ISSN: 2577 - 8005

Research Article

Medical & Clinical Research

Artificial Intelligence for Skin Prick Test Interpretation: A Systematic Review of Accuracy, Validation and Readiness

Shaymaa Al-tharwane[1] and Mohammed S Al Abadie[2]*

[1]*University Hospitals of Leicester NHS Foundation Trust & The Midlands Medical Academy United Kingdom*

[2]*University of Wolverhampton, North Cumbria Integrated Care NHS Foundation Trust, United Kingdom*

***Corresponding Author**
Mohammed S Al Abadie, PhD, FRCP, Clinical Professor of Dermatology, University of Wolverhampton, United Kingdom.

Submitted: 2026, Jul 27; **Accepted:** 2026, Aug 11; **Published:** 2026, Aug 17

Citation: Al-tharwane,S., Al Abadie, M.S., (2026). Artificial Intelligence for Skin Prick Test Interpretation: A Systematic Review of Accuracy, Validation and Readiness. *Med Clin Res, 11*(8), 1-9.

Abstract

Background: *The skin prick test (SPT) is the first-line in-vivo investigation for immunoglobulin E-mediated sensitisation, but its reading is constrained by interobserver variability, subjectivity in delineating irregular reactions, and the time burden of measuring multiple wheals. Artificial intelligence (AI), particularly deep learning and computer vision, offers a potential means of standardising measurement and expanding capacity. Current gaps in validation, calibration and reporting, however, confine these systems to exploratory roles rather than deployable clinical instruments.*

Aim: *This review aims to critically evaluate the application of AI to skin prick testing, focusing on diagnostic and measurement performance, methodological strengths, and current limitations regarding clinical integration.*
Materials and Methods: Following PRISMA 2020 guidelines, MEDLINE (PubMed) was searched for records from inception to July 2026. Eligible studies included original research involving human participants that utilised recognised AI or computational techniques, such as convolutional neural networks (CNNs), fully convolutional networks or classical image-processing pipelines, for the detection, segmentation, measurement or interpretation of SPT reactions, and reported quantifiable performance metrics.

Results: *Nine reports, representing eight unique studies, met the inclusion criteria Methodologies spanned automated wheal segmentation, smartphone and three-dimensional imaging, thermography-based classification, and one device-integrated AI-assisted readout workflow. Segmentation performance was moderate, with Dice similarity coefficients ranging from 0.66 to 0.81, while lesion-detection sensitivity was frequently low, from 0.56 to 0.70, despite reported accuracies exceeding 0.96 that reflect class imbalance rather than clinical performance. The largest study reported 85.0% sensitivity and 98.4% specificity against physician measurement, falling to 77.4% sensitivity for allergen wheals, with a 3.7-fold faster readout. Critical limitations included the complete absence of external validation and calibration, small single-centre datasets, weak or single-reader reference standards, and no quantification of performance across skin phototypes.*

Conclusion: *AI demonstrates credible technical capability in wheal detection and measurement but is currently constrained by dataset heterogeneity, internal-only validation and inadequate reporting. Consequently, AI currently functions best as a decision-support system rather than an autonomous diagnostic tool, and no included system is ready for routine clinical implementation.*

Keywords: Artificial intelligence, Skin prick test, Allergy diagnosis, Computer vision, Deep Learning, machine learning, Clinical decision support, Medical imaging, Allergy testing, Diagnostic accuracy

1. Introduction

Allergic disease occupies an important space in clinical practice: common, clinically impactful, and, despite decades of methodological refinement, still constrained by the subjectivity and labour intensity of skin prick testing [1]. The SPT remains the first-line in-vivo investigation for identifying immunoglobulin E-mediated sensitisation, and its interpretation rests on measuring the wheal, and in some protocols the surrounding erythema, provoked by an allergen relative to positive and negative controls [2]. Clinical relevance is conventionally attributed to the longest wheal diameter and to allergen-specific wheal size [3,4]. Yet interobserver variability persists even among experienced clinicians, with discrepancies reported in both reaction measurement and final interpretation [5]. Test devices and technique introduce further variability, irregular reactions bearing pseudopodia are difficult to delineate reproducibly, and reading multiple wheals imposes a time burden that creates structural barriers to timely diagnosis [6,7]. These limitations have made the field increasingly receptive to computational support systems capable of standardising interpretation, reducing human error and expanding diagnostic capacity.

Attempts to automate SPT reading span four decades, progressing from adhesive-tape planimetry to digital photography, three-dimensional surface imaging and infrared thermography [8]. Contemporary computer vision, and in particular encoder–decoder segmentation architectures such as U-Net, has renewed this ambition, while ubiquitous smartphone cameras make image acquisition inexpensive and scalable [9]. Recent studies confirm that convolutional neural networks can detect and measure wheals with clinically useful agreement: a large multicentre evaluation of a device-integrated readout reported 85.0% sensitivity and 98.4% specificity against physician measurement, with a 3.7-fold reduction in reading time [10]. Others have segmented wheal and erythema simultaneously from smartphone photographs, or proposed wheal area, rather than diameter, as a more faithful index of the reaction [11, 12]. These findings suggest that AI could meaningfully reduce subjectivity in SPT readings, standardise assessments across centres, and support scalable or remote evaluation workflows.

