Eurasia Specialized Veterinary Publication

International Journal of Veterinary Research and Allied Science

ISSN:3062-357X

2026, Volume 6, Issue 1, Page No: 181-192

Available online at: www.esvpub.com/

From Computer Vision to Clinical Decision Support in Veterinary Medicine: A Critical Review of Diagnostic Performance, Dataset Representativeness, Distribution Shift, Explainability, Calibration, Workflow Integration, Human Oversight, and Patient-Safety Risks, 2017–2026

Noah Williams[1]*, Grace Collins[1], Ethan Brooks[2], Daniel Green[3]

[1]Department of Veterinary Computer Vision and Clinical Decision Support, Faculty of Veterinary Science, University of Sydney, Sydney, Australia.
[2]Department of Veterinary Diagnostic Performance and AI Calibration, Faculty of Veterinary Science, University of Queensland, Brisbane, Australia.
[3]Department of Veterinary AI Safety and Workflow Integration, Faculty of Veterinary Science, University of New South Wales, Sydney, Australia.

***E-mail** ✉ noah.williams@sydney.edu.au

ABSTRACT

Computer vision is moving from experimental veterinary image analysis toward clinical decision support, yet high reported diagnostic performance does not establish transportability, reliable uncertainty estimation, workflow value, or patient safety. This critical review evaluates veterinary computer-vision evidence from 2017–2026 across acquisition, annotation and ground-truth construction, classification, detection, segmentation, measurement, diagnostic performance, validation, representativeness, distribution shift, explainability, calibration, abstention, workflow integration, human–AI interaction, oversight, safety, and governance. A transparent targeted selection captured 63 records; nine exact duplicates were removed, 54 unique records were screened, six were excluded at title/abstract assessment, 48 underwent detailed eligibility assessment, 11 were excluded, and 37 evidence units were retained. The literature shows widening veterinary applications in radiography, ultrasound, pathology, ophthalmic imaging, and movement analysis, but uneven evidential maturity. Controlled comparisons can yield strong task-level results, whereas independent practice-sourced validation can reveal substantial degradation and clinically important misses. The review therefore proposes that clinical translation be judged across distinct boundaries: input validity, target validity, transportability, uncertainty reliability, workflow fit, meaningful human oversight, and fail-safe response. These boundaries form an analytical architecture, not a validated score. Evidence for external generalization, calibration, human-factor effects, prospective workflow benefit, and patient-safety outcomes remains thinner than evidence for technical performance. Methodological principles transferred from human medical AI also require explicit veterinary validation because species structure, acquisition environments, professional roles, and regulatory contexts differ. Clinical use should therefore remain claim-specific, supervised, and bounded by the populations, devices, reference standards, and settings actually tested. The review does not establish causal effectiveness or universal readiness and emphasizes task-, species-, device-, and setting-specific validation.

Keywords: Veterinary artificial intelligence, Computer vision, Diagnostic imaging, External validation, Calibration, Clinical decision support

Received: 07 April 2026
Revised: 20 June 2026
Accepted: 23 June 2026

How to Cite This Article: Williams N, Collins G, Brooks E, Green D. From Computer Vision to Clinical Decision Support in Veterinary Medicine: A Critical Review of Diagnostic Performance, Dataset Representativeness, Distribution Shift, Explainability, Calibration, Workflow Integration, Human Oversight, and Patient-Safety Risks, 2017–2026. Int J Vet Res Allied Sci. 2026;6(1):181-92. https://doi.org/10.51847/PjJpgxHbqb

Introduction

Computer vision has expanded across veterinary diagnostics and animal-health research, but the literature remains heterogeneous in species, modalities, datasets, targets, and validation designs [1–3]. This matters because a model

that accurately labels a research image is not automatically a clinically reliable decision-support system. Veterinary reviews increasingly identify the same translational gap: strong technical results can coexist with uncertain representativeness, limited external validation, weakly characterized uncertainty, and unresolved workflow consequences. A current 2026 synthesis confirms continued expansion of veterinary imaging AI while emphasizing that implementation evidence has not matured at the same pace [4].

This review examines the chain from image acquisition to clinical action rather than treating accuracy as the endpoint. Its contribution is a critical separation of reported performance from transportability, calibration, explanation, human interaction, governance, and patient-safety boundaries. These distinctions are used to constrain claims, preserve species and setting limits, and identify what must be demonstrated before an algorithm can reasonably support veterinary decisions.

Scope and critical-review approach

The review treats computer vision as a clinical system problem spanning data generation, annotation, modeling, validation, interpretation, workflow, and oversight. Veterinary diagnostic-imaging scholarship shows that limitations arise beyond classifier accuracy, including dataset design, generalization, interpretability, and implementation [3]. Direct veterinary evidence is therefore distinguished from methodological principles transferred from human medical AI, and neither is treated as evidence of clinical effectiveness unless that outcome was evaluated.

Appraisal is claim-specific. A study may strongly support a narrow detection, segmentation, or measurement claim while providing little evidence for transportability or decision consequence. Contradictory findings are retained rather than averaged across incompatible tasks. Original classifications introduced here are explicitly proposed syntheses: they organize observed evidence but are not validated clinical rules. Species, device, setting, reference-standard, temporal, measurement, and actionability boundaries remain visible throughout.

