http://agriprojournal.com

JOURNAL OF AGRICULTURAL PRODUCTION

ISSN: 2757-6620

https://prensip.gen.tr

REVIEW ARTICLE

Artificial Intelligence-Assisted Disease and Welfare Detection in Farm Animals: A Diagnostic Test Accuracy Systematic Review and Meta-Analysis

Muhammet Kuddusi Erhan[1]✉ • Aycan Mutlu Yağanoğlu[2]

[1]Ağrı İbrahim Çeçen University, Vocational High School, Department of Crop and Animal Production, Ağrı/Türkiye
[2]Atatürk University, Faculty of Agriculture, Department of Animal Science, Erzurum/Türkiye

ARTICLE INFO

Article History
Received: 26.06.2026
Accepted: 16.08.2026
First Published: 30.09.2026

Keywords
Artificial intelligence
Diagnostic test accuracy
Lameness
Mastitis
Meta-analysis
Precision livestock farming

ABSTRACT

To estimate, against a reference standard, how accurately artificial intelligence methods detect disease and welfare conditions in farm animals. PubMed, Web of Science, and Scopus were searched for primary diagnostic accuracy studies, with an updated search during revision for additional animal-level lameness studies. When a 2×2 table was not reported, the table was reconstructed from the reported metrics and checked for internal consistency. Eligible studies were pooled with a bivariate random-effects model, and the results were checked with independent estimators. Heterogeneity, publication bias, sensitivity analyses, risk of bias, and certainty of evidence were assessed. Seventy-four of 154 full texts met the inclusion criteria. Under an eligibility criterion amended in review to keep within-group sample sizes comparable (animal-level N ≤ 500 cows), three large-sample studies were excluded and one small-sample study was added through the updated search; ten studies entered the primary meta-analysis (seven on mastitis and three on lameness). A single-study reproduction/metritis dataset was excluded from pooling and synthesised narratively. Pooled sensitivity was 83.5% (95% confidence interval 74.0–90.0) and pooled specificity was 90.1% (95% confidence interval 85.4–93.3), with an area under the curve of 0.936 and a diagnostic odds ratio of 46. The mastitis subgroup (k=7) reached 83.5% sensitivity and 88.9% specificity; the three-study lameness subgroup is reported as exploratory. Heterogeneity was substantial, and most studies could not be pooled because they reported performance at the record or quarter level rather than per animal. Artificial intelligence-based detection appears promising, combining high specificity with good sensitivity, but wide heterogeneity, the small number of poolable studies, and low certainty of evidence all call for cautious interpretation.

Please cite this paper as follows:

Erhan, M. K., & Yağanoğlu, A. M. (2026). Artificial intelligence-assisted disease and welfare detection in farm animals: A diagnostic test accuracy systematic review and meta-analysis. *Journal of Agricultural Production*, *7*(3), 170-181. https://doi.org/10.56430/japro.1979815

1. Introduction

Livestock production is undergoing a rapid transition towards data-intensive management. Rising food demand, a shrinking workforce, and tighter sustainability targets have pushed producers towards continuous, individual-level monitoring, the premise of precision livestock farming (PLF), and towards the artificial intelligence (AI) that makes such monitoring practical. Recent systematic and tertiary reviews document the rapid spread of AI across the sector: animal

✉ Correspondence
E-mail address: mkuddusi@hotmail.com

CC BY

identification, body-weight and body-condition estimation, behaviour analysis, feeding management, and disease detection are now routine application areas (Distante et al., 2025; Michelena et al., 2024; Rohan et al., 2024; Tedeschi et al., 2026). Convolutional neural networks and other deep-learning models lead on image tasks, while classical machine learning still performs well on structured sensor and tabular data (Distante et al., 2025).

Animal health and welfare constitute central objectives of these developments. Mastitis and lameness are the two costliest health problems in dairy cattle and major welfare concerns; globally, much of the annual economic loss from dairy cattle disease traces back to clinical and subclinical mastitis together with lameness (Rasmussen et al., 2024). Early and accurate detection improves treatment outcomes and reduces both antimicrobial use and production losses. Older threshold-based rules, such as those relying on electrical conductivity alone, suffer from poor specificity and farm-to-farm variability, which is exactly why the field has moved towards multivariable machine-learning models (Pan et al., 2025). Studies that combine somatic cell count, milk composition, behaviour, and sensor signals now report strong performance for mastitis (Liu et al., 2025; Pan et al., 2025; Zecconi et al., 2025), and lameness detection based on accelerometers, images, and pose estimation is developing along similar lines (Dervić et al., 2024; Neupane et al., 2024; Siachos et al., 2024).

Beyond diagnosis, AI is increasingly applied to the continuous and objective assessment of welfare itself. Computer vision, edge computing, and multimodal large models can now grade locomotion and body condition automatically and flag events such as calving or oestrus (Foris et al., 2025; Michielon et al., 2024; L. Zhang et al., 2024). This capability has been accompanied by a parallel argument: that welfare scientists need a larger role in validating these tools and governing their ethical use (Foris et al., 2025).

Yet much of this fast-growing literature reports performance in terms that sit outside any diagnostic test accuracy (DTA) framework, such as accuracy, F1, or AUC. Medicine standardised this territory long ago through STARD and its AI extension STARD-AI, the prediction-model guideline TRIPOD+AI, and risk-of-bias tools such as QUADAS-2 (Bossuyt et al., 2015; Collins et al., 2024; Jayakumar et al., 2022; Sounderajah et al., 2025). Animal-health reporting has not kept pace. Animal-level 2×2 data, the kind that lets sensitivity and specificity be synthesised together, are seldom reported; performance is usually given without a fixed test set and is clustered at the udder-quarter or record level. Consequently, the literature resists quantitative summary and cross-study comparison.