Despite accelerating progress, current evidence is limited by small, single-centre datasets, inconsistent reporting of skin phototypes, and the frequent evaluation of models on the same data used for their development, which risks overfitting. Most studies lack external validation and do not address crucial challenges such as harmonised imaging protocols, probability calibration, algorithmic transparency, or real-world integration into clinical workflows. Because wheal and erythema are recognised principally by colour and contrast against surrounding skin, algorithmic performance may plausibly vary with skin pigmentation, an equity consideration that has attracted little attention in this field. These recurrent deficiencies prompted the development of appraisal and reporting instruments specific to AI, including QUADAS-2 for diagnostic accuracy, PROBAST and its AI extension for risk of bias in prediction models,TRIPOD+AI for reporting model development and validation [13-15]. CLAIM for medical-imaging AI and DECIDE-AI for the early clinical evaluation of AI decision-support systems [16,17]. These gaps underscore why AI systems in SPT currently function as exploratory tools rather than deployable clinical instruments.

This article aims to critically evaluate how artificial intelligence is being applied to skin prick testing, with emphasis on methodological strengths, diagnostic and measurement performance, and current limitations. By integrating the available evidence, the review will establish a balanced, evidence-based perspective on how AI might realistically augment, not replace, clinical expertise in the diagnosis of IgE-mediated allergic disease.

2. Methods

Following PRISMA 2020, we searched MEDLINE (PubMed) for records from inception to July 2026 [18]. using controlled vocabulary and free-text terms such as “skin prick test”, “wheal”, “artificial intelligence” and “machine learning” (Supplementary Data 1). To capture methodological work published outside the biomedical literature, the strategy was adapted for Embase, Web of Science, Scopus, IEEE Xplore and arXiv, and reference lists of all included reports were hand-searched. A PRISMA flow diagram (Supplementary Figure 1) summarises the selection.

Category	Search Terms
Skin Prick Test	(“Skin Tests”[Mesh] OR “Intradermal Tests”[Mesh] OR “skin prick test” OR “skin prick testing” OR “skin prick tests” OR “prick test” OR “prick testing” OR “prick tests” OR “prick-puncture test” OR “puncture test” OR “allergy skin test” OR “allergy skin testing” OR “allergic skin test” OR “skin allergy test” OR “cutaneous allergy test” OR “skin prick automated test” OR SPAT)
Wheal and allergic skin reaction	(wheal OR wheals OR weal OR weals OR flare OR “wheal and flare” OR erythema OR “skin reaction” OR “cutaneous reaction” OR “allergic reaction” OR urticaria OR “Urticaria”[Mesh] OR “Hypersensitivity, Immediate”[Mesh] OR “Allergens”[Mesh] OR “Immunoglobulin E”[Mesh] OR “IgE-mediated” OR sensitization OR sensitisation OR aeroallergen OR “Rhinitis, Allergic”[Mesh])

Machine learning	("Artificial Intelligence"[Mesh] OR "Machine Learning"[Mesh] OR "Deep Learning"[Mesh] OR "Supervised Machine Learning"[Mesh] OR "Unsupervised Machine Learning"[Mesh] OR "Support Vector Machine"[Mesh] OR "Neural Networks, Computer"[Mesh] OR "Expert Systems"[Mesh] OR "Fuzzy Logic"[Mesh] OR "Pattern Recognition, Automated"[Mesh] OR "Decision Support Systems, Clinical"[Mesh] OR "Artificial Intelligence" OR "Machine Learning" OR "Deep Learning" OR "Neural Networks" OR "Convolutional Neural Networks" OR "U-Net" OR "Fully Convolutional Network" OR "Supervised Learning" OR "Unsupervised Learning" OR "Transfer Learning" OR "Ensemble Learning" OR "Reinforcement Learning" OR "Generative Adversarial Network" OR "Random Forest" OR "Support Vector Machine" OR "Kernel Fisher" OR "Predictive Modeling" OR "Decision Support Systems" OR "Diagnostic Decision Support" OR "Computer-Aided Diagnosis" OR "Automated Diagnosis" OR "Machine Intelligence" OR "Augmented Intelligence" OR "Large Language Models" OR "Foundation Models")
Imaging and automation	("Image Processing, Computer-Assisted"[Mesh] OR "Diagnosis, Computer-Assisted"[Mesh] OR "Photography"[Mesh] OR "Smartphone"[Mesh] OR "Thermography"[Mesh] OR "Imaging, Three-Dimensional"[Mesh] OR "Algorithms"[Mesh] OR "Computer Vision" OR "Image Analysis" OR "Image Segmentation" OR "Semantic Segmentation" OR "Object Detection" OR "Image Recognition" OR "Pattern Recognition" OR "Computer-Aided Measurement" OR "Automated Reading" OR "Automated Readout" OR "Automated Measurement" OR "Automated Detection" OR "Digital Photography" OR smartphone OR thermography OR "infrared imaging" OR "3D imaging" OR "three-dimensional imaging" OR planimetry OR "wheal detection" OR "wheal segmentation" OR "wheal measurement" OR "wheal area" OR "wheal diameter")

