

Artificial Intelligence in Nailfold Capillaroscopy: A Scoping Review¹ of Validation, Reproducibility, and Clinical Translation

Zahra Emrani

University of Tehran

Seyed Kamaledin Setarehdan

University of Tehran

Habibollah Jafari-Varzaneh

Islamic Azad University of Najafabad

Abdolamir Karbalaie

`abdolamir.karbalaie@ki.se`

Karolinska Institutet

Systematic Review

Keywords: nailfold videocapillaroscopy, machine learning, deep learning, computer vision, microvascular imaging, capillary segmentation, quantitative image analysis, participant-level data splitting

Posted Date: July 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10348706/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Artificial Intelligence in Nailfold Capillaroscopy: A Scoping Review of Validation, Reproducibility, and Clinical Translation

Zahra Emrani¹, Seyed Kamaledin Setarehdan¹, Habibollah Jafari-Varzaneh², Alimohammad Fatemi³,
Abdolamir Karbalaie^{4,*}

¹Control and Intelligent Processing centre of Excellence, School of Electrical and Computer Engineering, College of
Engineering, University of Tehran, Tehran, Iran

²Department of Computer Engineering, Na.C., Islamic Azad University, Najafabad, Iran

³Department of Rheumatology, Alzahra Hospital, Isfahan University of Medical Sciences, Isfahan, Iran

⁴Department of Clinical Science, Intervention and Technology (CLINTEC), Karolinska Institutet, 17177 Stockholm,
Sweden

Corresponding author: Abdolamir Karbalaie, Department of Clinical Science, Intervention and Technology
(CLINTEC), Karolinska Institutet. Email: abdolamir.karbalaie@ki.se

Abstract

Background: Nailfold capillaroscopy (NFC) and nailfold videocapillaroscopy (NVC) provide a non-invasive view of peripheral microvascular architecture, but manual interpretation is time-consuming, observer-dependent, and sensitive to image-acquisition conditions. Artificial intelligence (AI) has increasingly been used for capillary detection, segmentation, quantitative measurement, disease classification, and deployment-oriented analysis. This scoping review mapped AI applications in NFC/NVC and examined disease scope, task families, validation design, data-splitting practice, reproducibility, device portability, and barriers to clinical translation.

Methods: A scoping review was conducted in accordance with the PRISMA Extension for Scoping Reviews (PRISMA-ScR). PubMed, Ovid MEDLINE, Scopus, the Web of Science Core Collection, and IEEE Xplore were searched from inception to 28 March 2026. Records were managed in Covidence. Eligible studies were peer-reviewed journal articles or full conference proceedings reporting original, extractable AI results. Abstract-only, non-original, preprint-only, duplicate or companion reports, and insufficiently reported records were excluded from the primary evidence set.

Results: The searches yielded 3,059 records. After removal of 1,754 duplicates, 1,305 records underwent title/abstract screening and 1,202 were excluded. All 103 reports progressing to full-text/source assessment were reviewed; 60 were excluded from the primary evidence set and 43 studies were included, comprising 37 journal articles and six full conference proceedings. Principal task families were classification/prediction (n=19), detection/quantification (n=8), segmentation (n=7), dynamic/flow analysis (n=6), image-quality/preprocessing (n=2), and dataset/benchmark development (n=1). Fourteen studies addressed systemic sclerosis/Raynaud phenomenon-spectrum applications and 29 addressed general microcirculation

or diseases beyond systemic sclerosis. Under strict coding, one study provided independent external clinical validation, four were multicentre, 11 directly compared outputs with experts, four provided public code, two provided public data, and five explicitly confirmed participant-level separation, and 28 of the 43 included studies (65.1%) met none of the six indicators.

Conclusions: AI-NFC has expanded markedly in architecture, clinical scope, and deployment ambition, including new applications in autoimmune disease, diabetes, neonatal retinopathy, amyloidosis, and dynamic blood-cell analysis. Nevertheless, explicitly reported clinical-translation evidence remains constrained by sparse independent external validation, infrequent confirmation of participant-level splitting, incomplete reporting, limited open resources, and heterogeneous devices and reference standards.

Keywords: nailfold videocapillaroscopy; machine learning; deep learning; computer vision; microvascular imaging; capillary segmentation; quantitative image analysis; participant-level data splitting

1 Background

Microvascular dysfunction is a central feature of many systemic, autoimmune, and metabolic diseases. Because capillaries form the interface between the arterial and venous circulation, structural and functional abnormalities in these vessels can provide early information about tissue perfusion, endothelial injury, inflammation, and disease progression. The nailfold is a particularly informative anatomical window because the distal-row capillary loops run nearly parallel to the skin surface and can therefore be visualised non-invasively. Nailfold capillaroscopy (NFC) and nailfold videocapillaroscopy (NVC) have consequently become important techniques for assessing peripheral microvascular architecture in clinical rheumatology and microcirculation research [1, 2].

The strongest clinical role of NVC is in systemic sclerosis (SSc) and Raynaud’s phenomenon (RP). In SSc, characteristic capillaroscopic abnormalities—including giant capillaries, dilated loops, microhaemorrhages, avascular areas, capillary loss, and disorganised architecture—are used to support early diagnosis, classification, prognosis, and disease monitoring [3–5]. The early, active, and late scleroderma patterns described by Cutolo and colleagues provide the morphological taxonomy that underpins much of the clinical and computational literature in this field [3, 4]. Beyond SSc, NFC/NVC has also been investigated in RP, juvenile dermatomyositis (JDM), systemic lupus erythematosus (SLE), diabetes mellitus, hypertension, and other conditions involving microvascular impairment [1, 6–8].

Despite this clinical value, conventional NFC/NVC interpretation remains challenging. Although the method appears simple and repeatable, its interpretation depends on image quality, acquisition protocol, examiner training, device characteristics, and the definitions used for morphological and quantitative parameters. Many clinically meaningful variables—including capillary density, capillary width, loop shape, haemorrhage, tortuosity, ramification, and avascularity—depend on the consistent identification of individual capillary structures and their anatomical position within the distal row [1, 2]. Among these variables, capillary density is particularly important because it provides a direct measure of capillary loss and has been linked to disease severity and complications in SSc and other microvascular conditions [2]. However, density estimation also illustrates the

69 practical difficulty of manual assessment: even the apparently simple task of counting capillaries per millimeter
70 can become ambiguous when loops are crossed, branched, tortuous, partially visible, or poorly contrasted.

71 These ambiguities in capillary identification and density estimation have motivated earlier efforts to improve
72 standardisation and reliability. Previous work on capillary counting proposed structured approaches such as
73 the elliptical broken line method, which aimed to define clearer rules for identifying capillary apices and
74 reducing ambiguity in density estimation [9]. Similarly, image-enhancement work showed that improving NVC
75 image quality can increase intra- and inter-observer reliability, indicating that measurement variability reflects
76 examiner training as well as image-processing and acquisition factors [10]. Together, these studies illustrate
77 a broader methodological issue: before NFC/NVC can be widely used as a quantitative biomarker, the field
78 requires more reproducible acquisition, preprocessing, annotation, segmentation, and measurement pipelines.

79 Artificial intelligence (AI) has emerged as natural extensions of this methodological trajectory. Rather than
80 relying only on visual inspection or manual measurement, computational models can support image-quality con-
81 trol, capillary detection, segmentation, density estimation, morphometric quantification, pattern classification,
82 and disease prediction. Early automated and semi-automated approaches were based on handcrafted features
83 and classical image-processing pipelines, whereas recent studies have used convolutional neural networks (CNNs),
84 U-Net-based segmentation, YOLO-family detectors, ResNet classifiers, Vision Transformers (ViTs), gradient-
85 boosted feature classifiers, optoacoustic three-dimensional imaging, and edge-deployment systems [11–15]. This
86 technical progression suggests that AI may help reduce subjectivity, improve measurement speed, and extend
87 capillaroscopy beyond specialist centres.

88 However, the progress of AI in NFC/NVC should not be interpreted only through headline accuracy. The
89 available studies differ substantially in clinical population, disease label, imaging device, acquisition protocol,
90 annotation source, unit of analysis, validation design, and reporting standard. This concern is particularly
91 important because AI-NFC datasets often contain multiple images, videos, fingers, or fields from the same
92 participant. In such repeated-measures settings, validation strategies that ignore participant identity can allow
93 correlated observations from the same individual to appear in both training and test folds, producing data
94 leakage and inflated performance estimates [16]. Therefore, for NFC/NVC models, reporting whether splitting
95 was performed at the image, video, finger, patient, centre, or device level is not a technical detail but a core
96 requirement for judging generalisability.

97 A prior systematic review identified 10 AI-NFC studies from 105 retrieved records and highlighted substan-
98 tial heterogeneity in study populations, model development, validation, feature extraction, and performance
99 reporting [17]. Since then, the evidence base has expanded to include additional disease applications, low-cost
100 and portable devices, edge-deployment systems, open datasets, and externally validated software. Adjacent
101 reviews have focused on broader nail diagnostics, general microcirculation imaging, or rheumatology AI rather
102 than systematically mapping AI-NFC validation, split-unit reporting, reproducibility, device portability, and
103 clinical-translation readiness. Table 1 summarises how the present scoping review extends the prior literature.

Table 1: Positioning of the present scoping review relative to the most directly relevant prior review literature.