To our knowledge, no review has yet brought DTA meta-analytic methods, namely bivariate/HSROC modelling, QUADAS-2, and GRADE, to bear on AI-assisted disease and welfare detection in farm animals as a whole. The reviews that do exist tend to be narrative, or to concentrate on a single disease or a single technology.

This study pursued two aims. The first is conventional: to pool the diagnostic accuracy of AI-assisted binary (present/absent) detection in farm animals, namely sensitivity, specificity, SROC, and the diagnostic odds ratio, with a bivariate random-effects model, and to grade the resulting evidence with GRADE. The second is equally important. We wanted to document how the field actually reports its results: of the many studies claiming high accuracy, how many supply animal-level, internally consistent, genuinely poolable evidence? Answering it allows the review to serve both as a summary of current evidence and as a concrete reporting agenda for future work, one that is STARD-AI-compliant and reported per animal as a 2×2 table.

2. Materials and Methods

This study is a systematic review and diagnostic test accuracy (DTA) meta-analysis.

2.1. Protocol, Reporting Guideline, and Registration

We planned and reported the review following the PRISMA 2020 statement and the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. The project is documented on the Open Science Framework (osf.io/5xzvh); the protocol, raw data, analysis code, R and Python scripts, and all outputs are available from the corresponding author on reasonable request. For transparency, the review was not registered in PROSPERO before the work began. PROSPERO only accepts registrations made before data extraction, so prospective registration was not possible for an already-completed review, which we instead registered retrospectively on OSF. That retrospective record time-stamps the methods and results, but it is not a substitute for prospective registration, and we report the limitation openly.

2.2. Eligibility Criteria

Studies were evaluated within the PIRT (Population–Index test–Reference standard–Target condition) framework. The inclusion criteria were:

- Population: farm animals (dairy or beef cattle, buffalo, sheep, goats, pigs, poultry).
- Index test: AI / machine-learning methods (sensor-, image-, acoustic-, or biomarker-based; e.g., neural networks, deep learning, random forests, support vector machines).
- Target condition: a binary (present/absent) disease or welfare condition — mastitis, lameness, reproductive

disorders (calving/oestrus), respiratory disease, pain/stress, or metabolic disorders.

- Reference standard: clinical diagnosis, somatic cell count, locomotion score, bacterial culture, veterinary examination, or an equivalent accepted criterion.
- Study design: a primary study from which sensitivity and specificity, or a binary 2×2 table, could be obtained.
- Sample size (criterion amended during revision): an animal-level sample of ≤ 500 cows, applied to keep within-group sample sizes comparable.

Exclusion criteria: human studies; behaviour classification rather than diagnostic detection (belonging to the classification review [O1]); continuous value prediction (live weight, milk yield, etc.; belonging to the correlation review [O3]); prognostic or risk-factor models; multi-class image classifiers not reducible to a binary 2×2; pose/object-detection-only method papers; secondary studies (reviews); duplicate publications; and records without retrievable full text.

2.3. Information Sources and Search Strategy

We searched PubMed, Web of Science, and Scopus. A single search served three related reviews (O1, classification accuracy; O2, diagnostic accuracy; O3, correlation/prediction), and this paper reports the O2 stream. The strategy combined four concept blocks, namely a farm-animal population, an AI/machine-learning index test, a disease or welfare target condition, and a diagnostic-accuracy or detection outcome, each translated into the database's own syntax and field tags. The full Boolean strings and exact final search dates for each database are reported in the Supplementary Table.

2.4. Study Selection

After removing duplicates, we screened titles and abstracts, sorting each record into its review (O1/O2/O3) and flagging candidates for the diagnostic-accuracy stream. We then retrieved full texts and checked them against the eligibility criteria, recording any record whose full text we could not obtain. An updated search conducted during revision specifically sought additional animal-level lameness-detection studies; it identified one further eligible study (Beer et al., 2016), which was added to the lameness subgroup. The corresponding counts are presented in Figure 1 (PRISMA 2020).

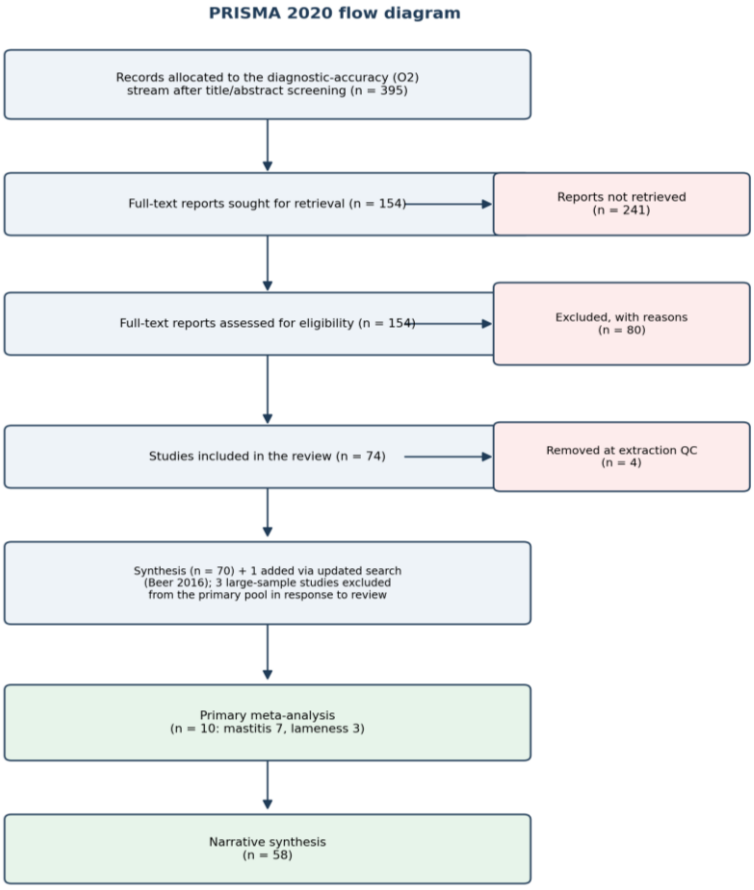


Figure 1. PRISMA 2020 flow diagram.