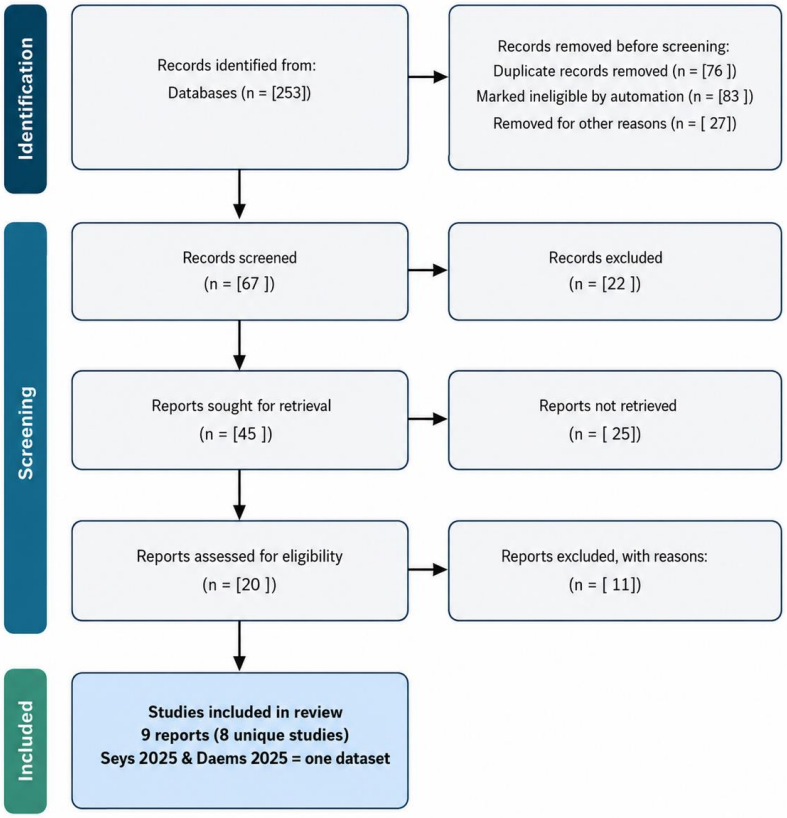
2.1. Inclusion and Exclusion Criteria

Studies were eligible for inclusion if they presented original research applying artificial intelligence or computational methods to skin prick testing. Eligible work had to involve human participants or human-derived data, including clinical SPT images acquired by smartphone, dedicated device, three-dimensional imaging or thermography. Only studies that implemented recognised AI or automated image-analysis techniques—such as deep learning models including convolutional neural networks and U-Net variants, fully convolutional networks, kernel methods, or classical image-processing pipelines applied to automated reading—were included. To ensure evaluability, studies were required to report quantifiable performance metrics, such as accuracy, sensitivity, specificity, area under the curve, Dice similarity coefficient, intersection over union, or measurement agreement. Both fully automated and semi-automated methods were eligible. Full conference papers and preprints reporting original SPT performance data in sufficient methodological detail to appraise algorithm development and validation were considered, and their peer-review status was recorded.

Studies were excluded if they did not apply AI or computational methodologies, relied solely on traditional statistical analyses, or failed to report measurable algorithm outputs. Non-original research, including reviews, commentaries, editorials, conference abstracts without primary data, and expert opinion pieces, was excluded. Work based on animal models, cadaveric skin, phantoms or wholly synthetic datasets without clinical linkage was not considered. Studies of patch testing were excluded unless SPT was also assessed, as were AI allergy-prediction studies using only serum specific IgE or clinical records without SPT. Studies were also excluded if they lacked relevance to skin prick testing despite incorporating AI, if they did not involve human-derived data, or if AI was mentioned only speculatively without development, application or evaluation of a method. Where two or more reports described the same underlying dataset, all were retained for extraction but were treated as linked companion reports of a single study and were not counted as independent.

Supplementary Figure 1

Figure 1. PRISMA 2020 flow diagram (template)



3. Results

Nine reports met the inclusion criteria and reported the application of AI or computational methods to skin prick test detection, measurement and interpretation (Table 1). Two of these, Seys et al. and Daems et al., report the same multicentre automated-device cohort and are treated as linked companion reports of a single study; the nine reports therefore represent eight unique studies [10,19]. The included studies spanned multiple methodological domains, including automated segmentation of wheals and erythema, smartphone-based image analysis, three-dimensional surface imaging, thermography-based reaction classification, and device-integrated readout workflows.