Review source	Main focus	Gap addressed by the present review
Helmy et al. (2023) [18]	Technical review of deep learning and computer vision for microcirculation image analysis (> 50 papers surveyed).	Provides a dedicated NFC/NVC evidence map with clinical scope, AI task families, validation design, split-unit reporting, reproducibility, and translation.
Emam et al. (2024) [17]	Direct systematic review of AI in nailfold capillaroscopy (105 records; 10 included studies; medRxiv preprint).	Updates the direct evidence base to 43 studies and evaluates external validation, participant-level separation, open resources, device portability, sample-size reporting, and translation readiness.
Koker et al. (2024) [19]	Structured review of AI in pediatric rheumatology (four databases to May 2024; final included-study = 33).	Provides an NFC/NVC-specific synthesis, including JDM evidence, and extends the scope to adult and non-rheumatologic microvascular applications.
Doncheva-Dilova et al. (2025) [20]	Narrative overview of AI-assisted capillaroscopy and SSc/Raynaud pattern classification; no formal selection protocol reported.	Adds PRISMA-ScR selection, study-level extraction, validation indicators, reproducibility assessment, and a structured synthesis of translation gaps.
Gaurav et al. (2025) [21]	Narrative review of AI in nail disorders, including NFC; PubMed and Google Scholar searches reported, but no final included-study corpus stated.	Separates nailfold microvascular imaging from broad nail-disease AI and assesses capillaroscopy-specific tasks, validation, devices, reproducibility, and clinical translation.
Marode et al. (2026) [22]	Structured narrative technology review of AI-enabled nail diagnostics and sensing platforms; no final included-study count reported.	Focuses on AI-NFC/NVC rather than general nail diagnostics and evaluates task families, devices, validation, reproducibility, and translation barriers.
Santos et al. (2026) [23]	Structured narrative review of AI in systemic sclerosis (54 studies in qualitative synthesis).	Extends beyond SSc and provides a focused audit of AI-NFC/NVC validation, reproducibility, and clinical-translation evidence.
Present scoping review	PRISMA-ScR review of AI in NFC/NVC (3,059 records; 43 included primary studies).	Provides a cross-domain translational evidence map spanning task families, clinical scope, validation, reproducibility, devices, open resources, sample-size denominators, equity, governance, and regulatory-readiness gaps.

Note. AI = artificial intelligence; NFC = nailfold capillaroscopy; NVC = nailfold videocapillaroscopy; SSc = systemic sclerosis; JDM = juvenile dermatomyositis; PRISMA-ScR = Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews. Counts are shown only when explicitly reported by the source.

Framing AI-NFC in these terms places it squarely within medical informatics and clinical decision-making rather than image processing alone. An AI-NFC model is, in practice, a candidate clinical decision-support component, its value depends not on peak reported accuracy but on whether it can be developed, validated, implemented, and monitored safely within a real diagnostic workflow. The evidence that determines this—transparent validation design, leakage-controlled data splitting, external and multicentre testing, calibration and uncertainty reporting, open code and data, device robustness, and equity and governance safeguards—is precisely the evidence that health-information-technology evaluation and reporting frameworks were created to capture. A synthesis that maps how completely this evidence is reported, rather than which model scored highest, therefore speaks directly to the design, development, implementation, use, and evaluation of decision-support technologies for microvascular assessment. To our knowledge, no prior review has audited AI-NFC through this implementation and reporting lens.

Therefore, this scoping review aims to map the current landscape of AI applications in NFC/NVC and to evaluate the field from both technical and translational perspectives. Specifically, we addressed four questions: (1) what AI task families have been applied to NFC/NVC images; (2) which diseases, devices, and datasets dominate the evidence base; (3) how robust are validation, cross-validation, reproducibility, and external-validation practices; and (4) what methodological and translational barriers limit clinical deployment? By framing AI-NFC as a problem of validation, reproducibility, leakage control, device generalisation, and clinical

121 utility, this review seeks to move the discussion beyond model accuracy toward trustworthy implementation in
122 microvascular imaging.

123 **2 Methods**

124 **2.1 Review Design and Reporting Framework**

125 This scoping review mapped the clinical applications, technical approaches, validation practices, reproducibility,
126 and translational readiness of AI methods applied to NFC/NVC. The review was conducted and reported in
127 accordance with the PRISMA Extension for Scoping Reviews (PRISMA-ScR) [24] and the completed PRISMA-
128 ScR checklist is provided as Additional file 1. A scoping-review design was selected because the evidence base
129 spans heterogeneous clinical populations, imaging devices, AI tasks, reference standards, validation designs, and
130 performance metrics that are not suitable for a single pooled effect estimate. Generative AI (large language
131 model) tools were used only for language and organisational support during manuscript preparation and were
132 not used for any screening, eligibility, extraction, indicator-coding, or analytical decision, full details are given
133 in the Declarations (Use of generative artificial intelligence tools).

134 **2.2 Protocol and registration**

135 This review was not prospectively registered. The review nevertheless followed a predefined internal plan spec-
136 ifying the review questions, eligibility criteria, information sources, screening procedure, extraction variables,
137 evidence-classification rules, and synthesis strategy. Because no prospectively registered protocol is available,
138 this manuscript avoids treating post hoc clarifications as externally registered methods; any reconstruction of
139 screening, extraction, or coding decisions should be interpreted through the audit files accompanying the review.

140 **2.3 Eligibility Criteria**

141 Studies were eligible for the primary synthesis when they: (1) reported original research using human NFC/NVC
142 images, videos, or quantitative features derived from these images; (2) applied AI, machine learning, deep
143 learning, or a learning-based computer-vision method; (3) provided sufficient information on the dataset, model,
144 validation procedure, and outcomes for extraction; and (4) were published as a full peer-reviewed journal
145 article or a full peer-reviewed conference proceeding. Non-English full-text reports were eligible when sufficient
146 information could be extracted through translation and verification against the original source.

147 For operational consistency, “sufficient information” required extractable information in four minimum do-
148 mains: (i) the study population or image/video source, (ii) the AI or learning-based computer-vision method,
149 (iii) the validation or testing procedure, and (iv) at least one extractable outcome or performance result. Full
150 reports lacking one or more of these minimum domains were excluded from the primary evidence set only when
151 the missing information prevented study-level charting. Abstract-only records were counted separately from
152 full reports excluded for insufficiently extractable methods or results.

153 Eligible task families included capillary detection and localisation, segmentation, morphological or haemo-
154 dynamic quantification, image or patient-level classification, prediction using image-derived variables, image-
155 quality enhancement, blood-cell or flow analysis, dataset/benchmark development, and deployment-oriented
156 systems. Feature-level machine-learning studies using NFC/NVC-derived variables were retained but distin-
157 guished from end-to-end image or video modelling.

158 Conference abstracts without a complete report were excluded because their methods and results could
159 not be extracted reliably. Reviews, commentaries, editorials, protocols, book chapters, preprint-only reports,
160 non-learning deterministic algorithms, studies using an ineligible modality without optical NFC/NVC data,
161 duplicate or companion reports, and reports without extractable methods or outcomes were excluded from
162 the primary denominator. Such sources could be retained for background or citation tracing. When multiple
163 publications described the same underlying study, they were linked as a study family and counted once, using
164 the most complete eligible peer-reviewed report.

165 2.4 Information Sources and Search Strategy

166 PubMed, Ovid MEDLINE, Scopus, the Web of Science Core Collection, and IEEE Xplore were searched from
167 database inception to 28 March 2026. The core strategy combined terms for nailfold imaging (for example,
168 “nailfold”, “capillaroscopy”, “videocapillaroscopy”, “nailfold capillary”, and “microcirculation”) with terms for AI
169 and computer vision (for example, “artificial intelligence”, “machine learning”, “deep learning”, “computer vi-
170 sion”, “segmentation”, “object detection”, “classification”, “convolutional neural network”, “U-Net”, “YOLO”, and
171 “Vision Transformer”). Database-specific subject headings, field tags, truncation, and syntax were adapted
172 while retaining the same conceptual structure. Search results were exported in RIS, PubMed, or EndNote XML
173 format and imported into Covidence. Covidence was used for duplicate detection, screening, full-text/source de-
174 cision management, and PRISMA-flow tracking. Full platform-specific strategies are provided in Supplementary
175 Appendix A.

176 2.5 Study Selection

177 After import into Covidence, duplicate records were identified automatically and additional duplicates were
178 removed manually when detected during screening. Title/abstract screening was performed by AK and ZE
179 against the predefined eligibility criteria, with records judged relevant or uncertain advanced to full-text/source
180 assessment. Full-text/source assessment was also performed by AK and ZE. HJ and ZE independently verified
181 the final full-text/source decisions, study-family links, extracted study variables, and validation/reproducibility
182 indicator codes against the source documents and extraction workbook.

183 At the full-text/source stage, each report was assessed for eligibility for the primary evidence set. Exclu-
184 sion decisions were assigned according to the predefined criteria described above, including publication type,
185 modality, AI-method eligibility, duplicate or companion-report status, and availability of extractable methods
186 and outcomes. Reports that did not meet the primary-synthesis criteria could be retained only for background

context or citation tracing. Two eligible non-English full-text/source reports required translation for eligibility assessment and data charting: one Korean-language report and one Spanish-language report [25, 26]. Both were translated into English for full-text/source assessment, and the extracted information was checked against the original tables, figures, abstracts, and bibliographic metadata.

Discrepancies in eligibility, extraction, or indicator coding were resolved by discussion, re-examination of the original reports, and consensus among AK, ZE, and HJ. Cohen’s kappa was not calculated because a complete paired record of independently coded decisions was not retained for a predefined screening or extraction stage; the review therefore reports a consensus-based screening and independent-verification workflow rather than a formal duplicate-screening reliability estimate.

2.6 Data Charting

A standardised extraction form was used to collect bibliographic information, publication form, disease or application domain, study design, participant and image/video counts, clinical setting, imaging modality and device, unit of analysis, annotation and reference standard, model family, input and output type, preprocessing, training and validation design, split unit, prevention of participant overlap, external-validation category, multicenter design, comparison with experts, reported metrics, confidence intervals, inference hardware and runtime, code and data availability, contribution, and limitations. Where explicitly reported, the extraction also captured translation-relevant items including calibration, uncertainty reporting, prospective clinical-utility or workflow endpoints, demographic or skin-phototype subgroup analyses, governance or monitoring statements, and regulatory-readiness information. Missing information was coded as not reported rather than inferred. The final article version replaced a corresponding preprint when both represented the same study, and duplicate study reports were linked rather than counted independently.