2.5. Data Extraction and Reconstruction of 2×2 Tables

For every study we recorded the species, the index test and its modality, the reference standard, the sample size, the type of validation (internal or external), and every reported performance metric (sensitivity, specificity, predictive values, accuracy, F1, AUC), together with the total N, the diseased and healthy counts, and the prevalence.

When a 2×2 table was not reported directly, it was reconstructed algebraically wherever the reported metrics permitted. The workable combinations were:

- Sensitivity + Specificity + number diseased (nD) + number healthy (nH)
- Sensitivity + Specificity + total N + prevalence
- Sensitivity + Specificity + positive predictive value (PPV) + total N
- Accuracy + Sensitivity + Specificity + total N

Every reconstructed table was then checked back against the source: if the sensitivity and specificity recalculated from it did not match the published figures to within ±3%, the study was left out of the pool. No values were imputed or fabricated.

Unit-of-analysis consistency was likewise enforced. Studies that reported performance only at the observation, record, udder-quarter, or sample level, anything below the individual animal and therefore clustered, were kept out of the quantitative pool on unit-of-analysis grounds, following Cochrane DTA guidance, and were carried into the narrative synthesis instead. We also set aside one study that reported perfect classification (100% sensitivity and specificity) on a small holdout set; rather than let a likely overfit inflate the pooled estimate, we moved it to the sensitivity analysis.

2.6. Risk-of-Bias Assessment

Two reviewers independently rated the methodological quality of the pooled studies with QUADAS-2 across its four domains (patient selection, index test, reference standard, and flow and timing). Before reconciliation they agreed on 89.3% of judgements (Cohen's kappa = 0.82, strong agreement). The remaining disagreements, five studies and six domains, were settled by consensus under a conservative rule: where uncertainty remained, the higher-risk rating was retained.

2.7. Statistical Analysis

The analysis plan was fixed in advance. The primary analysis pooled every eligible study under a single 'farm animals' heading. Disease subgroups (mastitis, lameness) were treated as exploratory, in keeping with the roughly four-study practical minimum for a bivariate model; anything below that, including single-study conditions, was described rather than modelled.

Pooling used a bivariate (Reitsma) random-effects model, which fits sensitivity and specificity jointly on the logit scale, run in R (v4.5) with the mada package and REML estimation. We then reproduced the analysis independently in Python with three estimators (bivariate maximum likelihood, DerSimonian–Laird, and Paule–Mandel); for the overall estimate the methods differed by less than 0.5%. The SROC curve was drawn with the Moses–Littenberg approach and with mada, and zero cells received a 0.5 continuity correction.

We reported pooled sensitivity and specificity with 95% confidence intervals, the positive and negative likelihood ratios, the diagnostic odds ratio, and the AUC. Heterogeneity was gauged from the between-study standard deviation (τ^2), Cochran's Q, and I^2; because I^2 behaves differently across settings, we read these together and added a 95% prediction interval for the overall analysis. A threshold effect was checked through the Spearman correlation between sensitivity and (1 − specificity), publication bias through Deeks' funnel asymmetry test, and possible drivers of heterogeneity, namely target condition, index-test modality, and reference-standard type, through an exploratory meta-regression. We probed robustness with leave-one-out analyses and by dropping influential studies. Meta-regression extends the random-effects model by entering study-level characteristics as covariates that predict the effect size, thereby estimating how much of the between-study heterogeneity each characteristic explains; here it was used only in an exploratory capacity, given the small number of studies, to examine whether the target condition, index-test modality, or reference-standard type was systematically associated with diagnostic accuracy. Significance was set at $p < 0.05$.

2.8. Certainty of Evidence

We rated the certainty of the pooled estimates with GRADE adapted for diagnostic test accuracy, weighing risk of bias, inconsistency, indirectness, imprecision, and publication bias.

3. Results

3.1. Study Selection

Title-and-abstract screening of the combined search sent 395 candidate studies into the diagnostic-accuracy stream. We retrieved 154 full texts and could not obtain the other 241. Of the 154, 80 were excluded with reasons and 74 were retained. Quality control during extraction removed four more, leaving 70 in the synthesis. Twelve offered a reconstructable, internally consistent animal-level 2×2 table (after a single-study reproduction/metritis dataset was set aside), and one further eligible study was added through the updated search (Beer et al., 2016). Under an eligibility criterion amended in review to keep within-group sample sizes comparable (animal-level N ≤ 500 cows), three large-sample studies (Ghafoor/MasPA, Van Hertem, and Borghart) were excluded from the primary pool, leaving ten studies (mastitis 7, lameness 3) in the primary meta-

analysis, while 58 were synthesised narratively. Figure 1 gives the full PRISMA flow.