Literature-selection strategy

The literature selection used a transparent, targeted critical-review strategy rather than an exhaustive systematic-search claim. Three discovery streams captured 63 records. After nine exact DOI duplicates were removed, 54 unique records were screened; six were excluded at title/abstract assessment, leaving 48 for detailed eligibility assessment. Eleven further records were excluded for eligibility or claim fit, redundancy, or non-primary commentary status, yielding 37 included evidence units. No reports were recorded as not retrieved.

Selection prioritized peer-reviewed journal sources from 2017–2026 with verifiable DOI metadata and a direct role in one or more review domains. Veterinary evidence was supplemented only where translational questions—particularly calibration, explainability, human factors, or governance—required methodological foundations not adequately represented in veterinary studies [1]. Such cross-domain evidence supports methods and boundaries, not veterinary effect estimates.

Veterinary computer-vision use cases

Veterinary computer vision now addresses more than static radiograph classification. Automated canine gait analysis illustrates functional phenotyping in which video is converted into movement measurements rather than a disease label [5]. A quantified gait feature may support assessment while remaining separate from orthopedic or neurologic diagnosis.

Image-based tasks are also diversifying. Deep learning has been applied to morphological segmentation in canine diffuse large B-cell lymphoma, demonstrating region-level analysis rather than image-level categorization [6]. In canine point-of-care ultrasound, AI models target diagnostic interpretation under an acquisition process that is operator- and device-dependent [7]. These examples motivate a use-case distinction by information function: classification or detection, segmentation or measurement, and functional observation. Each produces different evidence requirements. A successful narrow task should therefore be interpreted according to what it actually outputs, not according to the generic label “AI diagnosis.”

Image acquisition

Image acquisition is an upstream component of model validity because positioning, device characteristics, protocol, operator technique, and image quality determine the input distribution. Point-of-care ultrasound makes this dependence especially visible: the input is jointly produced by anatomy, probe placement, operator behavior,

and hardware rather than passively recorded [7]. Similar concerns arise when radiographic exposure, positioning, collimation, or acquisition systems differ across clinics.

For evaluation, acquisition failure should not be silently merged with interpretive error. This review therefore proposes separating three questions: whether an image is technically adequate, whether it lies within the validated acquisition domain, and whether the model interprets an adequate in-domain image correctly. This distinction does not assume that acquisition effects are always large; it prevents downstream accuracy from obscuring an upstream input problem. Magnitude and clinical consequence remain task-, device-, and setting-specific and require direct testing.

Annotation and ground-truth construction

Ground truth in veterinary computer vision is constructed, not automatically given. Machine learning has been used to classify technical quality of canine and feline thoracic radiographs, showing that positioning and adequacy can themselves become explicit reference labels [8]. Such labels answer a different question from disease truth and may vary with professional standards or intended use.

Annotation strategy also changes how evidence is produced. A deep active-learning study of canine thoracic radiograph segmentation shows how models can target informative cases for expert labeling [9]. Broader medical-image literature likewise identifies active learning and human-in-the-loop interaction as mechanisms for concentrating expert effort [10]. These approaches may improve efficiency, but they do not remove disagreement, selection bias, or imperfect reference standards. The review therefore separates annotation provenance, reader ambiguity, and sample-selection strategy. Efficient labeling is useful only if the resulting reference is sufficiently reliable for the claim tested.

Classification and detection tasks

Classification and detection convert an image or region into a predefined label, but the clinical meaning of that label depends on target definition. A point-of-care ultrasound classifier can be evaluated against a specified diagnostic target without thereby validating broader clinical reasoning [7]. This is particularly important when datasets use descriptive findings, proxy labels, or expert impressions rather than independently confirmed disease states.

This review treats label scope as a claim boundary. Image-level normal/abnormal classification, lesion detection, disease classification, and action-oriented triage are not interchangeable endpoints. As outputs become more actionable, stronger reference standards and representative validation become more important. This is a proposed evidential hierarchy, not an established score. Expert consensus may be appropriate for visual findings, whereas definitive clinical confirmation may be preferable for disease-state claims. Performance should be interpreted only within the label definition and decision context actually evaluated.

Segmentation and measurement tasks

Segmentation and measurement expose intermediate structure that can often be inspected separately from final interpretation. Morphological segmentation in canine lymphoma provides an example in which the model delineates image regions or cellular features rather than returning only a diagnostic class [6]. Active-learning segmentation of canine thoracic radiographs similarly emphasizes dependence on expert reference masks and labeling strategy [9].

Visibility can improve auditability, but visual plausibility is not proof of correctness. A segmentation may appear anatomically reasonable while introducing systematic boundary error, and a small measurement error may be negligible in one context but consequential near a decision threshold. This review therefore separates segmentation accuracy, measurement agreement, and downstream decision validity. Quantitative tasks should be compared with appropriate expert or validated reference measurements, while threshold-related consequences require separate testing. No universal acceptable measurement error is inferred from the selected evidence.