For each included primary study, we also extracted two sample-size denominators when available: the participant count and the number of image-based sample units, including still images, videos, video clips, extracted video frames, or annotated frames, depending on how each study defined its dataset. Studies that did not report one of these two denominators were retained in the descriptive synthesis but excluded from the supplementary denominator comparison. Because image-based units were not harmonised across studies, the participant–image/video/frame comparison was treated as an exploratory descriptive analysis rather than as a standardised measure of statistical sample size. Spearman’s rank correlation coefficient was reported only as a supplementary cross-study summary of the reported denominators.

2.7 Evidence Classification and Synthesis

After eligibility assessment, each included primary study was classified for descriptive synthesis according to publication track, principal AI task family, and clinical or application domain. Publication track distinguished peer-reviewed journal articles from full peer-reviewed conference proceedings, allowing conference evidence to be analysed separately.

Each study was assigned to one principal AI task family based on its primary contribution: classification or prediction, detection or quantification, segmentation, dynamic or flow analysis, image quality or preprocessing, or dataset and benchmark development. Studies with multiple technical components were classified according to their primary objective. Clinical or application domains were determined from the target population, disease indication, or physiological application.

Because the included studies differed in their tasks, units of analysis, reference standards, and performance metrics, results were summarised descriptively within comparable task families. No pooled diagnostic-accuracy estimate or cross-task performance ranking was performed.

2.8 Validation and Reproducibility Indicators

From the charted variables, six validation and reproducibility indicators were derived for each included primary study: independent external clinical validation, multicenter design, direct comparison with experts, publicly accessible code, publicly accessible data, and explicit participant-level separation. These indicators were selected to describe reporting practices relevant to generalisability, leakage control, expert-reference comparison, and independent reproducibility.

Indicators were coded using strict affirmative criteria. A study was coded as “Yes” only when the relevant information was explicitly reported in the article or an accompanying public resource. Availability only on request was not counted as publicly accessible code or data. Unclear, incomplete, or absent reporting was coded as not reported/unclear rather than inferred. Participant-level separation required explicit evidence that images, videos, fingers, fields, or frames from the same participant were not split across training, validation, or test partitions.

The indicators were analysed separately and were not combined into a composite quality score, maturity tier, or risk-of-bias rating, because their relevance differs across AI task families and study designs. A study-level indicator matrix is provided in Supplementary Table S5. For these indicators, “Yes” means explicitly affirmative reporting; “NR/unclear” means that the report did not provide enough information to determine the indicator; and “No” was reserved for cases where the study design or text made the absence of the indicator clear under the review definition. For independent external clinical validation, internal random splits, internal cross-validation, internal held-out images, and device holdout within the same development programme were not coded as independent external clinical validation. For public code or public data, access on request, private institutional storage, or unavailable links were not coded as public.

3 Results

3.1 Study Selection

The database search and Covidence workflow identified 3,059 records from five sources. After de-duplication, 1,305 records underwent title/abstract screening, and 103 reports progressed to full-text/source assessment

(Figure 1). Following full-text/source review, 43 primary studies met the criteria for the scoping synthesis, comprising 37 peer-reviewed journal articles and six full peer-reviewed conference proceedings. Sixty reports were excluded from the primary evidence set using mutually exclusive primary exclusion reasons (Figure 1). Eighteen non-primary reports were retained only for background context or citation tracing and were not included in the primary denominator.

3.2 Characteristics and Evolution of the Evidence Base

The 43 included primary studies were published between 2014 and 2026 and comprised 37 journal articles and six full conference proceedings. Figure 1 summarises study identification and selection and provides an overview of the final evidence distribution by principal AI task family.

Publication activity increased after 2022, with studies spanning all six AI task families. Classification/prediction was the largest group, whereas image-quality/preprocessing and dataset/benchmark studies were least frequent. Some systems performed more than one technical function; for descriptive synthesis, each study was assigned one principal task according to its main reported contribution. Figure 2 shows how task families and model architectures evolved over time, including classical machine-learning or handcrafted pipelines, CNN-based models, U-Net-family segmentation networks, YOLO/object-detection models, transformer-based approaches, and hybrid or ensemble systems.

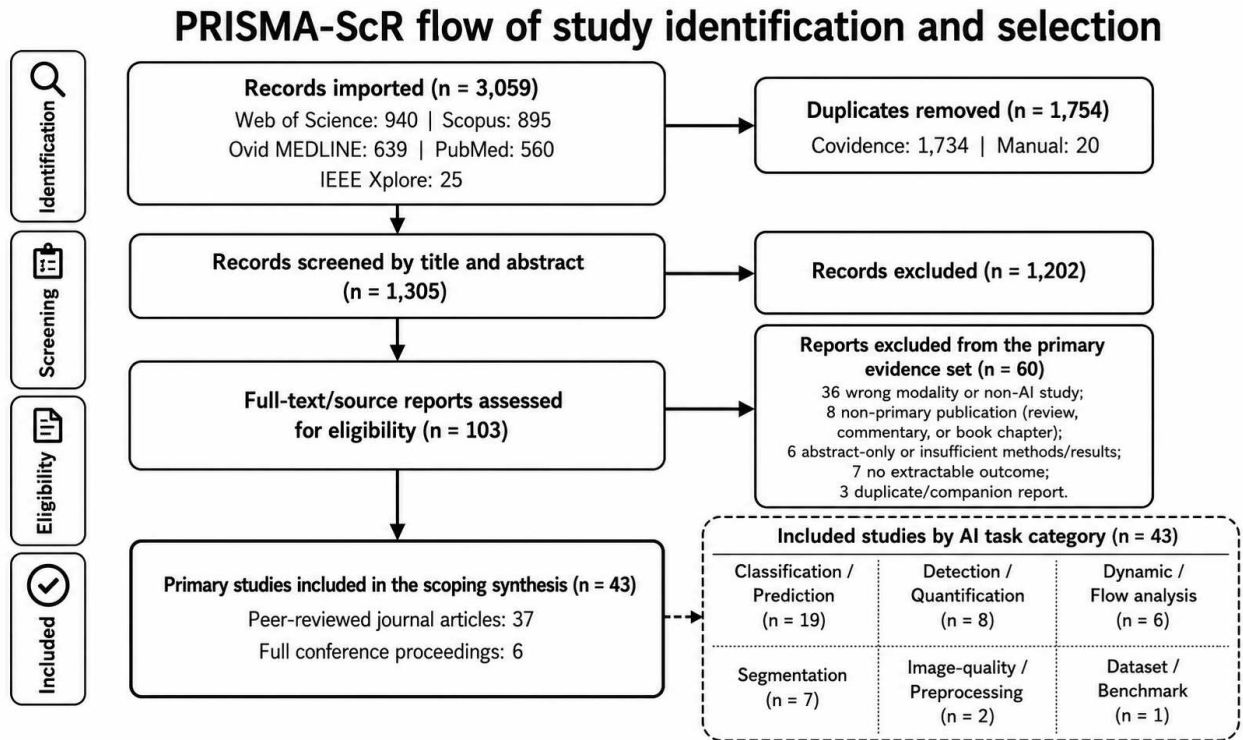


Figure 1: Study identification, selection, and final AI task-family distribution. The diagram summarises database retrieval, duplicate removal, title/abstract screening, and full-text/source eligibility assessment. Counts were derived from the Covidence import and screening history. The final evidence set comprised 43 primary studies (37 journal articles and six full conference proceedings). The task-category panel assigns each included study to one principal AI task family based on its main reported contribution; studies with multiple technical functions were counted once. When multiple reports described the same study family, the most complete eligible report was retained.

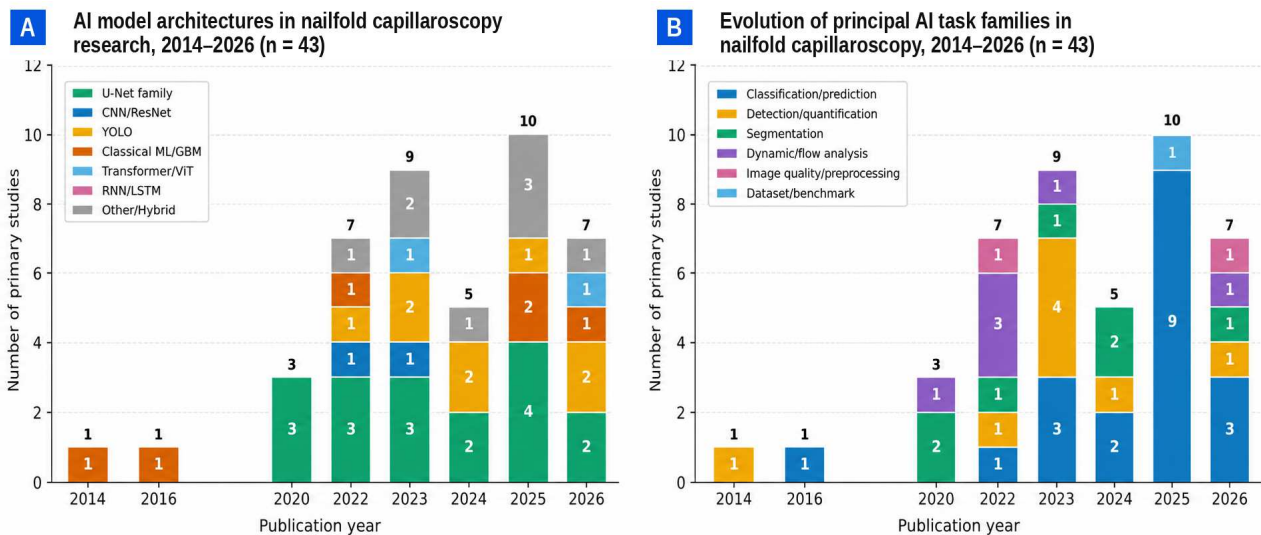


Figure 2: Evolution of AI model architectures and task families in AI–NFC research, 2014–2026 (n = 43). (A) **Principal AI model architecture families over time.** Each study was assigned one principal architecture category, illustrating the transition from classical ML and handcrafted pipelines to deep learning, object-detection, transformer-based, ensemble, and deployment-oriented approaches. (B) **Principal AI task families over time.** Each study was assigned one principal task family for descriptive mapping, showing the expansion from early detection/quantification and classification work to segmentation, dynamic/flow analysis, preprocessing, and benchmark development. Colours are meaningful in both panels and denote the corresponding task or architecture categories shown in the figure legends.