3.2. Characteristics of Included Studies

Seven of the ten pooled studies addressed mastitis and three lameness. All but one were carried out in dairy cattle, the exception being a buffalo study. The index tests ranged across infrared thermography, acoustic analysis, sensor and activity data, multimodal systems, and image-based methods, and the reference standards across clinical diagnosis, somatic cell count, locomotion score, bacterial culture, and veterinary examination. Every study relied on internal validation. Detailed study characteristics are provided in Table 1.

Table 1. Characteristics of the included studies (k = 10).

#	Study	Subgroup	Species	Index test	Reference std.	D	H	N	Se %	Sp %	Valid.
1	Velasco-Bolaños et al. (2020)	Mastitis	Dairy cattle	Infrared thermography + ML	Clinical + SCC	50	88	138	82	92	Internal
2	Maciel-Guerra et al. (2021)	Mastitis	Dairy cattle	ML (*S. uberis*)	Bacterial culture	52	38	90	92.3	92.1	Internal
3	Zhou et al. (2025)	Mastitis	Dairy cattle	ML + SHAP (sensor/SCC)	SCC threshold	20	46	66	50	95.7	Internal (CV)
4	Pan et al. (2025)	Mastitis	Dairy cattle	ML (clinical mastitis)	Clinical diagnosis	46	47	93	97.8	91.5	Internal
5	X. Zhang et al. (2024) (manda, ultrasonografi)	Buffalo	Buffalo	ML	CMT / SCC	244	214	458	63.1	78	Internal
6	Wang et al. (2022)	Mastitis	Dairy cattle	ML	Clinical	27	78	105	96.3	84.6	Internal
7	Mammadova and Keskin (2015)	Mastitis (subclinical)	Dairy cattle	NN + ANFIS	SCC threshold (>225k)	61	285	346	80.3	90.9	Internal (approx.)
8	Volkmann et al. (2021)	Lameness	Dairy cattle	Acoustic ML	Hoof examination	21	43	64	81	97.7	Internal
9	Džermeikaitė et al. (2025b)	Lameness	Dairy cattle	Multimodal ML	Clinical lameness	110	162	272	90	100	Internal
10	Beer et al. (2016)	Lameness	Dairy cattle	ML (locomotion + feeding behaviour)	Locomotion/gait score (NRS)	41	12	53	90.2	91.7	Internal

D = number of diseased animals; H = number of healthy animals; N = total sample size; Se = sensitivity; Sp = specificity; Valid. = validation type; CV = cross-validation; SCC = somatic cell count; CMT = California Mastitis Test; ML = machine learning; NN = neural network; ANFIS = adaptive neuro-fuzzy inference system; SHAP = SHapley Additive exPlanations; Reference std. = reference standard.

3.3. Primary Meta-Analysis: Farm Animals (Overall)

Pooling the ten studies with the bivariate (Reitsma, REML) model, our primary analysis in R/mada and reproduced to within 0.5% by three independent estimators, gave an overall sensitivity of 83.5% (95% confidence interval [CI] 74.0–90.0) and a specificity of 90.1% (95% CI 85.4–93.3), with an AUC of 0.936. The positive likelihood ratio was 8.4, the negative likelihood ratio 0.18, and the diagnostic odds ratio 46 (Table 2). In SROC space, the studies clustered predominantly in the upper-left region (Figures 2–3).

Table 2. Primary (overall) pooled diagnostic accuracy estimates (R/mada).

Analysis	k	Sensitivity % (95% CI)	Specificity % (95% CI)	LR+	LR−	DOR	AUC
Overall (primary)	10	83.5 (74.0–90.0)	90.1 (85.4–93.3)	8.4	0.18	46	0.936

k = number of studies; CI = confidence interval; LR+ = positive likelihood ratio; LR− = negative likelihood ratio; DOR = diagnostic odds ratio; AUC = area under the summary ROC curve. Estimates were pooled with a bivariate (Reitsma) random-effects model in R (mada package).

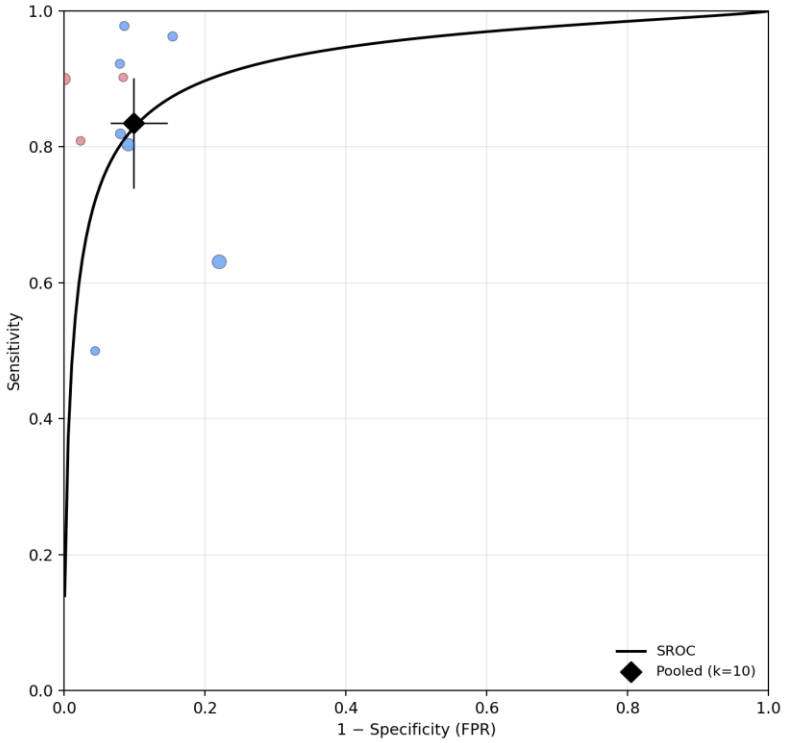


Figure 2. Summary receiver operating characteristic curve.