Reported diagnostic performance

Veterinary AI performance is heterogeneous across tasks and study designs [11–13]. In a 2025 comparison of 50 canine and feline radiographic studies, a commercial system approached the best radiologist in descriptive-finding accuracy, with greater specificity but lower sensitivity and no equivalent differential-diagnosis capability [11]. In contrast, a 2026 independent external validation of six commercial platforms on 53 general-practice canine

abdominal cases found mostly low-to-moderate performance, substantial platform variability, and frequent clinically important misses [12]. An earlier thoracic study reported a lower overall error rate for AI than veterinarians, or veterinarians assisted by AI, for the selected primary-lesion task [13].

These findings should not be pooled into a single estimate or winner. They differ in body region, target definition, platform, reference standard, case mix, and workflow. The critical signal is the dependence of performance on what was tested and under which conditions.

Internal versus external validation

Internal validation estimates performance within a development-related environment but does not establish transportability to different clinics, devices, case mixes, or prevalences. The 2025 commercial comparison shows that strong performance can occur in a bounded head-to-head setting [11]. The independent 2026 study shows that performance can be much less favorable when multiple commercial systems are tested on practice-sourced canine abdominal cases with definitive clinical reference diagnoses [12]. These studies are contrasting, not direct replications, because platforms, body regions, targets, and sampling differ.

External validation is therefore treated as a separate evidence boundary rather than another metric. The question is whether clinically relevant behavior persists when the data-generating environment changes in ways resembling intended use. **Table 1** presents computer-vision transportability claims under internal testing, bounded external validation, degradation, and label mismatch.

Table 1. Computer-vision transportability claims under internal testing, bounded external validation, degradation, and label mismatch

Validation configuration	Data-generating environments compared	What changed	Observed error behavior	Claim permitted	Next transport test
Internal concordance only	**Internal:** Strong held-out result **External:** Not demonstrated	Same site/device/case mix	Stable internal metrics	**Use:** Research claim only **Limit:** Transportability unresolved	**Test:** Test changed distributions **Basis:** Proposed
Bounded external transport	**Internal:** Internal result retained **External:** Independent task confirmation	Similar species/target	Comparable error profile	**Use:** Claim remains task-bounded **Limit:** Wider transfer unknown	**Test:** Monitor new sites/devices **Basis:** Proposed
External degradation	**Internal:** Internal result optimistic **External:** Lower or variable performance	Practice prevalence/device differs	Misses or asymmetry emerge	**Use:** Confidence decreases **Limit:** Cause may be multifactorial	**Test:** Revalidate before use **Basis:** Proposed
Reference-standard mismatch	**Internal:** Labels fit training target **External:** Clinical target differs	Confirmation method changes	Apparent disagreement	**Use:** Narrow the claim **Limit:** Label validity unresolved	**Test:** Rebuild target/reference **Basis:** Proposed

Dataset representativeness

Representativeness is not a single dataset property; it is relative to the claim and intended use. The independent evaluation of commercial veterinary radiology systems used general-practice canine abdominal radiographs with definitive diagnoses and exposed important misses that would be difficult to infer from a curated comparison alone [12]. Practice sourcing, however, does not automatically represent every clinic, breed distribution, device, referral population, or disease prevalence.

Two distinctions are useful. Prevalence representativeness concerns whether the spectrum and frequency of cases resemble the intended population. Decision-unit representativeness concerns whether labels correspond to the unit of clinical inference. A dataset of descriptive findings may support finding detection yet remain insufficient for patient-level diagnosis when relevant history, examination, laboratory data, or disease confirmation are absent. Dataset adequacy should therefore be judged against the permitted claim, not against size or apparent diversity alone.

Species, breed, device, and setting bias

Veterinary imaging contains biologically and technically distinct domains that should not be assumed interchangeable. A canine thoracic deep-learning study used radiographs from two acquisition systems,

illustrating device differences within one species and body region [14]. A separate feline thoracic classifier reinforces the need to test species transfer rather than infer it from canine performance [15]. Neither study quantifies the magnitude of cross-device or cross-species bias.

Setting also changes case mix. Deep learning evaluated in dogs screened for elbow dysplasia operates in a screening population whose disease spectrum and prevalence may differ from symptomatic general-practice or referral cohorts [16]. More broadly, medical-AI evidence shows that distribution shift can alter subgroup performance and fairness, although its magnitude cannot be transferred directly to veterinary breeds or species [17]. A veterinary bias audit should therefore consider species, breed or body-size structure, device, institution, and care setting as separate, potentially interacting domains requiring empirical validation.

Domain and dataset shift

Domain shift occurs when deployment data differ meaningfully from those used for model development. In veterinary imaging, plausible shifts include species, anatomy, disease prevalence, device, institution, positioning, operator technique, and referral setting. Weak external performance on general-practice canine abdominal radiographs illustrates why changed case mix can matter [12]. Device differences were explicit within a canine thoracic dataset [14], while separate feline thoracic modeling reinforces the need to test rather than assume species transfer [15]. Medical-AI evidence further shows that subgroup behavior can change under distribution shift [17]. This review therefore proposes three analytically separate forms: source shift, label shift, and decision-context shift. They organize possible failure mechanisms but do not identify the cause of any individual performance change. Robustness requires testing on genuinely changed distributions, not only random splits of the original dataset.