3.3 Clinical and Application Domains

The included studies covered seven clinical or application-domain categories (Figure 3). Fourteen studies primarily addressed SSc/RP-spectrum applications (33%). Eighteen studies focused on general or physiological microcirculation, including capillary morphology, image quality, flow, and cellular monitoring (42%). Five studies addressed diabetes or metabolic applications (12%), two addressed JDM (5%), two addressed other autoimmune or inflammatory diseases (5%), one addressed retinopathy of prematurity (ROP; 2%), and one addressed wild-type transthyretin (ATTRwt) amyloidosis (2%).

Figure 3A shows the distribution of clinical/application domains across principal AI task families. Figure 3B shows the year in which each disease or application domain appeared in the included primary evidence set.

3.4 Validation, Reproducibility, and Reporting

Figure 4B summarises the six validation and reproducibility indicators across the 43 included primary studies, and Supplementary Table S6 provides the corresponding task-family stratification. Under the strict definition used in this review, one study provided independent external clinical validation: the multicentre evaluation by Tello et al. [27]. Bharathi et al. reported independent held-out subject groups and a low-cost device cohort, but these cohorts were not classified as an independently sourced external clinical validation dataset [11].

The stratified summary showed that affirmative reporting was concentrated in classification/prediction and detection/quantification studies, whereas segmentation, dynamic/flow-analysis, and image-quality/preprocessing studies rarely reported validation and reproducibility indicators beyond internal technical evaluation. Explicit participant-level separation was reported in five studies, comprising four journal

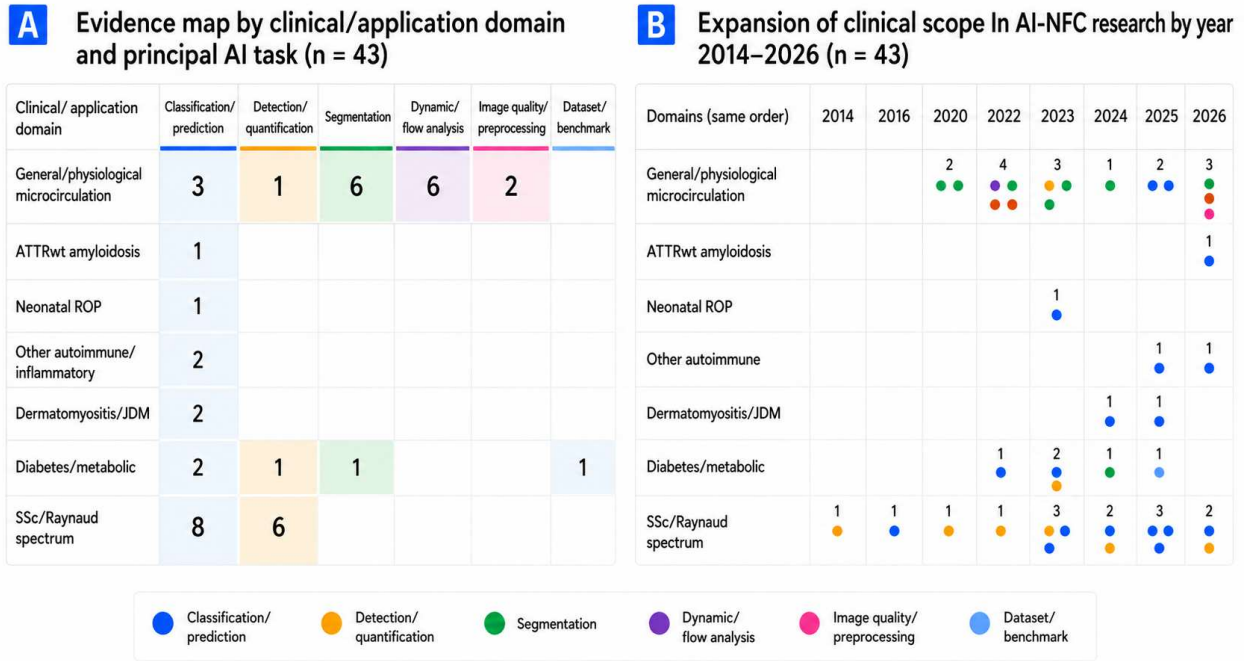


Figure 3: Clinical scope and temporal expansion of AI–NFC research, 2014–2026 (n = 43; colours are meaningful). (A) Evidence map by clinical/application domain and principal AI task. Cell values show the number of included primary studies, and colour intensity represents the study count in each cell. (B) Timeline of disease/application-domain expansion. Each dot represents one primary study; dot colour denotes the principal AI task family. Arrows mark the year each disease/application domain first appeared in the primary evidence set. SSc/RP studies have been present throughout, while most non-SSc domains emerged after 2022.

articles and one conference proceeding. Publicly accessible code and publicly accessible data were reported in four and two studies, respectively. Taken together, 28 of the 43 included studies (65.1%) met none of the six indicators, underscoring how concentrated affirmative reporting was among a small subset of studies. The corresponding study-level matrix appears in Supplementary Table S5 (Methods) and Supplementary Figure S2.

3.5 Performance Reporting and Cross-Study Comparability

Reported performance metrics varied by AI task family. Classification studies commonly reported accuracy, sensitivity, specificity, F1 score, AUROC, or AUPRC; segmentation studies reported Dice, intersection-over-union, or pixel accuracy; object-detection studies reported average precision or mean average precision; quantitative and flow-analysis studies reported correlation, error, agreement, or velocity estimates; and deployment-oriented studies additionally reported runtime or hardware characteristics. Detailed study characteristics and reported performance are provided in Supplementary Tables S3 and S4.

Participant-level cohort sizes varied across task families (Figure 4A). Among the 31 studies reporting a participant-level denominator, the overall median cohort size was 30 participants. Participant counts were not reported in 12 studies.

In the supplementary denominator comparison, 24 studies reported both participant counts and an image/video/frame denominator. This 24-study subset is smaller than, and not identical to, the 31 studies that reported a participant count, so its median participant count (70) differs from the 30-participant median reported above for the full 31-study set; the two medians are computed over different subsets and are not in conflict.

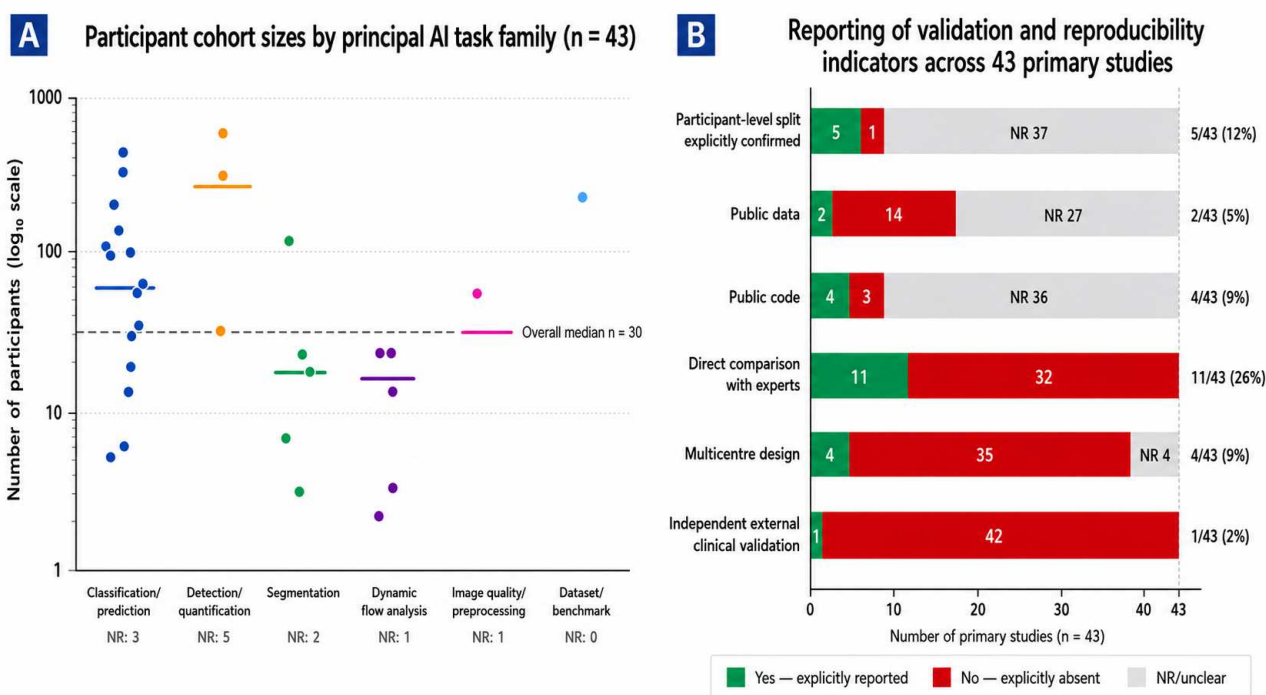


Figure 4: Cohort size, validation, reproducibility, and reporting across AI-NFC primary studies (n = 43). (A) **Participant cohort sizes by principal AI task family.** Each point represents one primary study reporting a participant-level denominator. The vertical axis is logarithmic because cohort sizes varied substantially. Coloured horizontal lines indicate within-task medians. “NR” below each task family denotes studies for which a participant-level denominator was not reported. The dashed horizontal line indicates the overall median cohort size of 30 participants among the 31 studies that reported a participant-level denominator. (B) **Reporting of validation and reproducibility indicators across 43 primary studies.** Bars show studies reporting explicit presence, explicit absence, or not-reported/unclear status for each indicator. Public code and public data required immediately accessible resources; availability only on request was not counted as public availability.

