integrated into the clinical setting. For example, XAI can be evaluated by the overlap between XAI method segmentation and expert segmentation.[43] Clinical XAI guidelines were proposed, and 16 commonly used heatmap XAI techniques were found to be insufficient for clinical use.[44] Increased collaboration with end users can help design human-centred, transparent deep learning systems for explainable medical imaging AI.[45]

Our quality assessment utilising the CLEAR Derm checklist reveals that most studies lacked adequate descriptions of checklist items. This deficiency in quality was more prominent in studies published in conference proceedings than in journals, which may be due to the more rigorous peer-review process and fewer space constraints in journal publications. Although guidelines are currently being developed for reporting studies assessing deep learning models in diagnostic accuracy studies,[46,47] these guidelines primarily emphasise best practices and the absence of a comprehensive framework for practical implementation. To address this gap, we have devised a suggested roadmap for the design, development, and evaluation of deep learning algorithms in the early detection of STIs (figure 3). This roadmap offers a step-by-step, hands-on guide that emphasises key principles such as data quality, inclusive representation, transparent reporting, rigorous validation, and real-world applicability. Each step includes practical recommendations and highlights crucial considerations to help mitigate potential biases and enhance generalisability. By following this framework, researchers and developers can better align their work with established best practices in image-based AI development and evaluation for STI detection.