The studies utilised AI and machine learning (ML) to address several specific stages of the SPT workflow, spanning reaction detection, boundary delineation, dimensional measurement and diagnostic interpretation.

A primary objective was the automated detection and delineation of the wheal from photographic images. Bulan aimed to improve wheal detection from images acquired in uncontrolled settings by maximising the contrast between wheal and surrounding erythema. Bulan and Artan pursued the same goal using graph-based segmentation, selecting the wheal-containing region by its distinctive colour, texture and shape attributes [20, 21]. Huttunen et al. set a related but narrower objective, seeking an optimal greyscale projection and an elliptic shape model that would approximate the ellipse a clinician draws when measuring by ruler [22].

Lee et al. aimed to segment wheals and erythema simultaneously, noting that no prior study had reported the segmentation of both elements individually, and that erythema carries interpretive information distinct from wheal size [11]. Gomes et al. focused on inferring wheal dimensions from smartphone photographs without specialised equipment, and additionally investigated whether wheal area is a better indicator of the reaction than the diameters used by the standard protocol [12]. Peña et al. sought to replace the parameter-tuned detection stage of an existing three-dimensional reading system with a learned model [23].

Seys et al. and Daems et al. addressed the full readout chain for an automated pricking device, the former validating an AI-assisted readout against physician measurement and evaluating its operational use, the latter developing the underlying segmentation model and quantifying the contribution of the device's multi-illumination image capture [10, 19]. Neumann et al. set a different objective, classifying each allergen site as positive or negative from correlated thermal and visible-spectrum images, rather than measuring a wheal at all [24].

The utilised AI techniques included classical and kernel-based image processing, deep learning architectures, and hybrid pipelines combining a learned segmentation stage with rule-based measurement, depending on the data modality.

Bulan applied a colour-space transformation from RGB to YCbCr, discarding luminance, followed by principal component analysis on the chrominance channels and morphological operations [21]. Bulan and Artan employed normalized cuts with contour cues for initial segmentation, then selected the wheal region by optimising a cost function over colour-name, Leung–Malik and MR8 texture features [21]. Huttunen et al. used a kernel Fisher discriminant with a radial basis function kernel for greyscale projection, followed by least-squares fitting of a Gaussian surface whose isosurfaces provided candidate ellipses [22].

Convolutional neural networks were established as the standard architecture for the more recent image analyses [10, 11,12, 19, 23, 24]. Lee et al. selected Attention U-Net for its capacity to prioritise small regions of interest and suppress the skin background, preceded by contrast-limited adaptive histogram equalisation [11]. Gomes et al. used a fully convolutional network with a VGG-16 encoder rather than U-Net [12]. Peña et al. and Neumann et al. both used U-Net, the former on depth images following removal of the global arm surface, the latter for segmenting allergen application sites prior to classification by a custom convolutional classifier [22, 23]. Daems et al. designed a custom U-Net accepting the 32 differently illuminated images of a single test concatenated as 96 input channels, and deliberately separated pattern recognition from wheal delineation by following the network with an interpretable, rule-based detection step [19]. Seys et al. combined this segmentation map with white-box algorithms that match detected regions to known prick locations and estimate the longest diameter using the Feret maximum diameter [10].

The studies analysed employed three principal imaging modalities: conventional two-dimensional photography, including consumer smartphones; specialised acquisition hardware; and thermal imaging.

Photographic studies varied widely in scale and standardisation. Bulan analysed images from 36 patients and Bulan and Artan from 50 patients, both using an iPhone 5 under uncontrolled indoor lighting with drawn calibration marks and histamine as the sole challenge [20, 21]. Lee et al. analysed 46 smartphone images from 33 participants at two hospitals, standardised to 512 × 512 pixels [11]. Gomes et al. deliberately avoided standardising camera, resolution, focus or lighting, collecting 1,461 photographs across multiple smartphone brands and models, augmented fourfold to 5,844 images, with a 3 cm reference tag providing scale [12].

Specialised acquisition was used by three reports. Peña et al. reconstructed the forearm surface using a fringe-projection system of two cameras, two lasers and a projector, training on 14 images containing 47 wheals [23]. Seys et al. and Daems et al. used the Skin Prick Automated Test device, which delivers twelve pressure-controlled pricks simultaneously and then captures 32 images of the forearm under distinct illumination angles, exploiting the shadow lines cast at wheal edges; their shared cohort comprised 868 patients with 10,416 manually annotated wheals, with a further 95-patient prospective cohort in the clinical report [10, 19]. Neumann et al. used a long-wave infrared thermographic system pixel-correlated with a visible camera, acquiring 404 thermal and 404 visible images from 100 patients across 1,584 allergen sites [24].