Explainability

Explainability should not be treated as a certificate of clinical validity. Plausible post-hoc visualizations can create confidence without demonstrating that highlighted image features faithfully caused the model output or that the output is clinically correct [18]. Human-centered evaluation is therefore important: explanations should be tested with intended users for whether they improve appropriate understanding without encouraging overtrust [19]. Methodological reviews also show that different explainability techniques answer different questions and have different limitations, so “explainable” is not a unitary property [20]. Calibration adds another distinction because an intelligible explanation may accompany unreliable probabilities [21]. For veterinary decision support, explanation plausibility, faithfulness, user usefulness, and probability reliability should remain analytically separate. Direct veterinary evidence comparing these dimensions is limited, so their integration here is a critical synthesis rather than a validated veterinary standard.

Calibration and uncertainty

Discrimination describes separation between classes; calibration asks whether predicted probabilities correspond to observed frequencies. Strong discrimination can coexist with overconfident or underconfident probabilities [21]. This matters whenever veterinary AI displays confidence values, applies thresholds, or prioritizes cases. The 2025 commercial comparison reported accuracy, sensitivity, and specificity trade-offs [11]. The 2026 external validation exposed clinically important misses across platforms [12]. Neither pattern should be translated into reliable probabilities without dedicated calibration evidence. Uncertainty also arises from ambiguous images, weak reference standards, unfamiliar devices, or changed case mix. This review therefore treats “performance reported” and “uncertainty characterized” as separate evidence states. Calibration does not establish clinical utility, because even reliable probabilities require an appropriate threshold, workflow, and response. Veterinary studies should also distinguish model uncertainty from uncertainty created by the data-generating process.

Model abstention

A system that must always return a label can convert uncertainty into false precision. This review therefore proposes model abstention as a safety function: when input quality, data-domain fit, or model confidence falls outside validated conditions, the system may defer rather than force classification. The rationale follows from the distinction between discrimination and calibrated probability reliability [21], but veterinary abstention policies remain insufficiently validated.

Abstention is not equivalent to safety. Confidence can be miscalibrated, shift may go undetected, and excessive deferral can increase workload or delay care. A clinically meaningful policy must specify what triggers deferral, what information is communicated, who receives the case, and what alternative assessment is available. Specialist review may also be inaccessible. The proposed rule is therefore conditional: abstention should create an explicit escalation pathway, not simply suppress output. Prospective veterinary studies are needed to test whether such policies reduce harmful error without unacceptable delay or workload.

Workflow integration

Clinical integration changes the unit of evaluation from an algorithm to a sociotechnical system. Veterinary radiology guidance emphasizes intended use, data context, validation design, and clinically meaningful errors when judging algorithms for practice [22]. Early-stage AI evaluation guidance recommends reporting workflow position, human interaction, safety events, and implementation conditions rather than technical metrics alone [23]. Trustworthy-AI guidance adds lifecycle robustness, traceability, usability, fairness, and explainability [24]. Human-factor evidence also shows why interface design matters: incorrect AI suggestions can degrade expert interpretation rather than being harmlessly ignored [25].

Workflow integration should therefore be evaluated by role, timing, and fallback. AI may act as triage, concurrent reader, second reader, measurement assistant, or alerting system, creating different opportunities for correction and failure. The following ledger distinguishes clinically consequential workflow states without assigning an unsupported readiness score. **Table 2** presents recoverability of veterinary AI errors across assistive, concurrent, deferred, and failed-fallback workflows.

Table 2. Recoverability of veterinary AI errors across assistive, concurrent, deferred, and failed-fallback workflows

AI position in workflow	When and how the human sees output	Workflow conditions	How error can be caught	Failure when recovery is unavailable
Bounded assistive use	**Clinical role:** AI supports a defined subtask **Observed interaction:** Output fits intended role	Validated species/device/setting	Human retains decision	**Limit:** Wider benefit unknown **Audit:** Recheck after workflow change **Basis:** Proposed
Concurrent influence	**Clinical role:** AI is visible during interpretation **Observed interaction:** Clinician–AI agreement or disagreement	Expertise/workload matter	Anchoring or correction possible	**Limit:** Interaction effect uncertain **Audit:** Test correct and incorrect advice **Basis:** Proposed
Deferred review	**Clinical role:** AI triages or flags cases **Observed interaction:** Cases reordered or escalated	Urgency/access shape risk	Review timing changes	**Limit:** Delay may offset benefit **Audit:** Audit misses and delays **Basis:** Proposed
Fallback failure	**Clinical role:** No effective escalation **Observed interaction:** Forced output or unresolved disagreement	Staffing/specialist access limited	Error may propagate	**Limit:** Recoverability compromised **Audit:** Redesign routing/oversight **Basis:** Proposed

Human–AI interaction

Human–AI interaction is not a simple addition of human judgment to machine output. Early clinical evaluation guidance treats interaction design as part of the intervention because timing, visibility, authority, and feedback determine how recommendations enter decisions [23]. Controlled reader evidence shows that incorrect AI suggestions can impair interpretation, including among experienced readers [25]. Veterinary systems therefore require direct evaluation of how clinicians use, reject, or defer to outputs.

This review proposes four interaction roles: assistive measurement, concurrent interpretation, confirmatory second reading, and triage. Their risks differ. Measurement can propagate quantitative error; concurrent suggestions can anchor interpretation; second reading may expose missed findings; triage can alter case order before human review. These are not validated performance categories. They clarify why standalone accuracy cannot predict combined human–AI performance and why interfaces should be tested under both correct and incorrect machine advice.