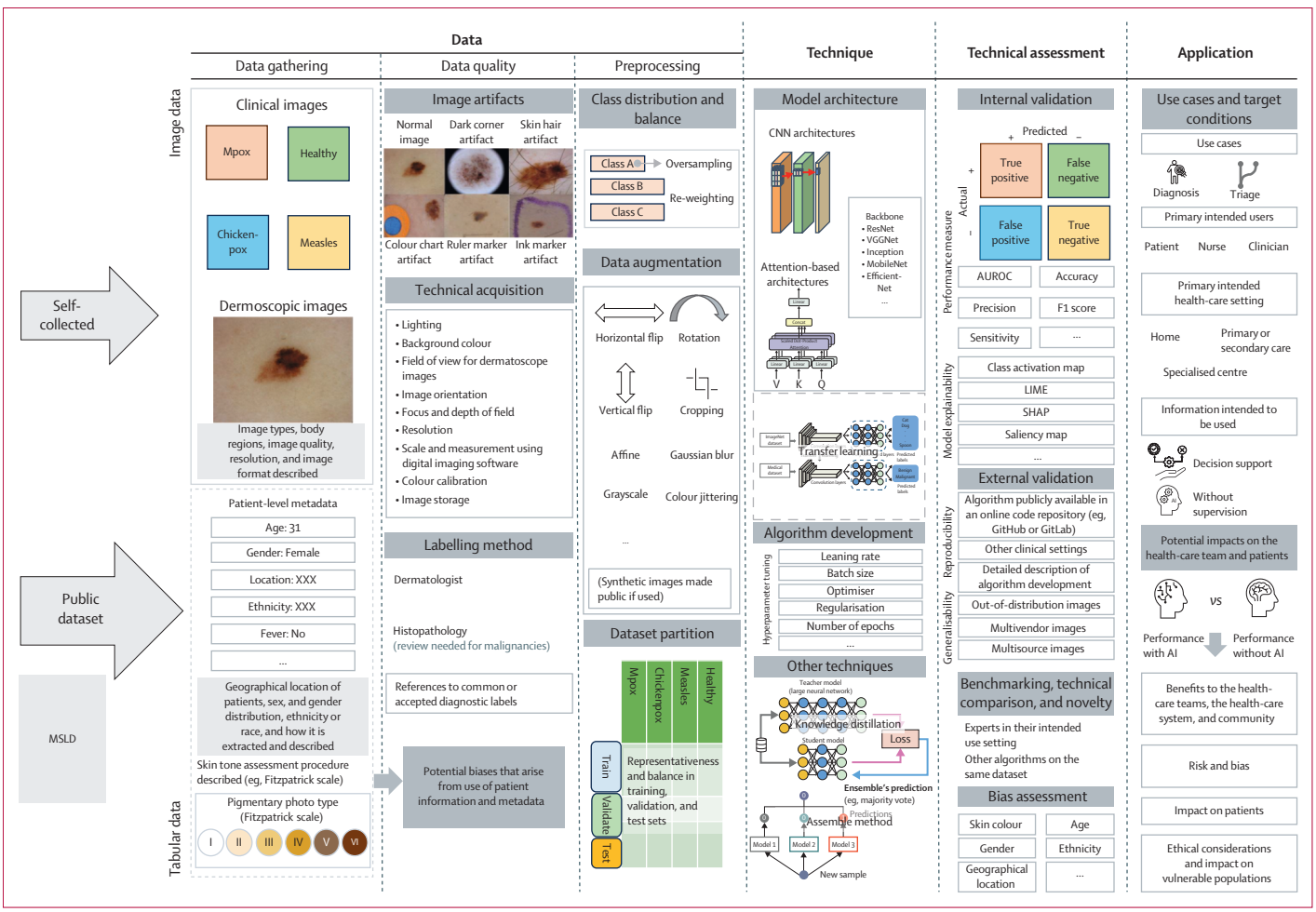


Figure 3: **Proposed roadmap for deep learning algorithm development for the early detection of STIs**
All photographs were obtained from the publicly available International Skin Imaging Collaboration 2017 Challenge dataset, which is released under the CC0 1.0 public-domain dedication. AI=artificial intelligence. AUROC=area under the receiver operating characteristic curve. CNN=convolutional neural networks. K=key. LIME=Local Interpretable Model-agnostic Explanations. MSLD=mpox skin lesion dataset. Q=query. SHAP=SHapley Additive exPlanations. STI=sexually transmitted infection. V=value.

Many studies have used traditional convolutional neural networks architectures such as different variations of ResNet and VGGNet; however, certain attention-based architectures such as Vision Transformer have shown superior performance compared with traditional architectures.[36,38,39] Additionally, integrating clinical images with tabular data such as demographic information and patient medical history within deep learning models can lead to enhanced performance.[14,48] Furthermore, various types of metadata can impact model accuracy in different ways. For example, metadata such as self-reported skin problems (eg, acne, hair loss, or rash) and a history of psoriasis were found to significantly affect accuracy. Training models with the input of both images and metadata yielded superior performance compared with training with images alone.[14] Furthermore, data on sexual behaviour is valuable for STI occurrence prediction.[49,50]

Nevertheless, our study has some limitations. Differences in datasets, partitioning methods, and classification tasks among the studies inhibited direct comparisons and introduced uncertainty in the estimates of diagnostic performance. The classification difficulty varied across included studies, with deep learning algorithms trained to distinguish a larger number of diseases and visually similar skin lesions using prospective study cohorts potentially performing worse than those focused on binary classification or fewer classes with less visually distracting images and retrospective datasets.[51,52] This raises concerns about the generalisability of the reported diagnostic performance to real-world settings and the overall estimates might overestimate diagnostic performance. Additionally, the majority of included studies focused on the diagnosis of mpox and scabies, with fewer evaluating deep learning for other STIs. This might limit the generalisability of our findings to the broader context of STI diagnosis. This bias could be due to the common asymptomatic nature of many STIs, and laboratory diagnosis generally being considered the best way to detect STIs.[53] Furthermore, we restricted our

search to selected databases and Google Scholar, which could have resulted in the omission of relevant studies not included in these sources. Additionally, the choice of keywords used in our search queries might have also limited the scope of our searches.

When implementing deep learning models in mobile applications, optimising model size and reducing parameters are crucial for rapid screening. In addition, local implementation of algorithms and the integration of privacy preserving techniques are imperative, given that some skin lesions occur in sensitive parts of the body. To date, deep learning models have predominantly been evaluated on limited datasets that lack diversity in skin tones, body locations, and geographical regions. It is therefore essential to create standardised datasets that encompass diversity to ensure equitable comparisons and reliable model generalisability. Moreover, establishing medical guidelines for AI-driven STI screening that outline specific use cases and target conditions is critical for the efficient implementation of such models.