Segmentation performance was moderate across all reports that measured it. Lee et al. reported Dice similarity coefficients of 0.7079 for wheals and 0.6636 for erythema [11]. Gomes et al. reported a Dice coefficient of 0.8085 and an intersection over union of 0.6965 on 144 correctly detected wheals [12]. Daems et al. reported a Dice coefficient of 0.787 when using all 32 device images, compared with 0.717 for a single uniformly lit image, but detection accuracy fell steeply as the overlap criterion tightened, from 91.8% at an intersection-over-union threshold of 0.5 to 74.3% at 0.7 and 38.8% at 0.8 [13].

Lesion-detection sensitivity was consistently modest relative to headline accuracy. Lee et al. reported sensitivities of 0.5621 for wheals and 0.5787 for erythema against accuracies of 0.9985 and 0.9660 respectively, and Gomes et al. reported a sensitivity of 70.29% against a segmentation accuracy of 85.88% [11, 12]. This dissociation reflects severe class imbalance: wheals occupy a small fraction of the image, so a model predicting background almost everywhere attains very high pixel accuracy while missing much of the lesion. Bulan reported a mean pixel accuracy of 94%, exceeding the 87% of a previously published hue-based method, without reporting sensitivity or specificity [20]. Bulan and Artan reported a detection accuracy of 0.9000 with a standard deviation of 0.3030, alongside a sensitivity of 0.6723 and specificity of 0.9396; the magnitude of the standard deviation indicates frequent outright detection failures [21].

Diagnostic accuracy against a clinical reference standard was reported by one study. Seys et al. compared the AI-derived longest wheal diameter with the treating physician's measurement across 2,604 wheals, applying a previously validated 4.5 mm positivity threshold, and observed a sensitivity of 85.0%, a specificity of 98.4%, a positive predictive value of 93.3% and a negative predictive value of 96.2%; restricted to allergen rather than control wheals, sensitivity fell to 77.4% [10]. Agreement between AI and physician diameters was moderate to strong (Pearson r = 0.83; median absolute difference 0.2 mm). In a separate prospective cohort, physicians adjusted 5.8% of AI measurements, changing the test interpretation in only 0.5% of cases, and readout was 3.7 times faster than manual measurement with significantly reduced inter- and intra-observer variability.

The highest discriminatory performance was reported for a task that does not measure the wheal at all. Neumann et al. reported a classification accuracy of 93.6%, an area under the receiver operating characteristic curve of 0.98 and an average precision of 0.97 under patient-level ten-fold and leave-one-out cross-validation, with thermal-only input performing at least as well as combined thermal and visible input [24]. Measurement agreement was reported by Peña et al., who observed a mean absolute difference of 0.966 mm and a mean coefficient of variation of 7.29% against

physician measurement, although three of twelve tabulated wheals were not detected and coefficients of variation reached 15 to 17% for irregular reactions [23]. Gomes et al. reported strong correlation between the model and an allergist-guided reference ($\rho = 0.88$), with Bland–Altman analysis showing bias-centred differences uncorrelated with lesion size, and demonstrated that circular and elliptical approximations of the standard protocol deviate by 13.44 ± 13.95% for irregular wheals bearing pseudopodia [12]. Huttunen et al. reported only a bespoke area-based error metric, with a best value of approximately 0.185, on seven wheals [22].

Notably, no included report performed external validation, none reported probability calibration or decision-curve analysis, only two quantified inter-rater reliability, and none made code, model weights or data available in usable form.

S/NO	Author	Study aim	AI Algorithm Used	Sample size	Materials used	Accuracy
1	Huttunen et al. [22]	To develop a semi-automatic segmentation method for allergic reactions using an optimal greyscale projection and an elliptic shape model	Kernel Fisher discriminant (RBF kernel) + Gaussian/ elliptic shape model	7 wheals	2D colour photography; allergens not reported	Not reported (custom area error ≈0.185)
2	Bulan [21]	To improve wheal detection from prick test images captured by digital cameras in an uncontrolled environment	Classical: RGB→YCbCr transformation, principal component analysis on Cb/Cr, morphological operations	36 patients	Histamine only; iPhone 5, 8 MP	Accuracy: 94% (vs 87% for prior hue-based method)
3	Bulan and Artan [22]	To detect the wheal by segmenting the image and selecting the region with distinctive colour, texture and shape	Normalized cuts + region selection (colour names, LM and MR8 texture filters)	50 patients	Histamine only; iPhone 5, 8 MP	Accuracy: 0.9000 ± 0.3030, sensitivity: 0.6723 ± 0.2399, specificity: 0.9396 ± 0.0489
4	Peña et al. [23]	To improve the automated reading of the SPT by replacing the wheal detection stage with a deep learning model	Convolutional neural network (U-Net) + ellipse fitting	14 training and 4 test images (47 wheals)	3D fringe projection (2 cameras, 2 lasers, 1 projector)	Not reported; mean diameter difference 0.966 mm, mean CV 7.29%
5	Neumann et al. [24]	To classify prick allergic reactions using correlated visible-spectrum and thermal images of the forearm	U-Net (segmentation) + custom convolutional classifier	100 patients (1,584 allergen sites)	12 allergens + histamine + negative control; LWIR thermography and visible imaging	Accuracy: 93.6%, ROC AUC: 0.98, average precision: 0.97
6	Gomes et al. [12]	To infer wheal dimensions by deep-learning segmentation of smartphone photographs, reducing dependence on human interpretation	Fully convolutional network (VGG-16 encoder) + pixel clustering	1,461 photographs (5,844 augmented); validation on 30 images, 150 wheals; patients not reported	EAACI protocol, panel not specified; mixed smartphones with 3 cm reference tag	Detection accuracy: 94.12%, segmentation accuracy: 85.88%, sensitivity: 70.29%, specificity: 98.94%, DSC: 0.8085, IoU: 0.6965
7	Lee et al. [11]	To segment wheals and erythema simultaneously in SPT images captured by smartphone	Attention U-Net (two models) with CLAHE preprocessing	46 images from 33 participants	Allergens not reported; smartphone photography	Wheal: accuracy 0.9985, sensitivity 0.5621, specificity 0.9995, DSC 0.7079. Erythema: accuracy 0.9660, sensitivity 0.5787, specificity 0.9797, DSC 0.6636