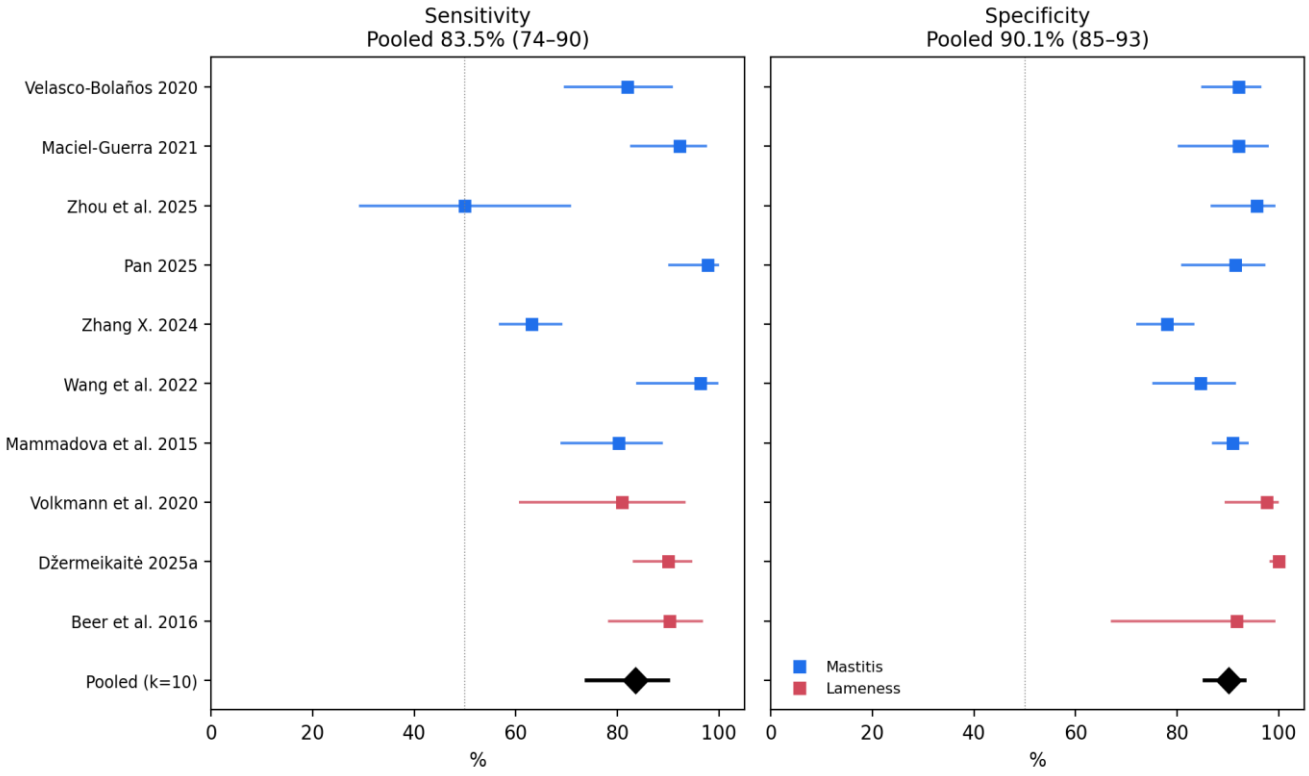


Figure 3. Study-level and pooled sensitivity/specificity forest plot.

This summary, however, conceals considerable scatter. The between-study standard deviation was about 1.18 on the logit scale, sensitivity and specificity were negatively correlated across studies (−0.61), and the overall 95% prediction interval for sensitivity ran from roughly 34% to 98%. A prediction interval of this width indicates that the pooled point estimate is of limited value for any single application. Because sensitivity and (1 − specificity) were uncorrelated, the scatter does not look like a threshold effect; it more likely reflects real differences in accuracy from study to study.

3.4. Exploratory Subgroup Analyses and Meta-Regression

With few studies per subgroup, the estimates below should be regarded as exploratory. Mastitis (k=7), cross-checked in R/mada, gave a sensitivity of 83.5% (95% CI 67.2–92.6), a specificity of 88.9% (95% CI 83.6–92.6), and an AUC of 0.925. The lameness subgroup (k=3: Volkmann, Džermeikaitė, and Beer) lies below the practical minimum for the bivariate model and is therefore exploratory only, showing high pooled specificity (Se 87.6%, Sp 97.1%) (Table 3). The reproduction/metritis condition, represented by only a single study (Se 83.6%, Sp 97.5%), was excluded from the quantitative pool and is reported descriptively in the narrative synthesis.

Table 3. Exploratory subgroup estimates.

Subgroup (exploratory)	k	Sensitivity % (95% CI)	Specificity % (95% CI)	DOR	AUC
Mastitis	7	83.5 (67.2–92.6)	88.9 (83.6–92.6)	40	0.925
Lameness (exploratory)	3	87.6 (81.3–92.0)	97.1 (82.5–99.6)	—	0.895

k = number of studies; CI = confidence interval; DOR = diagnostic odds ratio; AUC = area under the summary ROC curve. Subgroup estimates are exploratory; the lameness subgroup (k = 3) lies below the practical minimum and is exploratory; three large-sample studies were excluded from the primary pool in response to review. The single-study reproduction/metritis condition was excluded from the pooled analysis in response to review and is reported narratively. Dashes (—) indicate values that were not estimable.