Contributors

ML conceived and designed the study. W-KM supervised the study. ML and X-YY searched, screened, and assessed the publications. ML, X-YY, YZ-C, M-NL, and YY-Z extracted and analysed the data. ML, YZ-C, X-YY, MN-L, and YY-Z designed the figures. ML, X-YY, YZ-C, M-NL, and YY-Z wrote the first draft of the manuscript. CJPZ and JH revised the manuscript. All of the authors had access to the data presented in the manuscript. The raw data were collected and verified by ML, X-YY, YZ-C, M-NL, and YY-Z. All of the authors read the manuscript, provided feedback, and approved the final version.

Declaration of interests

We declare no competing interests.

Data sharing

Search terms, study characteristics (data, model, and performance metrics), quality assessment criteria and results, risk of bias assessment criteria and results, and meta-analysis findings are available in the appendix.

Acknowledgments

During the preparation of this work, the authors used ChatGPT to improve language and readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication. This study was partially supported by a strategic interdisciplinary research grant (7020093) from City University of Hong Kong.

References

1 Zheng Y, Yu Q, Lin Y, et al. Global burden and trends of sexually transmitted infections from 1990 to 2019: an observational trend study. *Lancet Infect Dis* 2022; **22:** 541–51.

2 Reddy K. Why early detection of STIs is critical? https://drsafehands.com/blog/early-detection-stis/ (accessed March 24, 2024).

3 de Vries HJ. Skin as an indicator for sexually transmitted infections. *Clin Dermatol* 2014; **32:** 196–208.

4 Asif S, Zhao M, Li Y, Tang F, Ur Rehman Khan S, Zhu Y. AI-Based approaches for the diagnosis of mpox: challenges and future prospects. *Arch Comput Methods Eng* 2024; **31:** 1–33.

5 Tuddenham S, Hamill MM, Ghanem KG. Diagnosis and treatment of sexually transmitted infections: a review. *JAMA* 2022; **327:** 161–72.

6 WHO. Guidelines for the management of symptomatic sexually transmitted infections. July 15, 2021. https://www.who.int/publications/i/item/9789240024168 (accessed June 20, 2025).

7 Kabir K, Kapten BS, Thet AM, Schairer J. Misdiagnosis and MSM: a systematic review assessing the diagnosis of IBD and sexually transmitted infectious colitis in gay men. *Am J Gastroenterol* 2023; **118:** S775.

8 Summers BMY. Sexually transmitted infection or urinary tract infection? Misdiagnosis of *chlamydia trachomatis* and *neisseria gonorrhoeae* infections in primary care practice. PhD thesis, University of Northern Colorado, 2017: 1–105.

9 Towns JM, Leslie DE, Denham I, Azzato F, Fairley CK, Chen M. Painful and multiple anogenital lesions are common in men with *Treponema pallidum* PCR-positive primary syphilis without herpes simplex virus coinfection: a cross-sectional clinic-based study. *Sex Transm Infect* 2016; **92:** 110–15.

10 Hood JE, Friedman AL. Unveiling the hidden epidemic: a review of stigma associated with sexually transmissible infections. *Sex Health* 2011; **8:** 159–70.

11 Layland EK, Carter JA, Perry NS, et al. A systematic review of stigma in sexual and gender minority health interventions. *Transl Behav Med* 2020; **10:** 1200–10.

12 Hauser K, Kurz A, Haggenmüller S, et al. Explainable artificial intelligence in skin cancer recognition: a systematic review. *Eur J Cancer* 2022; **167:** 54–69.

13 Chadaga K, Prabhu S, Sampathila N, et al. Application of artificial intelligence techniques for monkeypox: a systematic review. *Diagnostics (Basel)* 2023; **13:** 824.