8	Seys et al. [10]	To develop and validate an AI-assisted readout method to support physicians in interpreting skin reactions following automated prick testing	U-Net segmentation + rule-based algorithms (Feret maximum diameter)	963 patients (11,556 wheals)	Aeroallergen panel, histamine and glycerol–saline controls; SPAT device, 32 illumination angles	Accuracy: 95.4%, sensitivity: 85.0% (77.4% allergen wheals), specificity: 98.4%, PPV: 93.3%, NPV: 96.2%, Pearson r = 0.83
9	Daems et al. [19]	To detect and delineate wheals by leveraging the 32-image, multi-illumination data modality of the automated prick test device	Custom U-Net (96-channel input) + interpretable rule-based wheal detection	868 patients (10,416 annotated wheals)	Aeroallergen panel; SPAT device, 32 illumination angles	Dice: 0.787 (32 images) vs 0.717 (single image); accuracy at IoU 0.5: 91.8%, at 0.7: 74.3%, at 0.8: 38.8%

Table 1: Key characteristics of the included studies

4. Discussion

The collective evidence demonstrates a credible technical foundation for automating the reading of the skin prick test, extending from classical colour-space and graph-based segmentation to deep learning models embedded within a regulated device. Convolutional neural networks have shown consistent capability in delineating wheals from photographs, particularly when preprocessing enhances the contrast between lesion and skin, or when acquisition hardware is designed to cast informative shadows at the wheal boundary [11,19]. Where a learned segmentation stage is coupled to an interpretable, rule-based measurement step, the resulting longest-diameter estimate correlates well with physician measurement and reduces both inter- and intra-observer variability and reading time [10]. Similar performance patterns have been observed in other dermatological imaging tasks, where CNNs demonstrate strong pattern-recognition capacity in visually constrained problems [9]. At the level of the reaction itself, one study has shown that thermographic imaging can discriminate positive from negative allergen sites with high internal discrimination, while another has demonstrated that automated area measurement is invariant to wheal shape and departs substantially from the circular and elliptical approximations on which the standard protocol depends [12, 24].

Despite high reported performance metrics, the translational relevance of many of these models remains limited. Accuracies exceeding 0.96 are, in the segmentation studies, an artefact of class imbalance rather than a measure of clinical performance: the same models report lesion sensitivities between 0.56 and 0.70, meaning that a third to almost a half of the lesion, or of the lesions, is missed [11,12]. This phenomenon is well recognised across medical AI, where task simplification, curated datasets and internal validation inflate apparent performance while obscuring fragility under real-world conditions [16,17]. Even the strongest study, benchmarked against physician measurement in a prospective multicentre cohort, missed 15% of positive wheals overall and 22.6% of positive allergen wheals; for a test used in part to exclude sensitisation, a false-negative rate of this magnitude is clinically material rather than cosmetic [10]. Boundary agreement is likewise looser than headline figures suggest: detection accuracy in the best-resourced segmentation study fell from 91.8% to 38.8% as the required overlap rose from a permissive to a stringent threshold [19]. Furthermore, the reports that measure irregular reactions consistently find that algorithms struggle with pseudopodia, precisely the reactions in which manual measurement is least reliable, so that AI at present approximates rather than surpasses the human standard it is intended to replace [23,12].