Because these image-based units were not harmonised across studies, the correlation result is reported only as a descriptive cross-study summary (Spearman $\rho = 0.35$, $p = 0.095$; Supplementary Figure S1; Supplementary Tables S7–S8) and was not used as a pooled measure of effective sample size.

4 Discussion

4.1 The Evidence Base Has Expanded Substantially

This scoping review addressed four questions concerning the task families, clinical domains, validation practices, and translational barriers in AI-assisted NFC/NVC research. The main finding is that AI-NFC has expanded substantially in both technical scope and clinical ambition, but this expansion has not yet been matched by equally strong evidence for clinical generalisability, reproducibility, or implementation readiness.

This expansion should be interpreted as widening scope rather than as evidence of field maturity. Much of the growth occurred in recent years, and many included reports were proof-of-concept, engineering, segmentation, preprocessing, or benchmark studies rather than deployment-ready clinical evaluations. The validation and reproducibility indicators used here therefore describe the availability of translation-relevant evidence; they are not intended as a pass–fail judgement on every early-stage study.

The included studies show that AI-NFC is no longer limited to automated capillary counting or SSc-pattern

classification. The evidence base now spans classification and prediction, capillary detection and quantification, segmentation, dynamic flow and cellular analysis, image-quality enhancement, dataset and benchmark development, and deployment-oriented systems. This breadth indicates that AI-NFC is developing into a broader microvascular-computing field, with models designed to support quantitative biomarker extraction, diagnostic classification, image restoration, object-detection benchmarking, low-cost imaging, and real-time analysis.

The same expansion is visible at the level of model architecture. Early studies relied mainly on classical image-processing, handcrafted features, random forests, support-vector machines, and other traditional machine-learning pipelines. More recent studies increasingly used CNNs, U-Net-family segmentation models, YOLO-family object detectors, Vision Transformers, gradient-boosted models, attention-based networks, and hybrid or ensemble approaches. This architectural diversity is useful because it brings methods from computer vision, biomedical image analysis, and clinical decision-support research into the same field. At the same time, it also explains why a pooled diagnostic-accuracy meta-analysis was not appropriate, the included studies differed substantially in task definition, unit of analysis, reference standard, validation design, and reported metrics.

The clinical and application scope has also broadened beyond the traditional SSc/RP focus. SSc/RP remains the largest disease-specific domain, reflecting its established capillaroscopic taxonomy, expert-label tradition, and clinical dependence on NVC. However, many studies now address general microcirculation or non-SSc applications, including diabetes, JDM, rheumatoid arthritis (RA)/SLE, psoriasis/psoriatic arthritis (PsA), neonatal ROP, ATTRwt amyloidosis, image-quality assessment, flow analysis, and dynamic cellular monitoring. This broader scope is important because it suggests that AI-NFC is increasingly being considered as a general microvascular-imaging platform and a tool for scleroderma-pattern recognition.

This diversification should be interpreted with caution. Many non-SSc applications remain exploratory and lack disease-specific validation, transferable reference standards, and clear links to clinical decision-making. Compared with the prior direct AI-NFC review and adjacent reviews of microcirculation, nail diagnostics, rheumatology AI, and systemic sclerosis AI, the present review therefore adds a wider translational evidence map and updated list of models [17, 18, 23]. The key message from this first part of the Discussion is that AI-NFC has become technically and clinically broader, but the evidence base remains methodologically fragmented.

4.2 Technical Diversity Does Not Equal Clinical Generalisability

The expansion of model architectures and clinical targets is encouraging, but technical diversity should not be equated with clinical readiness. Which datasets and devices dominate the field, and how robust are the validation practices? Several studies reported strong internal performance, but many used small cohorts, selected images, image-level splits, or incompletely described split units. In repeated-measures imaging settings such as NFC/NVC, these design choices can make apparent performance difficult to interpret.

The distinction between clinical cohort size and technical image volume is particularly important. NFC/NVC datasets may contain multiple images, fingers, fields, videos, frames, or visits from the same participant. Large image or frame counts can therefore arise from repeated imaging, video-frame extraction, or dense annotation

rather than from a larger or more diverse clinical cohort. As a result, image-unit counts alone can overstate the effective sample size available for clinical generalisation. The supplementary denominator comparison is exploratory because it combines non-equivalent image-based units; its role is therefore to illustrate reporting heterogeneity and the need to report participant and image/video/frame denominators separately, not to provide a standardised estimate of clinical sample size.

Participant-level separation should be treated as a minimum requirement for model evaluation in this field. When correlated observations from the same participant are allowed to appear in both training and test sets, performance estimates may be inflated by participant overlap rather than true generalisable signal [16]. For stronger deployment claims, participant-level grouping should be extended to centre-, device-, temporal-, and geographic holdout designs. This is especially relevant for NFC/NVC because acquisition quality, magnification, illumination, field of view, operator experience, disease spectrum, and device type can all influence model performance.

Negative or modest findings are also informative and should remain visible in the evidence base. For example, the negative pretrained-ViT comparison illustrates that a powerful generic architecture does not automatically transfer to capillaroscopy without domain-specific training, task alignment, and appropriate clinical validation [28]. This finding is useful because it shifts attention away from model novelty alone and toward data quality, clinical framing, and validation design.

4.3 External Validation Is the Principal Evidence Gap

Independent external clinical validation remains the main evidence gap. Under the strict coding criteria, only one study met the definition of independent external clinical validation. Bharathi et al. provide valuable independent held-out and cross-device testing, while Tello et al. provide true external multicentre clinical validation [11, 27]. Both studies represent important progress, but they also show how carefully validation categories need to be distinguished.

The Tello et al. [27] study was coded as independent external clinical validation because the validation images were separate from the model-development data and were evaluated in a multicentre clinical setting against expert consensus. This coding refers to independence of the validation dataset/source under the review definition, it should not be interpreted as a fully independent third-party validation performed by an unrelated research team. The earlier Capillary.io report [29] from the same research programme was therefore treated as development-stage evidence, whereas the later Tello et al. report was treated as external clinical validation.

Multicentre data collection, internal holdout testing, cross-device evaluation, and independent external clinical validation are related but not interchangeable. Multicentre development studies such as CAPI-Detect, SCLEROCAP, and Capillary.io are important because they broaden the source population and clinical workflow context [29–31]. However, if the model-development process and evaluation data are not fully separated, multicentre development should not be interpreted as external validation. Future studies should therefore define the intended use before model development, lock the model before external testing, preserve participant-level

grouping, and evaluate performance across independent centres, devices, acquisition protocols, demographic groups, and disease subgroups.

External validation should also include clinically meaningful reporting beyond discrimination metrics. Confidence intervals, calibration, subgroup performance, failure modes, and uncertainty estimates are needed to determine whether a model is reliable enough for clinical use. Prospective studies should additionally measure workflow outcomes such as interpretation time, clinician decision change, diagnostic error, referral efficiency, usability, cost, and patient-relevant outcomes. Without these endpoints, high retrospective accuracy remains insufficient evidence for clinical implementation.

4.4 Reproducibility and Benchmarking Remain Limited

Public code and public data were uncommon under the strict coding criteria, and availability only on request was not treated as immediate public availability. Data or software shared upon request can still support collaboration, but it does not provide the same level of independent reproducibility as open code, model weights, preprocessing scripts, annotation rules, and reusable benchmark datasets.

Public datasets and benchmark studies are therefore valuable developments in the field, especially when they include participant-level grouping, clear licensing, device metadata, annotation documentation, and predefined train/validation/test partitions [32, 33]. However, benchmarks must be designed carefully. A benchmark that contains many images but limited participant diversity, unclear split units, or incomplete metadata may encourage technical optimisation without improving clinical generalisability. For AI-NFC, benchmark design should distinguish between still images, videos, frames, fingers, fields, visits, and participants, because these units are not statistically equivalent.

Annotation and reference-standard reporting also need improvement. Direct comparison with experts was reported in a minority of studies, and annotation protocols, consensus rules, disagreement handling, and inter-rater reliability were often incompletely described. This matters because the ground truth in capillaroscopy is not always self-evident: capillary boundaries, loop shape, haemorrhage, tortuosity, ramification, density, and scleroderma-pattern labels can vary across raters and acquisition conditions. Stronger studies should report the number and expertise of annotators, the annotation protocol, consensus or adjudication procedures, disagreement rates, and inter-rater agreement where applicable.

4.5 Priority Reporting and Translation Gaps

The evidence map identifies underdeveloped reporting and validation infrastructure as the main barrier to clinical deployment, model innovation is not the limiting factor. Future AI-NFC studies should move beyond retrospective performance reporting and provide stronger evidence on external generalisability, split-unit control, open reproducibility, benchmark governance, device robustness, clinical utility, equity, and lifecycle monitoring. Future reporting should also be aligned with established AI-specific reporting frameworks. For AI-NFC/NVC studies that develop or validate diagnostic or prognostic prediction models, TRIPOD+AI provides relevant

426 guidance for transparent reporting of model development, validation, predictors, outcomes, calibration, and
427 clinical-use context [34]. For imaging-based AI studies, CLAIM provides additional domain-specific guidance
428 on image acquisition, dataset construction, preprocessing, annotation, model training, validation, performance
429 reporting, and reproducibility [35]. For early clinical evaluation or workflow integration, DECIDE-AI is relevant,
430 whereas prospective randomized evaluations of AI-assisted clinical workflows should follow CONSORT-AI and,
431 at the protocol stage, SPIRIT-AI [36, 37]. Each framework addresses a distinct stage of development or evalu-
432 ation, and together they clarify the reporting expectations needed for trustworthy translation of AI-NFC/NVC
433 systems.