Automation bias

Automation bias becomes clinically relevant when users overweight machine advice or fail to search adequately for alternatives. In a controlled mammography experiment, incorrect AI suggestions reduced reader performance, showing that nominal human oversight does not guarantee correction of algorithmic error [25]. This mechanism is plausible in veterinary imaging, but its magnitude and consequences have not been established directly in veterinarians.

Veterinary studies should therefore test more than satisfaction. Relevant outcomes include detection of incorrect suggestions, confidence change after machine advice, experience-related override behavior, and whether explanation or alert design changes overreliance. Correct advice may improve performance, so automation bias should not be framed as inevitable. The critical question is conditional: under which task, interface, expertise level, and error type does AI appropriately support judgment, and when does it displace independent assessment? This remains a priority for prospective veterinary human-factor research.

Human oversight

Oversight should be an active control function rather than the formal presence of a veterinarian somewhere in the workflow. Comparative screening research shows that different allocations of work between humans and AI can produce different system-level trade-offs [26]. Veterinary clinical guidance recommends decision-support use, maintenance of diagnostic skills, team training, transparency, and quality assurance while retaining veterinarian responsibility for diagnosis and treatment decisions [27]. Veterinary analysis of diagnostic error further suggests that AI can alter the error profile rather than simply remove error [28].

The ACVR and ECVDI position statement adds a direct professional boundary: qualified veterinary professionals should remain involved, training and testing information should be transparent, external validation should be available, and postimplementation performance monitored [29]. Meaningful oversight is therefore proposed here as authority to reject output, information sufficient to recognize disagreement, an escalation route, and feedback from subsequent outcomes. Workload, staffing, and specialist access can constrain these functions.

Patient-safety failure modes

Patient-safety risk can arise between image acquisition and action. Independent external testing documented missed diagnoses and substantial variation among commercial radiology platforms [12]. Incorrect AI advice can also interact with human cognition, creating a pathway in which model error is amplified rather than corrected [25]. Veterinary literature on diagnostic-error reduction similarly indicates that AI may introduce new error structures while reducing selected perceptual errors [28].

This review proposes four safety-origin classes: input failure, model failure, interaction/workflow failure, and governance failure. A missed finding is not automatically patient harm; clinicians may correct it, and consequence depends on urgency and context. Conversely, absence of reported harm does not demonstrate safety. Veterinary best-practice guidance emphasizes quality assurance and continued clinician responsibility [27], while specialty guidance calls for postimplementation monitoring [29]. Direct veterinary evidence linking AI deployment to patient outcomes remains limited relative to technical-performance evidence.

Governance and accountability

Clinical translation requires auditable evidence. STARD-AI, PROBAST+AI, and TRIPOD+AI converge on transparent reporting of data sources, reference standards, model development, evaluation, risk of bias, and applicability [30–32]. These standards do not certify safety; they make assumptions and limitations more assessable. Veterinary adaptation remains necessary because species structure, clinical-data ownership, professional roles, and regulatory pathways differ from human healthcare.

Accountability also extends beyond the clinician operating the software. A 2026 systematic review of health-system AI governance frameworks emphasizes lifecycle responsibilities across development, procurement, deployment, monitoring, and incident response [33]. This review proposes an analogous veterinary responsibility chain involving developers, vendors, procuring organizations, clinicians, specialty bodies, and relevant regulators. The proposal is organizational rather than legal and does not assign jurisdiction-specific liability. Its purpose is to prevent patient safety from depending exclusively on individual vigilance at final use.

Critical appraisal of current evidence

The evidence base is strongest for narrow technical tasks and weaker for the full chain required for dependable clinical decision support. Veterinary reviews identify limitations in generalization and implementation [3]. Independent external testing shows that performance can deteriorate in practice-sourced data [12]. Structured appraisal methods also show that low apparent technical error does not resolve risk of bias or applicability concerns [31].

The literature therefore supports neither blanket rejection nor broad clinical-equivalence claims. Strong task-level performance is credible in selected settings, but major uncertainties concern transportability, calibration, failure recognition, interaction effects, workflow consequences, and patient safety. Contrasting controlled and external evaluations are useful because they expose context dependence rather than yielding a single pooled answer. **Figure 1** separates evidence domains often collapsed into "accuracy" and marks the proposed boundaries that must be crossed before technical performance can support a clinical-use claim.