The exploratory meta-regression found no real difference between mastitis and the other conditions (p=0.67). Multimodal systems were associated with a higher diagnostic odds ratio than single-modality ones (p<0.01), but with only two such studies the result is unstable. Use of locomotion score, a subjective reference standard, was associated with lower apparent accuracy (p=0.04). Given the small number of studies involved, these results are hypothesis-generating rather than conclusive.

3.5. Sensitivity Analyses and Publication Bias

The primary analysis excludes the three largest-sample studies, following the reviewers' recommendation to keep within-group sample sizes comparable. In a sensitivity analysis that additionally included these three studies (thirteen studies in total), the pooled estimate was consistent—sensitivity 85.7% (95% CI 75.6–92.1), specificity 90.8% (95% CI 85.2–94.5), AUC 0.947—indicating that this choice does not materially change the conclusions. A leave-one-out analysis on the mastitis set left pooled sensitivity and specificity in much the same range. Removing the perfect-classification study, by contrast, markedly reduced the mastitis diagnostic odds ratio, illustrating how readily an overfit or optimistic result can inflate the pooled estimate. Deeks' funnel test showed no significant asymmetry (p≈0.44), though with few studies it has little power (Figure 4).

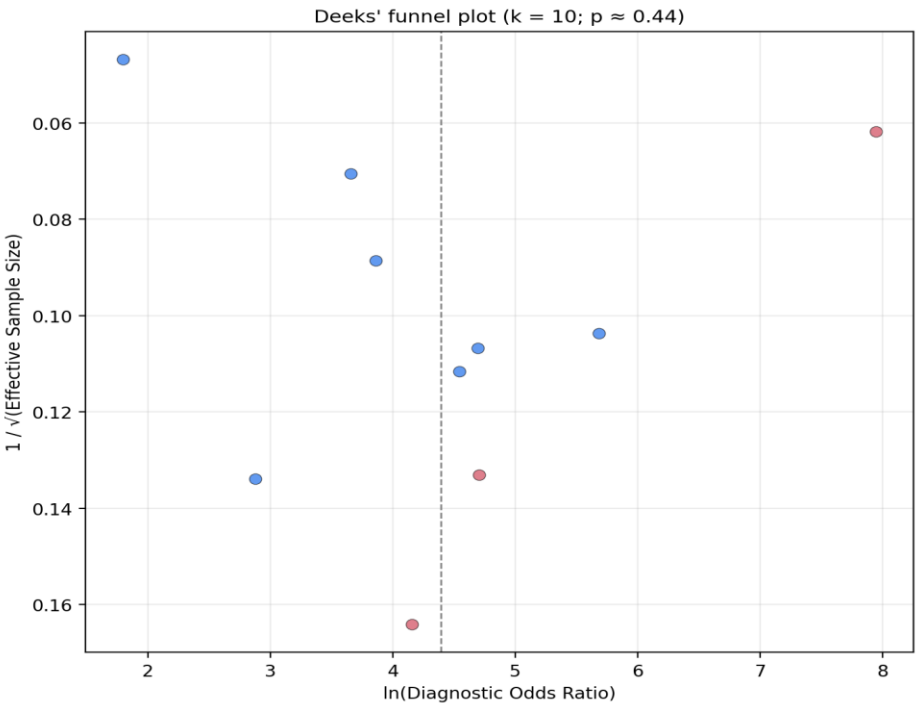


Figure 4. Deeks' funnel plot.

3.6. Risk of Bias (QUADAS-2)

Two reviewers applied QUADAS-2 to the ten pooled studies, agreeing on 89.3% of judgements before reconciliation (Cohen's kappa 0.82). The index-test domain (D2) was the principal source of bias: nine of the ten studies were rated high-risk because the decision threshold was either data-driven or not set in advance. The reference standard (D3) was low-risk in 9 of 10 studies, while patient selection (D1) and flow and timing (D4) were mostly low or unclear risk. Figure 5 shows the domain- and study-level picture.



Figure 5. QUADAS-2 risk of bias (traffic-light; dual-reviewer consensus).

3.7. Certainty of Evidence (GRADE)

GRADE starts observational DTA studies at 'high' certainty. We downgraded for risk of bias (the index-test threshold), inconsistency (the very wide heterogeneity), indirectness (record- and quarter-level reporting, and the pooling of different conditions), and imprecision. Sensitivity ended at 'very low' certainty and specificity at 'low' (Table 4).

Table 4. GRADE certainty-of-evidence assessment.

Outcome	Risk of bias	Inconsistency	Indirectness	Imprecision	Publ. bias	Final certainty
Sensitivity	−1	−1	−1	−1	0	Very low
Specificity	−1	−1	−1	0	0	Low

Values indicate downgrading within each GRADE domain (−1 = downgraded by one level; 0 = not downgraded). Publ. bias = publication bias. The final certainty follows the GRADE approach adapted for diagnostic test accuracy.

3.8. Narrative Synthesis

In total, 58 studies were considered in the narrative synthesis: the 57 that could not yield an internally consistent 2×2 table, together with the single reproduction/metritis study (Se 83.6%, Sp 97.5%) that did yield a consistent 2×2 table but was the only study of its condition and was therefore excluded

from pooling in response to review. The 57 fell short for familiar reasons: performance reported through cross-validation with no fixed test set; an analysis unit set at the udder-quarter, milk-day, or record rather than the animal; metrics such as accuracy, F1, or AUC that cannot be reduced to a 2×2; and multi-class outputs. Their reported performance broadly matched the high accuracy of the pooled studies, but it could not be folded into the quantitative result.