Dataset heterogeneity and bias substantially constrain generalisability across all modalities reviewed. Image-based studies show marked variability in acquisition, ranging from a purpose-built multi-illumination device to consumer smartphones photographed under uncontrolled indoor lighting, introducing inconsistencies in colour calibration, lighting and resolution that degrade cross-domain performance [11,19,20]. Demographic bias is particularly pronounced. Lee et al. studied an exclusively Asian cohort and observed qualitative failure of erythema detection in darker skin; Seys et al. and Daems et al. identify hyperpigmentation and darker skin tone among the principal causes of misclassification while acknowledging that such cases are under-represented in their dataset; and Bulan and Artan state that darker skin was absent from their sample altogether [10, 11, 19, 21]. Because wheal and erythema are recognised largely through colour contrast, this is not a peripheral concern but a foreseeable failure mode with equity implications, and it remains unquantified in every study identified. Two studies used histamine alone as the challenge (Bulan, Bulan and Artan, and three did not report which allergens were tested, further limiting transferability to clinical allergen panels. No study included children, despite the SPT being widely used in paediatric practice [2,11, 20-23].

From a translational standpoint, current AI applications in skin prick testing are best conceptualised as decision-support systems rather than autonomous diagnostic tools. The most advanced implementation is explicitly physician-in-the-loop, with the clinician reviewing and adjusting each measurement before validating the result, and the authors of the smartphone-based approach take the same position, stating that segmentation should

always be interpreted and validated by a health professional [10, 12]. Integration into routine practice is impeded by the absence of standardised imaging protocols, despite clear evidence that performance is highly sensitive to illumination and preprocessing (Daems et al.), and by a near-total absence of transparency: no included report provides usable code, model weights or an accessible dataset, with two explicitly withholding code to protect commercialisation [10,19]. Independent replication of any included result is therefore currently impossible. Only one system is described as a regulated medical device; the remainder are research prototypes. Regulatory frameworks increasingly emphasise intended use and clinical impact rather than algorithmic sophistication, and none of the reviewed systems has reported the human-factors evaluation, safety analysis under live use, or clinical-impact outcomes that early-stage clinical evaluation of an AI decision-support system requires [15, 17].

The evidence base is constrained by structural limitations that curtail robustness and external validity. Reference standards are heterogeneous and, in several cases, weak: they range from three trained annotators with 87% concordance (Daems et al.) through a single allergy specialist without any reliability estimate (Lee et al.) to a non-medical expert (Huttunen et al.), while one study derives its reference from a pen contour drawn by the same allergist whose judgement the model reproduces, raising the prospect of incorporation bias (Gomes et al.) [11, 12, 19, 22]. Sample sizes span seven wheals (Huttunen et al.) to 963 patients (Seys et al.), with several studies too small to support stable estimates [11,22,23]. Validation is internal in every study without exception; the most rigorous designs employ patient-level stratified splits or leave-one-out cross-validation (Neumann et al.; Daems et al.), but the weakest evaluates performance on the same split used for early stopping, with no held-out test set and unclear control of data leakage across augmented images from the same participant (Lee et al.) [11,24, 19]. Probability calibration and decision-curve analysis are reported by none, and correlation coefficients are in places presented as evidence of agreement, which they are not [10,12]. Across all domains, reliance on internal validation and inconsistent reporting of demographic and skin-type variables mirrors broader weaknesses in clinical AI research and precludes confident assessment of fairness and generalisability [13, 14, 16]. The strongest study, finally, was funded by the manufacturer of the device it evaluates, with several authors employed by or holding equity in that company, and the same is true of the thermographic [10, 19, 24].

5.Conclusion

Artificial intelligence has demonstrated credible technical capability in supporting selected components of skin prick testing, particularly wheal detection, boundary delineation and dimensional measurement, and in one instance the readout of a regulated automated device within a physician-in-the-loop workflow. However, its translational readiness is constrained by moderate segmentation performance, sensitivity that is far lower than headline accuracy implies once class imbalance is accounted for, uniformly internal validation, an absence of calibration, weak or single-reader reference standards, and unquantified performance across skin phototypes and in children. Current evidence supports AI as a means of reducing observer variability and improving efficiency, rather than as a means of redefining diagnostic authority, and no included system is ready for routine clinical implementation. Prospective, externally validated and calibrated studies, with transparent reporting, agreement analysis where measurement is the output, and performance stratified by skin tone and age, are required before these technologies can be recommended for clinical use. Overall, the clinical value of AI in skin prick testing will be determined less by headline performance metrics than by alignment between task definition, data provenance, reference-standard validity and regulatory context.