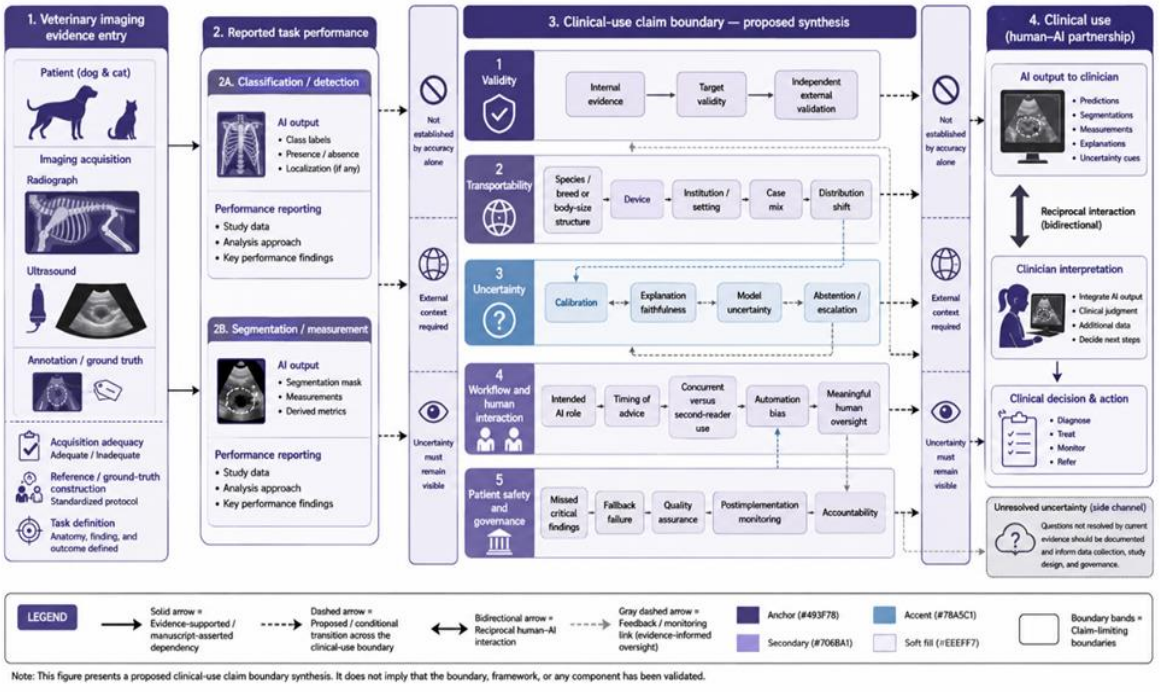

Figure 1. Performance evidence fractures before safe veterinary AI use

Minimum standards for clinical translation

Clinical translation should require more than a favorable internal metric. Veterinary guidance supports alignment between intended use and validation context [22]. Specialty guidance adds transparent datasets, independent external validation, qualified oversight, and postimplementation monitoring [29]. AI reporting and appraisal standards support explicit reference standards, model-development transparency, risk-of-bias assessment, and applicability analysis [30]. These boundaries remain important when models are updated or moved between devices or populations [31]. Transparent reporting also allows readers to judge transportability rather than infer it [32].

This review proposes seven translation gates: input validity, target validity, external transportability, uncertainty characterization, workflow fit, effective human oversight, and safety/governance response. Passing one gate does not compensate automatically for failure at another, and the set is not a validated scoring system. **Figure 2** specifies their conditional relationships, failure routes, and reassessment logic.

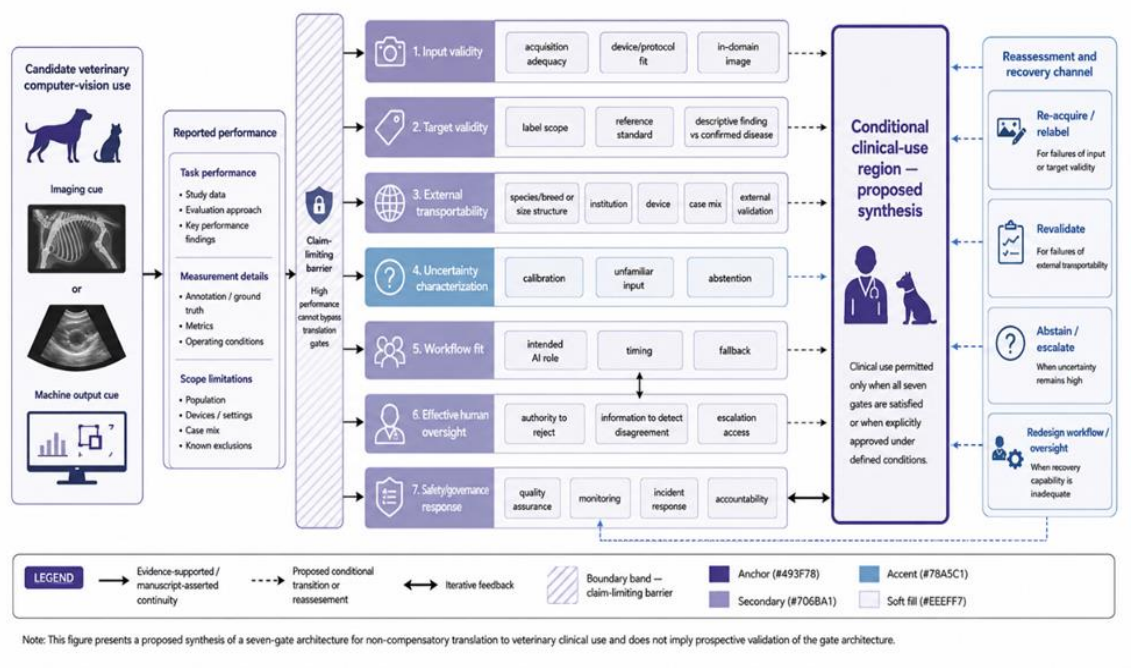

Figure 2. Translation gates for veterinary computer vision under clinical uncertainty

Research priorities

Future veterinary AI studies should target gaps most consequential for translation. A 2026 canine fundus-classification study combined held-out evaluation with explicit calibration analysis, showing that probability reliability can be reported alongside discrimination [34]. A 2025 multi-institution thoracic-radiograph framework estimated vertebral heart size and cardiothoracic ratio against expert measurements, illustrating interpretable intermediate outputs and cross-institution data use [35]. These are design exemplars rather than universal benchmarks.