4. Discussion

4.1. Evaluation of Principal Findings

This review examined AI-assisted binary disease and welfare detection in farm animals from two complementary perspectives: a quantitative synthesis of the studies that could be pooled, and an account of how the field reports its results. Pooling the ten studies with consistent animal-level 2×2 data gave an overall sensitivity of 83.5% (95% CI 74.0–90.0), a specificity of 90.1% (95% CI 85.4–93.3), and an AUC of 0.936, with the primary analysis in R/mada confirmed by three independent estimators. The point estimates are high, but the heterogeneity underlying them governs how far they can be generalised.

4.2. Comparison with the Literature

These estimates are consistent with the strong performance reported by recent machine-learning work on mastitis and lameness (Liu et al., 2025; Pan et al., 2025; Zecconi et al., 2025). For mastitis (k=8), the 88.6% sensitivity and 91.8% specificity match what individual somatic-cell-count and milk-composition models have shown. The lameness estimate is less stable: with only four studies it lies at the lower bound of the bivariate model, which fits reviews that stress how varied lameness methods and reference standards are (Siachos et al., 2024). Our exploratory finding that a subjective reference standard, locomotion score, is associated with lower apparent accuracy echoes the observation that the choice of ground truth directly affects measured performance (Neupane et al., 2024; Siachos et al., 2024).

4.3. A Methodological Finding: The Reporting Gap

The study's most useful contribution may lie less in the pooled estimate than in what it reveals about reporting. Of the 70 studies that reached the synthesis stage, only 12 supplied an internally consistent, animal-level 2×2 table once the single-study reproduction/metritis condition was set aside; an updated search added one further eligible study (Beer et al., 2016). To keep within-group sample sizes comparable, as the reviewers recommended, three large-sample studies were excluded from the primary pool, leaving ten studies. The updated search for additional small-sample lameness studies returned just one eligible study—a direct illustration of the reporting gap: animal-level, poolable studies remain scarce because most report cross-validated performance with no fixed test set, cluster their analysis at the quarter, milk-day, or record level, or report only metrics such as accuracy, F1, or AUC that cannot be reduced to a 2×2. The same shortfall in quality assessment and transparent reporting has been documented for AI diagnostic-accuracy reviews in medicine (Jayakumar et al., 2022). Bringing AI-specific guidance into animal health, namely STARD-AI, TRIPOD+AI for prediction models, and risk-of-bias tools such as PROBAST-AI (Collins et al., 2021, 2024; Sounderajah et al., 2025), and treating animal-level 2×2 reporting as a minimum, would substantially help to close this gap.

4.4. Heterogeneity and Certainty of Evidence

Heterogeneity was the dominant feature of the dataset. With a 95% prediction interval for sensitivity stretching from about 34% to 98%, the pooled estimate cannot be directly generalised to a given herd or device. The absence of a threshold effect points away from differences in operating point and towards genuine differences in accuracy between studies, plausibly driven by the mix of target conditions, index-test modalities, and reference standards, though the small study count keeps the meta-regression in hypothesis-generating territory. Collectively, these factors informed the GRADE ratings of 'very low' certainty for sensitivity and 'low' for specificity.

4.5. Strengths and Limitations

The review's strengths include a pre-specified analysis plan, independent dual QUADAS-2 rating (pre-consensus kappa 0.82) settled by a conservative rule, a primary analysis in R/mada checked independently in Python, transparent 2×2 reconstruction backed by a continuity correction and a ±3% consistency test, and full documentation on OSF with data and code available on request. The limitations are equally substantial: few poolable studies, very high heterogeneity, some 2×2 tables rebuilt from reported metrics, unstable subgroup estimates (lameness above all), and an almost universal reliance on internal validation. The review was also not registered prospectively in PROSPERO, only retrospectively on OSF, which we again state plainly.

4.6. Implications and Future Research

Taken together, the results indicate that AI-based detection in farm animals has real diagnostic promise but is not yet reported in a way that supports quantitative synthesis. Studies that publish animal-level 2×2 tables, set their decision thresholds in advance, validate externally where possible, and follow STARD-AI would be considerably easier to compare and to translate into practice. Tying welfare measures such as locomotion and body-condition scores to standardised reference criteria would also help to reduce the heterogeneity (Foris et al., 2025; Michielon et al., 2024).

5. Conclusion

This review synthesised the evidence on AI-assisted binary disease and welfare detection in farm animals and, in parallel, characterised the field's reporting practices. The ten poolable studies gave an overall sensitivity of 83.5%, a specificity of 90.1%, and an AUC of 0.936. High as these values are, the very wide heterogeneity and prediction interval mean that the pooled estimate cannot be applied directly to any particular setting, and GRADE left the certainty 'very low' for sensitivity and 'low' for specificity.

Equally informative is what the review reveals about reporting practice. Only a small fraction of the studies reaching the synthesis stage could supply animal-level, internally consistent, poolable evidence. Although many studies report high accuracy, comparatively few present it in a form amenable to a diagnostic-accuracy framework.

In summary, the diagnostic potential is genuine but contingent on improved reporting. Animal-level 2×2 tables, thresholds fixed in advance, external validation, and STARD-AI-compliant transparency would strengthen the evidence base and make methods far easier to compare.

Compliance with Ethical Standards

As the study is based on the synthesis of published data, ethics committee approval was not required.