434 Table 2 translates the main gaps identified in this review into practical reporting priorities. The table is
435 intended as a reporting-oriented synthesis, not as a scoring framework or risk-of-bias tool. Its purpose is to make
436 the translational message explicit: AI-NFC studies should report the clinical denominator, image/video/frame
437 denominator, split unit, device and acquisition details, reference-standard process, validation design, code/data
438 availability, calibration, subgroup performance, and intended-use context.

439 Device portability and acquisition generalisation deserve particular attention. Low-cost, smartphone,
440 portable, and edge-deployment systems are promising because they may extend capillaroscopy beyond spe-
441 cialist centres. However, portability also introduces additional sources of variation, including microscope type,
442 magnification, illumination, field of view, image quality, operator experience, and preprocessing. Claims of
443 portable or low-cost deployment should therefore be supported by device-stratified or device-external validation
444 rather than by internal testing alone.

445 Clinical implementation also requires equity and governance considerations. Optical imaging performance
446 may vary with skin phototype, age, disease severity, comorbidity, acquisition setting, and device configuration.
447 No included study provided skin-phototype-stratified performance, formal lifecycle monitoring, or regulatory-
448 readiness information in a form suitable for tabulation. Where demographic, subgroup, workflow, or governance-
449 relevant information was mentioned, reporting was usually insufficient for a systematic subgroup or governance
450 synthesis. These items should therefore be interpreted as priority recommendations arising from the evidence
451 map rather than as a formal risk-of-bias or regulatory-readiness assessment. Future studies should predefine
452 intended use, risk controls, failure modes, data governance, monitoring plans, and post-deployment evaluation
453 pathways.

454 4.6 Conference Evidence Requires a Separate Track

455 The inclusion of full conference proceedings was methodologically important because AI-NFC is partly driven
456 by engineering and computer-vision research. Conference proceedings captured foundational work in automated
457 vessel detection, segmentation, dataset development, and flow analysis that may be missed if the evidence base
458 is restricted to clinical journals. Including these reports therefore improves technical coverage and reduces the
459 risk of overlooking early innovation.

460 However, conference proceedings often provide less complete clinical detail than journal articles. They may

Table 2: Priority methodological and translational gaps in AI-NFC studies and recommended reporting directions.

Gap domain	Review finding	Why this limits interpretation	Recommended report-ing/action
External validation and generalisability	Strict independent external clinical validation was identified in 1/43 studies; multicentre design was reported in 4/43.	Internal or multicentre development performance does not establish generalisability across patients, centres, devices, operators, acquisition protocols, or disease spectrum.	Predefine intended use; lock the model before testing; evaluate on independent external cohorts; report centre-, device-, operator-, and disease-subgroup performance.
Split unit and leakage control	Participant-level separation was explicitly confirmed in 5/43 studies; most studies were NR/unclear, and image-, field-, finger-, or frame-level splits were often insufficiently described.	Repeated images, fingers, fields, visits, or frames from the same participant can leak across train/test partitions and inflate performance estimates.	State the split unit explicitly; preserve participant-level grouping at minimum; where possible, use centre-, device-, temporal-, or geographic holdout sets for deployment claims.
Sample-size denominators and dataset density	Only 24/43 studies reported both participant counts and an image/video/frame denominator; the image-based units were not harmonised across studies.	Large image datasets may reflect repeated imaging, video clips, frame extraction, or annotation strategy rather than large clinical cohorts; image-unit counts alone can overstate the effective sample size.	Report participant count, image/video/frame count, number of fingers/fields/visits, and image units per participant; distinguish clinical sample size from technical image volume.
Reproducibility and open science	Public code was available in 4/43 studies and public data in 2/43 under strict affirmative coding.	Independent verification, model reuse, benchmark comparison, and error analysis remain limited when code, weights, preprocessing scripts, and data are unavailable.	Release inference and training code, model weights, preprocessing scripts, annotation rules, data dictionaries, and licence/availability statements; distinguish public access from access on request.
Annotation and reference standard	Direct comparison with experts was reported in 11/43 studies, but annotation protocols, consensus rules, disagreement handling, and inter-rater reliability were inconsistently described.	Label variability affects model training, ground-truth reliability, and interpretation of apparent model performance.	Report number and expertise of raters, annotation protocol, consensus rule, adjudication process, disagreement rate, and inter-rater reliability where applicable.
Benchmarking and metric comparability	Tasks and metrics varied substantially across classification, detection, segmentation, flow analysis, image restoration, and deployment studies; no pooled ranking was appropriate.	Accuracy, Dice, mAP, correlation, runtime, and clinical diagnostic metrics are not directly comparable across task families or units of analysis.	Develop task-specific benchmark datasets with fixed participant-level splits, common metrics, device metadata, and reporting templates; avoid cross-task ranking from incompatible metrics.
Device portability and acquisition generalisation	Low-cost, portable, smartphone, and edge-deployment systems were reported, but device-stratified and device-external validation remained limited.	Clinical adoption requires robustness to microscope type, magnification, illumination, field of view, image quality, operator, and acquisition setting.	Report device model, magnification, field of view, lighting, preprocessing, image-quality exclusions, and device-stratified performance; test across vendor-diverse and low-cost devices.
Clinical scope beyond SS	SSc/Raynaud remained the largest disease-specific domain, although 29/43 studies addressed broader physiology or non-SSc applications.	Broader clinical relevance is promising, but many non-SSc applications remain exploratory and lack disease-specific validation or transferable reference standards.	Build disease-specific cohorts for diabetes, JDM/dermatomyositis, RA/SLE, psoriasis/PsA, neonatal ROP, amyloidosis, and general microcirculation; report clinically meaningful outcomes and reference standards.
Clinical utility and implementation	Prospective human-AI workflow evaluation, diagnostic decision-change analysis, cost-effectiveness, calibration, and patient-impact endpoints were not systematically reported.	High retrospective accuracy is insufficient to show clinical usefulness, safety, or workflow value.	Measure calibration, uncertainty, interpretation time, clinician decision change, referral efficiency, diagnostic error, cost, usability, and patient-relevant outcomes in prospective studies.
Equity, governance, and regulatory readiness	Phototype-stratified performance, demographic subgroup analysis, lifecycle monitoring, and regulatory-readiness plans were not consistently extractable from the current evidence base.	Optical imaging and AI performance may vary with skin tone, age, disease severity, device setting, and clinical environment; governance gaps limit safe deployment.	Report demographic and skin-phototype distributions where available; evaluate subgroup performance; predefine intended use, risk controls, failure modes, data governance, monitoring, and post-deployment evaluation.

Note. AI-NFC = artificial intelligence in nailfold capillaroscopy; NFC = nailfold capillaroscopy; NVC = nailfold videocapillaroscopy; SSc = systemic sclerosis; JDM = juvenile dermatomyositis; RA = rheumatoid arthritis; SLE = systemic lupus erythematosus; PsA = psoriatic arthritis; ROP = retinopathy of prematurity; NR = not reported or unclear. The indicators summarise descriptive evidence-mapping findings and should not be interpreted as a validated risk-of-bias or study-quality score.

omit participant-level denominators, acquisition metadata, annotation protocols, split-unit information, confidence intervals, or external-validation details. They may also precede a later journal version of the same underlying study. Treating full conference proceedings as a separate primary-evidence track is therefore appropriate, it preserves technical coverage while preventing conference evidence from silently inflating the journal-study denominator.

Study-family linking remains essential for future updates. Preprints, conference papers, abstracts, software papers, and final journal articles may describe overlapping datasets or model-development pipelines. Counting these records as independent studies can distort the evidence base. Future reviews should therefore preserve the title-level exclusion log, full-text/source decisions, and study-family linkage file alongside the extraction dataset.

4.7 Strengths and Limitations

This review has several strengths. It used a broad multidisciplinary search across biomedical and engineering databases, Covidence-based de-duplication and screening, assessment of all full-text/source reports that reached eligibility review, explicit study-family linking, and separation of journal articles from full conference proceedings. It also extended beyond algorithm-performance reporting by mapping task families, clinical scope, validation design, split-unit reporting, expert comparison, code availability, data availability, sample-size denominators, device portability, and translational readiness.

The review also avoided pooling incompatible metrics across heterogeneous task families. This is important because accuracy, AUROC, Dice, intersection-over-union, mean average precision, correlation, velocity estimates, runtime, and clinical diagnostic outcomes do not measure the same construct. Instead, the synthesis reports results descriptively and uses the validation and reproducibility indicators as evidence-mapping variables rather than as a validated quality score, maturity tier, or risk-of-bias rating.

Several limitations should be considered. The review was not prospectively registered, although it followed a predefined plan and PRISMA-ScR reporting principles. The boundary between direct image AI, feature-level machine learning, and contextual quantitative automation required judgement. Full conference proceedings were retained to capture engineering innovation, but their clinical reporting was often less complete than journal articles. In addition, no single risk-of-bias or critical-appraisal tool was applied across classification, segmentation, detection, flow analysis, image restoration, and dataset/benchmark studies. Accordingly, the review should be interpreted as a scoping evidence map of applications and reporting practices, not as a formal task-stratified risk-of-bias assessment. In addition, although this review extends beyond SSc/RP, the disease-specific clinical evidence remains concentrated around SSc/RP. Evidence for other rheumatic diseases was comparatively sparse and heterogeneous, limiting disease-specific conclusions across the broader rheumatology spectrum.

A further limitation is the possibility of positive-result and selective-reporting bias. Many included studies were retrospective, small, single-centre, not prospectively registered, and reported heterogeneous outcomes. Studies with poor performance, failed model transfer, or unsuccessful external validation may be less likely to be published. Reported performance values may therefore overestimate generalisable performance, independently

of the specific leakage risks created by image-, finger-, field-, video-, or frame-level splitting.