Actionable disease-oriented tasks require similarly careful target validation. A pilot study of AI-assisted canine urolith composition illustrates why broader representative validation is needed before preliminary image classification becomes definitive clinical inference [36]. A 2026 study evaluating detection of clinically confirmed heart failure in dogs and cats illustrates the value of clinically verified target states [37]. Priorities include prospective multisite validation, temporal and device-shift testing, calibration monitoring, abstention evaluation, version tracking, reader studies, and postdeployment incident surveillance.

Conclusion

Veterinary computer vision has progressed from narrow experimental classification toward diverse diagnostic and measurement applications, but technical performance and clinical readiness remain different evidential claims. This review argues for a boundary-based interpretation of translation: acquisition and target validity, transportability, calibrated uncertainty, workflow fit, meaningful oversight, and safety governance must remain distinguishable rather than being compressed into accuracy. Current evidence supports promising task-specific applications but leaves substantial uncertainty about generalization, human–AI behavior, prospective workflow benefit, and patient outcomes. Clinical use should therefore remain supervised, claim-specific, and restricted to validated populations and settings while stronger veterinary external-validation and implementation evidence develops.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Xiao S, Dhand NK, Wang Z, Hu K, Thomson PC, House JK, et al. Review of applications of deep learning in veterinary diagnostics and animal health. Front Vet Sci. 2025;12:1511522. doi:10.3389/fvets.2025.1511522
2. Hennessey E, DiFazio M, Hennessey R, Cassel N. Artificial intelligence in veterinary diagnostic imaging: A literature review. Vet Radiol Ultrasound. 2022;63(Suppl 1):851-70. doi:10.1111/vru.13163
3. Burti S, Banzato T, Coghlan S, Wodzinski M, Bendazzoli M, Zotti A. Artificial intelligence in veterinary diagnostic imaging: Perspectives and limitations. Res Vet Sci. 2024;175:105317. doi:10.1016/j.rvsc.2024.105317
4. Barillaro G, Minniti S, Mezzatesta S, Macrì F, Torino C, Vilasi AD. Artificial intelligence in veterinary diagnostic imaging: A narrative review of the state of the art and its applications. Vet J. 2026;318:106762. doi:10.1016/j.tvjl.2026.106762
5. Phelan B, Mc Nally T, Cuddy L, Lacey G. Automatic gait analysis in canines using computer vision. Front Vet Sci. 2026;13:1729697. doi:10.3389/fvets.2026.1729697
6. Ancheta K, Psifidi A, Yale AD, Le Calvez S, Williams J. Deep-learning based morphological segmentation of canine diffuse large B-cell lymphoma. Front Vet Sci. 2025;12:1656976. doi:10.3389/fvets.2025.1656976
7. Martinez R, Amezcua KL, Hernandez Torres SI, Winter T, Yankin I, Venn E, et al. Artificial intelligence models for point-of-care ultrasound diagnostics in dogs. Front Vet Sci. 2026;13:1729114. doi:10.3389/fvets.2026.1729114
8. Tahghighi P, Appleby RB, Norena N, Ukwatta E, Komeili A. Classification of the quality of canine and feline ventrodorsal and dorsoventral thoracic radiographs through machine learning. Vet Radiol Ultrasound. 2024;65(4):417-28. doi:10.1111/vru.13373
9. Norena N, Tahghighi P, Ukwatta E, James F, Monteith G, Komeili A, et al. Evaluation of a deep active learning model for the segmentation of canine thoracic radiographs. Vet Radiol Ultrasound. 2026;67(1). doi:10.1111/vru.70083
10. Budd S, Robinson EC, Kainz B. A survey on active learning and human-in-the-loop deep learning for medical image analysis. Med Image Anal. 2021;71:102062. doi:10.1016/j.media.2021.102062
11. Ndiaye YS, Cramton P, Chernev C, Ockenfels A, Schwarz T. Comparison of radiological interpretation made by veterinary radiologists and state-of-the-art commercial AI software for canine and feline radiographic studies. Front Vet Sci. 2025;12:1502790. doi:10.3389/fvets.2025.1502790
12. Ma D, Faulkner JE, Stander N, Raisis A, Joslyn S. Pilot study: External validation of commercial veterinary radiology artificial intelligence services shows deficiencies in interpretation of general practice-sourced canine abdominal radiographs. J Am Vet Med Assoc. 2026;264(8):1022-9. doi:10.2460/javma.25.10.0691
13. Boissady E, de La Comble A, Zhu X, Hespel AM. Artificial intelligence evaluating primary thoracic lesions has an overall lower error rate compared to veterinarians or veterinarians in conjunction with the artificial intelligence. Vet Radiol Ultrasound. 2020;61(6):619-27. doi:10.1111/vru.12912
14. Banzato T, Wodzinski M, Burti S, Longhin Osti V, Rossoni V, Atzori M, et al. Automatic classification of canine thoracic radiographs using deep learning. Sci Rep. 2021;11(1):3964. doi:10.1038/s41598-021-83515-3
15. Banzato T, Wodzinski M, Tauceri F, Donà C, Scavazza F, Müller H, et al. An AI-based algorithm for the automatic classification of thoracic radiographs in cats. Front Vet Sci. 2021;8:731936. doi:10.3389/fvets.2021.731936
16. Hauback MN, Huynh BN, Steiro SED, Groendahl AR, Bredal W, Tomic O, et al. Deep learning can detect elbow disease in dogs screened for elbow dysplasia. Vet Radiol Ultrasound. 2025;66(1). doi:10.1111/vru.13465
17. Ktena I, Wiles O, Albuquerque I, Rebuffi SA, Tanno R, Guha Roy A, et al. Generative models improve fairness of medical classifiers under distribution shifts. Nat Med. 2024;30(4):1166-73. doi:10.1038/s41591-024-02838-6
18. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11). doi:10.1016/S2589-7500(21)00208-9