References

1. Sánchez-Borges, M., Martin, B. L., Muraro, A. M., Wood, R. A., Agache, I. O., Ansotegui, I. J., ... & Rosenwasser, L. (2018). The importance of allergic disease in public health: an iCAALL statement. *World Allergy Organization Journal, 11*(1), 8.
2. Heinzerling, L., Mari, A., Bergmann, K. C., Bresciani, M., Burbach, G., Darsow, U., ... & Lockey, R. (2013). The skin prick test–European standards. *Clinical and translational allergy, 3*(1), 3.
3. Konstantinou, G. N., Bousquet, P. J., Zuberbier, T., & Papadopoulos, N. G. (2010). The longest wheal diameter is the optimal measurement for the evaluation of skin prick tests. *International archives of allergy and immunology, 151*(4), 343-345.
4. Haahtela, T., Burbach, G. J., Bachert, C., Bindslev-Jensen, C., Bonini, S., Bousquet, J., ... & Heinzerling, L. M. (2014). Clinical relevance is associated with allergen-specific wheal size in skin prick testing. *Clinical & Experimental Allergy, 44*(3), 407-416.
5. McCann, W. A., & Ownby, D. R. (2002). The reproducibility of the allergy skin test scoring and interpretation by board-certified/board-eligible allergists. *Annals of allergy, asthma & immunology, 89*(4), 368-371.
6. Nelson, H. S., Lahr, J., Buchmeier, A., & McCormick, D. (1998). Evaluation of devices for skin prick testing. *Journal of allergy and clinical immunology, 101*(2), 153-156.
7. Carr, W. W., Martin, B., Howard, R. S., Cox, L., Borish, L., & Immunotherapy Committee of the American Academy of Allergy, Asthma and Immunology. (2005). Comparison of test devices for skin prick testing. *Journal of allergy and clinical immunology, 116*(2), 341-346.
8. Justo, X., Díaz, I., Gil, J. J., & Gastaminza, G. (2016). Prick test: evolution towards automated reading. *Allergy, 71*(8), 1095-1102.
9. Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Cham: *Springer international publishing.*
10. Seys, S. F., Hox, V., Chaker, A. M., De Greve, G., Lemmens, W., Poirrier, A. L., ... & Van Gerven, L. (2025). Artificial

Intelligence (AI)-assisted readout method for the evaluation of skin prick automated test results. *Nature Communications, 16*(1), 8637.
11. Lee, Y. H., Shim, J. S., Kim, Y. J., Jeon, J. S., Kang, S. Y., Lee, S. P., ... & Kim, K. G. (2025). Allergy wheal and erythema segmentation using attention U-Net. *Journal of imaging informatics in medicine, 38*(1), 467-475.
12. Gomes, R. H. M., Perger, E. L. P., Vasques, L. H., Gagete, E., & Simões, R. P. (2024). Deep learning method applied to autonomous image diagnosis for prick test. *Life, 14*(10), 1256.
13. Whiting, P. F., Rutjes, A. W., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., ... & QUADAS-2 Group*. (2011). QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of internal medicine, 155*(8), 529-536.
14. Wolff, R. F., Moons, K. G., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., ... & PROBAST Group†. (2019). PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Annals of internal medicine, 170*(1), 51-58.
15. Collins, G. S., Moons, K. G., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., ... & Logullo, P. (2024). TRIPOD+ AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *bmj, 385.*
16. Mongan, J., Moy, L., & Kahn Jr, C. E. (2020). Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiology: *Artificial Intelligence, 2*(2), e200029.
17. Vasey, B., Nagendran, M., Campbell, B., Clifton, D. A., Collins, G. S., Denaxas, S., ... & McCulloch, P. (2022). Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nature medicine, 28*(5), 924-933.
18. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *bmj, 372*.
19. Daems, R., Seys, S., Hox, V., Chaker, A., De Greve, G., Lemmens, W., ... & Demeester, T. (2025, June). Improved Allergy wheal detection for the skin prick automated test device. *In International Conference on Artificial Intelligence in Medicine (pp. 116-120). Cham: Springer Nature Switzerland.*
20. Bulan, O. (2014, March). Improved wheal detection from skin prick test images. *In Image Processing: Machine Vision Applications VII* (Vol. 9024, pp. 138-147). SPIE.
21. Bulan, O., & Artan, Y. (2016, September). Wheal detection from skin prick test images using normalized-cuts and region selection. *In 2016 IEEE International Conference on Image Processing* (ICIP) (pp. 1250-1253). IEEE.
22. Huttunen, H., Ryynänen, J. P., Forsvik, H., Voipio, V., & Kikuchi, H. (2011, May). Kernel fisher discriminant and elliptic shape model for automatic measurement of allergic reactions. In Scandinavian Conference on Image Analysis (pp. 764-773). Berlin, Heidelberg: *Springer Berlin Heidelberg.*
23. Pena, J. C., Pacheco, J. A., & Marrugo, A. G. (2021, October). Skin prick test wheal detection in 3D images via convolutional neural networks. *In 2021 IEEE 2nd International Congress of Biomedical Engineering and Bioengineering (CI-IB&BI) (pp. 1-4). IEEE.*
24. Neumann, Ł., Nowak, R., Stępień, J., Chmielewska, E., Pankiewicz, P., Solan, R., & Jahnz-Różyk, K. (2022). Thermography based skin allergic reaction recognition by convolutional neural networks. *Scientific reports, 12*(1), 2648.