Finally, the conclusions are limited by the reporting quality of the primary studies. Many studies incompletely reported participant overlap, split units, calibration, annotation reliability, demographic characteristics, device metadata, or external-validation procedures. Therefore, absence of reporting should not always be interpreted as absence of the practice. Nevertheless, incomplete reporting itself is an important translational finding, because clinical users, reviewers, and regulators cannot assess generalisability or reproducibility without these details.

5 Conclusions

AI applications in NFC/NVC have expanded from early automated vessel analysis into a broader microvascular-imaging field that includes classification, detection and quantification, segmentation, flow and cellular analysis, image-quality enhancement, public datasets, and portable or deployment-oriented systems. The evidence base has also moved beyond the traditional SSc/RP focus, with emerging applications in diabetes, inflammatory and autoimmune diseases, neonatal retinopathy, amyloidosis, and general microcirculation.

The central translational conclusion is cautionary. Technical diversity and strong internal performance have not yet been matched by robust evidence of clinical generalisability. Under the review’s strict affirmative coding, independent external clinical validation was confirmed in only one of the 43 included studies, and participant-level separation, public code, and public data likewise remained uncommon (5/43, 4/43, and 2/43, respectively). Calibration, uncertainty reporting, and prospective clinical-utility endpoints were not consistently extractable and should be treated as priority reporting needs rather than as formally appraised evidence-quality domains. Multicentre model development, cross-device testing, and portable-device implementation are encouraging, but they should not be interpreted as evidence of deployment readiness unless models are locked and evaluated on independent clinical cohorts.

Future AI-NFC/NVC studies should prioritise locked external testing across centres, devices, acquisition protocols, and patient subgroups; explicit reporting of the split unit and participant-level separation; open or reusable benchmarks with clear annotation standards; clinically interpretable uncertainty and calibration; and prospective evaluation of whether AI-assisted capillaroscopy improves diagnosis, longitudinal disease monitoring, referral efficiency, treatment decisions, and workflow outcomes in routine rheumatology practice.

List of abbreviations

AI: artificial intelligence; AI-NFC: artificial intelligence in nailfold capillaroscopy; AUPRC: area under the precision–recall curve; AUROC: area under the receiver operating characteristic curve; CLAIM: Checklist for Artificial Intelligence in Medical Imaging; CNN: convolutional neural network; CONSORT-AI: Consolidated Standards of Reporting Trials–Artificial Intelligence; DECIDE-AI: Developmental and Exploratory Clinical Investigations of DEcision support systems driven by Artificial Intelligence; JDM: juvenile dermatomyositis;

mAP: mean average precision; ML: machine learning; NFC: nailfold capillaroscopy; NR: not reported or unclear; NVC: nailfold videocapillaroscopy; PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews; PsA: psoriatic arthritis; RA: rheumatoid arthritis; ROP: retinopathy of prematurity; RP: Raynaud's phenomenon; SLE: systemic lupus erythematosus; SPIRIT-AI: Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence; SSc: systemic sclerosis; TRIPOD+AI: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis plus Artificial Intelligence; U-Net: U-shaped convolutional neural network; ViT: Vision Transformer; YOLO: You Only Look Once.

Declarations

Ethics approval and consent to participate

Not applicable. This scoping review is based entirely on published literature and did not involve new human participants, human tissue, animal data, biological specimens, or identifiable individual-level patient data.

Consent for publication

Not applicable. This manuscript does not contain any individual person's data in any form.

Availability of data and materials

All data generated or analysed during this study are included in this published article and its supplementary information files. The full search strategies, PRISMA-ScR checklist, study-level extraction matrix, validation and reproducibility indicator matrix, supplementary tables, and supplementary denominator analyses are provided as Additional files 1–2. The complete study-level extraction workbook, the six-indicator coding matrix, the title-level exclusion log, the study-family linkage file, and the analysis scripts underlying the figures are additionally deposited in a retrospective, open-ended OSF registration documenting the review protocol as it was actually followed (consistent with the Methods, Section 2.2): Open Science Framework (OSF), <https://doi.org/10.17605/OSF.IO/2YEMD>, with the associated working project available at <https://osf.io/xjzsf2>. No individual participant-level data were generated or analysed in this review.

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no external funding.

557 **Authors' contributions**

558 ZE, AK and SKS conceptualised the review. ZE, AK, HJ and SKS designed the methodology. ZE, AK and HJ
559 performed the investigation and data charting. ZE, AK and HJ curated the data. ZE and AK performed the
560 formal analysis and visualisation. ZE and AK drafted the manuscript. All authors contributed to review and
561 editing. SKS, AF and AK supervised the work. All authors read and approved the final manuscript.

562 **Acknowledgements**

563 Not applicable.

564 **Authors' information**

565 Not applicable.

566 **Patient and public involvement**

567 No patient or public involvement was undertaken for this literature-based scoping review.

568 **Use of generative artificial intelligence tools**

569 Generative artificial intelligence tools were used only for language and organisational support during manuscript
570 preparation. No AI tool was used to screen records, determine eligibility, assign exclusion reasons, extract study
571 data, code validation or reproducibility indicators, generate analytical results, or serve as a cited source. All
572 content, citations, analytical conclusions, and final manuscript text were reviewed, verified, and approved by
573 the authors.

574 **Additional files**

575 **Additional file 1. PRISMA-ScR checklist.**

576 File format: PDF. Completed PRISMA Extension for Scoping Reviews checklist indicating where each reporting
577 item is addressed in the manuscript and supplementary material.

578 **Additional file 2. Supplementary material.**

579 File format: PDF. Supplementary material for “Artificial Intelligence in Nailfold Capillaroscopy: A Scoping Re-
580 view of Validation, Reproducibility, and Clinical Translation.” This file contains Supplementary Appendix A,
581 search strategy, information sources, and Covidence workflow; Supplementary Appendix B, evidence rules; Sup-
582plementary Appendix C, PRISMA and evidence-set reconciliation; Supplementary Tables 1–8; Supplementary
583 Figures 1–2; and supplementary references.

References

- [1] Mahnaz Etehad Tavakol, Alimohammad Fatemi, Abdolamir Karbalaie, Zahra Emrani, and Björn-Erik Erlandsson. Nailfold capillaroscopy in rheumatic diseases: Which parameters should be evaluated? *BioMed Research International*, 2015:974530, 2015. doi:10.1155/2015/974530.
- [2] Zahra Emrani, Abdolamir Karbalaie, Alimohammad Fatemi, Mahnaz Etehadtavakol, and Björn-Erik Erlandsson. Capillary density: An important parameter in nailfold capillaroscopy. *Microvascular Research*, 109:7–18, 2017. doi:10.1016/j.mvr.2016.09.001.
- [3] M. Cutolo, A. Sulli, C. Pizzorni, and S. Accardo. Nailfold videocapillaroscopy assessment of microvascular damage in systemic sclerosis. *The Journal of Rheumatology*, 27(1):155–160, 2000.
- [4] Maurizio Cutolo, Ariane L. Herrick, Oliver Distler, et al. Nailfold videocapillaroscopic features and other clinical risk factors for digital ulcers in systemic sclerosis: a multicenter, prospective cohort study. *Arthritis & Rheumatology*, 68(10):2527–2539, 2016. doi:10.1002/art.39730.
- [5] F. van den Hoogen, D. Khanna, J. Fransen, et al. 2013 classification criteria for systemic sclerosis: An american college of rheumatology/european league against rheumatism collaborative initiative. *Arthritis and Rheumatism*, 65(11):2737–2747, 2013. doi:10.1002/art.38098.
- [6] Alimohammad Fatemi, Björn-Erik Erlandsson, Zahra Emrani, Mahnaz Etehadtavakol, Abbas Smiley, and Abdolamir Karbalaie. Nailfold microvascular changes in patients with systemic lupus erythematosus and their associative factors. *Microvascular Research*, 126:103910, 2019. doi:10.1016/j.mvr.2019.103910.
- [7] Reema Shah, Jeremy Petch, Walter Nelson, Karsten Roth, Michael D. Noseworthy, Marzyeh Ghassemi, and Hertz C. Gerstein. Nailfold capillaroscopy and deep learning in diabetes. *Journal of Diabetes*, 15(2):145–151, 2023. doi:10.1111/1753-0407.13354.
- [8] Peyman Hosseinzadeh Kassani, Louis Ehwerhemuepha, Chloe Martin-King, Ryan Kassab, Ellie Gibbs, Gabrielle Morgan, and Lauren M. Pachman. Artificial intelligence for nailfold capillaroscopy analyses – a proof of concept application in juvenile dermatomyositis. *Pediatric Research*, 95:981–987, 2024. doi:10.1038/s41390-023-02894-7.
- [9] Abdolamir Karbalaie, Farhad Abtahi, Alimohammad Fatemi, Mahnaz Etehadtavakol, Zahra Emrani, and Björn-Erik Erlandsson. Elliptical broken line method for calculating capillary density in nailfold capillaroscopy: Proposal and evaluation. *Microvascular Research*, 113:1–8, 2017. doi:10.1016/j.mvr.2017.04.002.
- [10] Abdolamir Karbalaie, Mahnaz Etehadtavakol, Farhad Abtahi, Alimohammad Fatemi, Zahra Emrani, and Björn-Erik Erlandsson. Image enhancement effect on inter and intra-observer reliability of nailfold capillary assessment. *Microvascular Research*, 120:100–110, 2018. doi:10.1016/j.mvr.2018.06.005.