14 Liu Y, Jain A, Eng C, et al. A deep learning system for differential diagnosis of skin diseases. *Nat Med* 2020; **26:** 900–08.

15 Aziz MA, Pujiono P, Soeleman MA, Rosidin S, eds. Scabies skin classification using SVM method with GLCM feature extraction. International Seminar on Application for Technology of Information and Communication; Sept 16–17, 2023.

16 Alex Berg JD, Li F-F. ImageNet Large Scale Visual Recognition Challenge 2010 (ILSVRC2010). https://image-net.org/challenges/LSVRC/2010/index.php (accessed Jan 16, 2024).

17 Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Adv Neural Inf Process Syst* 2012; 25: 1097–105.

18 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv* 2014; published online Sept 4 (preprint).

19 He K, Zhang X, Ren S, Sun J, eds. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition; Jun 27–30, 2016.

20 Guo MH, Xu TX, Liu JJ, et al. Attention mechanisms in computer vision: a survey. *Comput Vis Media (Beijing)* 2022; **8:** 331–68.

21 Dosovitskiy A. An image is worth 16x16 words: transformers for image recognition at scale. *arXiv* 2020; published online Oct 22 (preprint).

22 Atasever S, Azginoglu N, Terzi DS, Terzi R. A comprehensive survey of deep learning research on medical image analysis with focus on transfer learning. *Clin Imaging* 2023; **94:** 18–41.

23 Exten C, Pinto CN, Gaynor AM, Meyerson B, Griner SB, Van Der Pol B, and the Board of Directors of the American Sexually Transmitted Diseases Association. Direct-to-consumer sexually transmitted infection testing services: a position statement from the American Sexually Transmitted Diseases Association. *Sex Transm Dis* 2021; **48:** e155–59.

24 Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Int J Surg* 2021; **88:** 105906.

25 Liu M, Yi X, Li M, et al. Early detection of sexually transmitted infections from skin lesions with deep learning: a systematic review and meta-analysis. https://www.crd.york.ac.uk/PROSPERO/view/CRD42024496966 (accessed June 5, 2025).

26 Papers with Code. Image classification on ImageNet. https://paperswithcode.com/sota/image-classification-on-imagenet (accessed Jan 22, 2024).

27 Daneshjou R, Barata C, Betz-Stablein B, et al. CheckList for Evaluation of image-based AI Reports in Dermatology: CLEAR Derm Consensus Guidelines from the International Skin Imaging Collaboration Artificial Intelligence Working Group. *JAMA Dermatol* 2022; **158:** 90.

28 Harbord RM, Whiting P, Sterne JA, et al. An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. *J Clin Epidemiol* 2008; **61:** 1095–103.

29 Fisranda F, Austin F, Sadrawi M. Comparison analysis of YOLO, ResNet, and EfficientNet for contagious skin diseases. International Conference on Cybernetics and Intelligent System; Oct 6–7, 2023.

30 Ahsan MM, Uddin MR, Ali MS, et al. Deep transfer learning approaches for monkeypox disease diagnosis. *Expert Syst Appl* 2023; **216:** 119483.

31 Yanagisawa Y, Shido K, Kojima K, Yamasaki K. Convolutional neural network-based skin image segmentation model to improve classification of skin diseases in conventional and non-standardized picture images. *J Dermatol Sci* 2023; **109:** 30–36.

32 Waldamichael FG, Kebede SR, Ayano YM, Demissie MT, Debelee TG. A multi-input architecture for the classification of skin lesions using resnets and metadata. Pan African Conference on Artificial Intelligence; Oct 4–5, 2022.

33 Rai S, Joshi RC, Dutta MK. PoxDetector: a deep convolutional neural network for skin lesion classification using android application. International Conference on Advancement in Computation & Computer Technologies; May 5–6, 2023, 363–68.

34 Thieme AH, Zheng Y, Machiraju G, et al. A deep-learning algorithm to classify skin lesions from mpox virus infection. *Nat Med* 2023; **29:** 738–47.