19. Chen H, Gomez C, Huang CM, Unberath M. Explainable medical imaging AI needs human-centered design: Guidelines and evidence from a systematic review. NPJ Digit Med. 2022;5(1):156. doi:10.1038/s41746-022-00699-2
20. van der Velden BHM, Kuijf HJ, Gilhuijs KGA, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal. 2022;79:102470. doi:10.1016/j.media.2022.102470
21. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: The Achilles heel of predictive analytics. BMC Med. 2019;17(1):230. doi:10.1186/s12916-019-1466-7
22. Joslyn S, Alexander K. Evaluating artificial intelligence algorithms for use in veterinary radiology. Vet Radiol Ultrasound. 2022;63(Suppl 1):871-9. doi:10.1111/vru.13159
23. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ. 2022;377. doi:10.1136/bmj-2022-070904
24. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, et al. FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ. 2025;388. doi:10.1136/bmj-2024-081554
25. Dratsch T, Chen X, Rezazade Mehrizi M, Kloeckner R, Mähringer-Kunz A, Püsken M, et al. Automation bias in mammography: The impact of artificial intelligence BI-RADS suggestions on reader performance. Radiology. 2023;307(4). doi:10.1148/radiol.222176
26. Frazer HML, Peña-Solorzano CA, Kwok CF, Elliott MS, Chen Y, Wang C, et al. Comparison of AI-integrated pathways with human-AI interaction in population mammographic screening for breast cancer. Nat Commun. 2024;15(1):7525. doi:10.1038/s41467-024-51725-8
27. Basran PS, Appleby R, Porter I. Artificial intelligence-assisted interpretation of veterinary radiographs: Opportunities, risks, and best practices for clinicians. Vet Clin North Am Small Anim Pract. 2026;56(5):1105-15. doi:10.1016/j.cvsm.2026.03.018
28. Burti S, Zotti A, Banzato T. Role of AI in diagnostic imaging error reduction. Front Vet Sci. 2024;11:1437284. doi:10.3389/fvets.2024.1437284
29. Appleby RB, Difazio M, Cassel N, Hennessey R, Basran PS. American College of Veterinary Radiology and European College of Veterinary Diagnostic Imaging position statement on artificial intelligence. J Am Vet Med Assoc. 2025;263(6):773-6. doi:10.2460/javma.25.01.0027
30. Sounderajah V, Guni A, Liu X, Collins GS, Karthikesalingam A, Markar SR, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. Nat Med. 2025;31(10):3283-9. doi:10.1038/s41591-025-03953-8
31. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388. doi:10.1136/bmj-2024-082505
32. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385. doi:10.1136/bmj-2023-078378
33. Alami H, Pozelli Sabio R, Pérez EJ, Gagnon MP, Langlois L, Denis JL, et al. Artificial intelligence governance in health systems: Systematic review of frameworks and integrative model proposal. J Med Internet Res. 2026;28. doi:10.2196/87448
34. Smith DA, Dufour VL, Beltran-Grémaud HGJ, Aguirre GD, Beltran WA. A deep learning-based framework for fundus image classification of inherited retinal degeneration in dogs. Sci Rep. 2026. doi:10.1038/s41598-026-60424-x
35. Mekonnen HT, Puig N, Elson A, López E, Martínez P, Campos S, et al. Deep learning framework for vertebral heart size and cardiothoracic ratio estimation in dogs and cats using thoracic radiographs. Front Vet Sci. 2025;12:1612338. doi:10.3389/fvets.2025.1612338
36. Canejo-Teixeira R, Carvalho M, Semião Teixeira G, Lima A, Crowell C, Kwok J, et al. A pilot study on using an artificial intelligence algorithm to identify urolith composition through abdominal radiographs in the dog. Vet Radiol Ultrasound. 2025;66(2). doi:10.1111/vru.70012

37. Ellingsberg L, Solano M, Hicks JM. Performance of an artificial intelligence convolution neural network software for the detection of confirmed heart failure in dogs and cats. Vet Radiol Ultrasound. 2026;67(2). doi:10.1111/vru.70144