- [11] Praveen Gurunath Bharathi, Michael Berks, Graham Dinsdale, Andrea Murray, Joanne Manning, Sarah Wilkinson, Maurizio Cutolo, Vanessa Smith, Ariane L. Herrick, and Chris J. Taylor. A deep learning system for quantitative assessment of microvascular abnormalities in nailfold capillary images. *Rheumatology (Oxford)*, 62(6):2325–2329, 2023. doi:10.1093/rheumatology/kead026.
- [12] Alexandru Garaiman, Farhad Nooralahzadeh, Carina Mihai, Nicolas Perez Gonzalez, Nikitas Gkikopoulos, Mike Oliver Becker, Oliver Distler, Michael Krauthammer, and Britta Maurer. Vision transformer assisting rheumatologists in screening for capillaroscopy changes in systemic sclerosis: an artificial intelligence model. *Rheumatology (Oxford)*, 62(3):1140–1150, 2023. doi:10.1093/rheumatology/keac541.
- [13] Hao Yin, Zhiwei Wu, An Huang, Jiaxiong Luo, Junzhao Liang, Jianan Lin, Qian Yao Ye, Mugui Xie, Cong Ye, Xiaosong Li, and Yanxiong Wu. Automated nailfold capillary density measurement method based on improved YOLOv5. *Microvascular Research*, 150:104593, 2023. doi:10.1016/j.mvr.2023.104593.
- [14] S.A. Tuncer, M. Yildirim, T. Tuncer, and M.K. Mulayim. YOLOv8-based system for nail capillary detection on a single-board computer. *Diagnostics*, 14(17):1843, 2024. doi:10.3390/diagnostics14171843.
- [15] L. Zhao, J. Tang, D. Chen, et al. A comprehensive dataset and automated pipeline for nailfold capillary analysis. In *Proceedings of the 21st IEEE International Symposium on Biomedical Imaging*, 2024. doi:10.1109/ISBI56537.2024.10635443.
- [16] Abdolamir Karbalaie, Farhad Abtahi, and Charlotte K. Häger. Participant-aware model validation for repeated-measures data: Comparative cross-validation study. *JMIR AI*, 5:e87728, 2026. doi:10.2196/87728.
- [17] Omar S. Emam, Mona Ebadi Jalal, Begonya Garcia-Zapirain, and Adel S. Elmaghraby. Artificial intelligence algorithms in nailfold capillaroscopy image analysis: A systematic review. *medRxiv*, 2024. Preprint; not peer-reviewed. doi:10.1101/2024.07.28.24311154.
- [18] Maged Helmy, Trung Tuyen Truong, Eric Jul, and Paulo Ferreira. Deep learning and computer vision techniques for microcirculation analysis: A review. *Patterns*, 4(1):100641, 2023. doi:10.1016/j.patter.2022.100641.
- [19] Oya Koker, Sezgin Sahin, Mehmet Yildiz, Amra Adrovic, and Ozgur Kasapcopur. The emerging paradigm in pediatric rheumatology: Harnessing the power of artificial intelligence. *Rheumatology International*, 2024. doi:10.1007/s00296-024-05661-x.
- [20] J. Doncheva-Dilova, N. Stoilov, and Vl. Boyadzhieva. Artificial intelligence (ai) and capillaroscopy: New horizons. *Rheumatology*, 33(3), 2025. doi:10.35465/rbj.rbj.406.
- [21] Vishal Gaurav, Chander Grover, Mehul Tyagi, and Suman Saurabh. Artificial intelligence in diagnosis and management of nail disorders: A narrative review. *Indian Dermatology Online Journal*, 2025. doi:10.4103/idoj.idoj_460_24.

- [22] Tejrao Panjabrao Marode, Vikas K. Bhangdiya, Shon Nemane, Dhiraj Tulaskar, Vaishnavi M. Sarad, K. Sankar, Sonam Chopade, Ankita Avthankar, Manish Bhaiyya, and Madhusudan B. Kulkarni. Artificial intelligence meets nail diagnostics: Emerging image-based sensing platforms for non-invasive disease detection. *Bioengineering*, 13(1):75, 2026. doi:10.3390/bioengineering13010075.
- [23] Cristiana Sieiro Santos, José Miguel Sequí-Sabater, and Diego Benavent. Artificial intelligence in systemic sclerosis: Clinical applications, challenges and future directions. *Arthritis Care & Research*, 2026. doi:10.1002/acr.80068.
- [24] Andrea C. Tricco, Erin Lillie, Wasifa Zarin, et al. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine*, 169(7):467–473, 2018. doi:10.7326/M18-0850.
- [25] Youngjin Park, Sehyun Kye, and Ojin Lee. Detection of nailfold capillaries using smartphone microscopy and deep learning analysis. *Transactions of the Korean Institute of Electrical Engineers*, 2026. doi:10.5370/KIEE.2026.75.1.166.
- [26] L. Pérez Abad, M.Á. Aibar Arregui, J. Aramburu Llorente, J. Ramón y Cajal Calvo, A. Andrés Gracia, P. Revilla Martí, S. Atienza Ayala, C. Lahuerta Pueyo, A. Campos Sáenz de Santamaría, E. Ramos Ibáñez, and B.D.C. Gracia Tello. Unlocking the potential of nailfold videocapillaroscopy in diagnosing and staging wild-type transthyretin amyloidosis: A preliminary approach. *Medicina Clínica*, 2026. doi:10.1016/j.medcli.2025.107283.
- [27] B.C. Gracia Tello, E. Ramos Ibáñez, L. Saez Comet, A. Guillén Del Castillo, C.P. Simeón Aznar, A. Selva O’Callaghan, G. Espinosa, G. Lledó, M. Freire Dapena, E. Martinez Robles, J.J. Ríos, J.A. Todolí Parra, B. Marí Alfonso, N. Ortego Centeno, A. Marín Ballvé, J.L. Callejas Rubio, V. Fonollosa Plá, and P. Fanlo. External clinical validation of automated software to identify structural abnormalities and microhaemorrhages in nailfold videocapillaroscopy images. *Clinical and Experimental Rheumatology*, 41(8):1605–1611, 2023. doi:10.55563/clinexprheumatol/m6ob13.
- [28] Neşe Çabuk Çelik and Elif Altunel Kılınç. Comparison of artificial intelligence and rheumatologists in nailfold capillaroscopic evaluation. *Revista da Associação Médica Brasileira*, 2026. doi:10.1590/1806-9282.20251389.
- [29] Borja Gracia Tello, Eduardo Ramos Ibáñez, Patricia Fanlo Mateo, Luis Sáez Cómet, Elena Martínez Robles, Juan José Ríos Blanco, Begoña Marí Alfonso, Gerard Espinosa Garriga, José Todolí-Parra, Norberto Ortego-Centeno, José Luis Callejas Rubio, Mayka Freire Dapena, Adela Marín Ballvé, Albert Selva O’Callaghan, Alfredo Guillén del Castillo, Carmen Pilar Simeón-Aznar, and Vicent Fonollosa Pla. The challenge of comprehensive nailfold videocapillaroscopy practice: a further contribution. *Clinical and Experimental Rheumatology*, 2022. doi:10.55563/clinexprheumatol/6usce8.
- [30] Gema M Lledó-Ibáñez, Luis Sáez Comet, Mayka Freire Dapena, Miguel Mesa Navas, Miguel Martín Cascón, Alfredo Guillén del Castillo, Carmen Pilar Simeon, Elena Martinez Robles, José Todolí Parra, Di-

ana Cristina Varela, Génesis Maldonado, Adela Marín, Laura Pérez Abad, Jimena Aramburu, Laura Vela, Eduardo Ramos Ibáñez, and Borja del Carmelo Gracia Tello. Capi-detect: machine learning in capillaroscopy reveals new variables influencing diagnosis. *Rheumatology*, 64(6):3667–3675, 06 2025. doi:10.1093/rheumatology/keaf073.

[31] L. Ozturk, C. Laclau, C. Boulon, M. Mangin, E. Braz-Ma, J. Constans, L. Dari, and C. Le Hello. Analysis of nailfold capillaroscopy images with artificial intelligence: data from literature and performance of machine learning and deep learning from images acquired in the SCLEROCAP study. *Microvascular Research*, 157:104753, 2025. doi:10.1016/j.mvr.2024.104753.

[32] Hang Thi Phuong Nguyen and Hieyong Jeong. Capillary dataset: A dataset of nail-fold capillaries captured by microscopy for diabetes detection. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD 2025)*, 2025. doi:10.1145/3711896.3737425.

[33] Mohamed Tawfik and Wael Badawy. A comparative evaluation of deep learning-based capillaroscopy image analysis using YOLOv8–YOLOv12 models for microvascular diagnostics. *Neural Computing and Applications*, 2026. doi:10.1007/s00521-026-11847-0.

[34] Gary S Collins, Karel G M Moons, Paula Dhiman, Richard D Riley, Andrew L Beam, Ben Van Calster, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385:e078378, 2024. doi:10.1136/bmj-2023-078378.

[35] John Mongan, Linda Moy, and Charles E. Kahn, Jr. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. *Radiology: Artificial Intelligence*, 2(2):e200029, 2020. doi:10.1148/ryai.2020200029.

[36] Baptiste Vasey, Myura Nagendran, Bruce Campbell, David A Clifton, Gary S Collins, Spiros Denaxas, Alastair K Denniston, Livia Faes, Bart Geerts, Mudathir Ibrahim, Xiaoxuan Liu, Bilal A Mateen, Piyush Mathur, Melissa D McCradden, Lauren Morgan, Johan Ordish, Campbell Rogers, Suchi Saria, Daniel S W Ting, Peter Watkinson, Wim Weber, Peter Wheatstone, and Peter McCulloch. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ*, 377:e070904, 2022. doi:10.1136/bmj-2022-070904.

[37] Xiaoxuan Liu, Samantha Cruz Rivera, David Moher, Melanie J. Calvert, and et al. Denniston. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nature Medicine*, 26:1364–1374, 2020. doi:10.1038/s41591-020-1034-x.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SuppAINFCmanuscriptBMCfinal.pdf](#)
- [PrismaAINFCmanuscriptBMCfinal.pdf](#)