35 Yotsu RR, Ding Z, Hamm J, Blanton RE. Deep learning for AI-based diagnosis of skin-related neglected tropical diseases: a pilot study. *PLoS Negl Trop Dis* 2023; **17:** e0011230.

36 Kundu D, Siddiqi UR, Rahman MM, eds. Vision transformer based deep learning model for monkeypox detection. International Conference on Computer and Information Technology; Dec 17–19, 2022.

37 Surati S, Trivedi H, Shrimali B, Bhatt C, Travieso-González CM. An enhanced diagnosis of monkeypox disease using deep learning and a novel attention model senet on diversified dataset. *Multimodal Technol Interact* 2023; **7:** 75.

38 Dan R, Wu Q, Ji X, et al, eds. LSANet: Lesion-Specific Attention Network for Monkeypox Categorization. 2022 IEEE Smartworld, Ubiquitous Intelligence & Computing, Scalable Computing & Communications, Digital Twin, Privacy Computing, Metaverse, Autonomous & Trusted Vehicles (SmartWorld/UIC/ScalCom/DigitalTwin/PriComp/Meta). Institute of Electrical and Electronics Engineer, 2022.

39 Haque ME, Ahmed MR, Nila RS, Islam S, eds. Human monkeypox disease detection using deep learning and attention mechanisms. International Conference on Computer and Information Technology; Dec 17–19, 2022.

40 Mehta N, Gupta S, Kularathne Y. The role and impact of artificial intelligence in addressing sexually transmitted infections, nonvenereal genital diseases, Sexual Health, and Wellness. *Indian Dermatol Online J* 2023; **14:** 793–98.

41 Goyal M, Knackstedt T, Yan S, Hassanpour S. Artificial intelligence-based image classification methods for diagnosis of skin cancer: challenges and opportunities. *Comput Biol Med* 2020; **127:** 104065.

42 Hauser K, Kurz A, Haggenmüller S, et al. Explainable artificial intelligence in skin cancer recognition: a systematic review. *Eur J Cancer* 2022; **167:** 54–69.

43 Saporta A, Gui X, Agrawal A, et al. Benchmarking saliency methods for chest x-ray interpretation. *Nat Mach Intell* 2022; **4:** 867–78.

44 Jin W, Li X, Fatehi M, Hamarneh G. Guidelines and evaluation of clinical explainable AI in medical image analysis. *Med Image Anal* 2023; **84:** 102684.

45 Chen H, Gomez C, Huang C-M, Unberath M. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. *NPJ Digit Med* 2022; **5:** 156.

46 Sounderajah V, Ashrafian H, Aggarwal R, et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. *Nat Med* 2020; **26:** 807–08.

47 Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open* 2016; **6:** e012799.

48 Soe NN, Yu Z, Latt PM, et al. Evaluation of artificial intelligence-powered screening for sexually transmitted infections-related skin lesions using clinical images and metadata. *BMC Med* 2024; **22:** 296.

49 Everett BG. Sexual orientation disparities in sexually transmitted infections: examining the intersection between sexual identity and sexual behavior. *Arch Sex Behav* 2013; **42:** 225–36.

50 Hogben M, Leichliter J, Aral SO. An overview of social and behavioral determinants of STI. In: Cristaudo A, Giuliani M, eds. Sexually transmitted infections: advances in understanding and management. Springer, 2020: 25–45.

51 Meng X, Yao J. Impact of classification difficulty on the weight matrices spectra in deep learning and application to early-stopping. *J Mach Learn Res* 2023; **24:** 1107–46.

52 Kleppe A, Skrede O-J, De Raedt S, Liestøl K, Kerr DJ, Danielsen HE. Designing deep learning studies in cancer diagnostics. *Nat Rev Cancer* 2021; **21:** 199–211.

53 Wagenlehner FM, Brockmeyer NH, Discher T, Friese K, Wichelhaus TA. The presentation, diagnosis, and treatment of sexually transmitted infections. *Dtsch Arztebl Int* 2016; **113:** 11–22.