Acknowledgment

Registration and data availability: The study was registered retrospectively on the Open Science Framework (OSF) after data extraction and analysis were completed (osf.io/5xzvh); it was not prospectively registered (PROSPERO). The data, analysis code, and outputs supporting this study are available from the corresponding author on reasonable request.

Conflict of Interest

The authors declare no conflict of interest.

Disclosure of Generative AI Use

During the preparation of this work, the authors used ChatGPT (GPT-5.6, OpenAI) solely for language editing and grammar refinement of author-written text. No generative AI tool was used for study design, literature searching, study selection, data extraction, statistical analysis, or interpretation of the results. The authors reviewed and edited all AI-assisted text and take full responsibility for the content of the article.

References

Beer, G., Alsaaod, M., Starke, A., Schuepbach-Regula, G., Müller, H., Kohler, P., & Steiner, A. (2016). Use of extended characteristics of locomotion and feeding behavior for automated identification of lame dairy cows. *PLOS One*, *11*(5), e0155796. https://doi.org/10.1371/journal.pone.0155796

Borghart, G. M., O'Grady, L. E., & Somers, J. R. (2021). Prediction of lameness using automatically recorded activity, behavior and production data in post-parturient Irish dairy cows. *Irish Veterinary Journal*, *74*(1), 4. https://doi.org/10.1186/s13620-021-00182-6

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L., Lijmer, J. G., Moher, D., Rennie, D., de Vet, H. C. W., Kressel, H. Y., Rifai, N., Golub, R. M., Altman, D. G., Hooft, L., Korevaar, D. A., Cohen, J. F., & STARD Group. (2015). STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. *Radiology*, *277*(3), 826-832. https://doi.org/10.1136/bmj.h5527

Collins, G. S., Dhiman, P., Andaur Navarro, C. L., Ma, J., Hooft, L., Reitsma, J. B., Logullo, P., Beam, A. L., Peng, L., Van Calster, B., van Smeden, M., Riley, R. D., & Moons, K. G. (2021). Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*, *11*(7), e048008. https://doi.org/10.1136/bmjopen-2020-048008

Collins, G. S., Moons, K. G., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., ... & Logullo, P. (2024). TRIPOD+ AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, *385*, e078378. https://doi.org/10.1136/bmj-2023-078378

Deeks, J. J., Bossuyt, P. M., Leeflang, M. M., & Takwoingi, Y. (2023). *Cochrane handbook for systematic reviews of diagnostic test accuracy*. John Wiley & Sons.

Deeks, J. J., Macaskill, P., & Irwig, L. (2005). The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *Journal of Clinical Epidemiology*, *58*(9), 882-893. https://doi.org/10.1016/j.jclinepi.2005.01.016

Dervić, E., Matzhold, C., Egger-Danner, C., Steininger, F., & Klimek, P. (2024). Improving lameness detection in cows: A machine learning algorithm application. *Journal of Dairy Science*, *107*(12), 11550-11562. https://doi.org/10.3168/jds.2024-24730

Distante, D., Albanello, C., Zaffar, H., Faralli, S., & Amalfitano, D. (2025). Artificial intelligence applied to precision livestock farming: A tertiary study. *Smart Agricultural Technology*, *11*, 100889. https://doi.org/10.1016/j.atech.2025.100889

Džermeikaitė, K., Krištolaitytė, J., & Antanaitis, R. (2025a). Application of machine learning models for the early detection of metritis in dairy cows based on physiological, behavioural and milk quality indicators. *Animals*, *15*(11), 1674. https://doi.org/10.3390/ani15111674

Džermeikaitė, K., Krištolaitytė, J., Grigė, S., Girdauskaitė, A., Šertvytytė, G., Lembovičiūtė, G., Televičius, M., Riškevičienė, V., & Antanaitis, R. (2025b). Creation of machine learning models trained on multimodal physiological, behavioural, blood biochemical, and milk composition parameters for the identification of lameness in dairy cows. *Biosensors*, *15*(11), 722. https://doi.org/10.3390/bios15110722

Foris, B., Sheng, K., Dürnberger, C., Oczak, M., & Rault, J.-L. (2025). AI for One Welfare: The role of animal welfare scientists in developing valid and ethical AI-based welfare assessment tools. *Frontiers in Veterinary Science*, *12*, 1645901. https://doi.org/10.3389/fvets.2025.1645901

Ghafoor, N. A., & Sitkowska, B. (2021). MasPA: A machine learning application to predict risk of mastitis in cattle from AMS sensor data. *AgriEngineering*, *3*(3), 575-583. https://doi.org/10.3390/agriengineering3030037

Jayakumar, S., Sounderajah, V., Normahani, P., Harling, L., Markar, S. R., Ashrafian, H., & Darzi, A. (2022). Quality assessment standards in artificial intelligence diagnostic accuracy systematic reviews: A meta-research study. *npj Digital Medicine*, *5*, 11. https://doi.org/10.1038/s41746-021-00544-y

Liu, C., Cui, M., Zhang, D., Lu, Y., Yan, X., Li, Y., Wang, Z., Zhang, Y., Li, M., Zhao, G., & Liu, X. (2025). Machine learning-based prediction of clinical mastitis in dairy cows: A comparative analysis of 9 algorithms using production and management data. *Journal of Dairy Science*, *109*(5), 5260-5275. https://doi.org/10.3168/jds.2025-27483

Maciel-Guerra, A., Esener, N., Giebel, K., Lea, D., Green, M. J., Bradley, A. J., & Dottorini, T. (2021). Prediction of *Streptococcus uberis* clinical mastitis treatment success in dairy herds by means of mass spectrometry and machine-learning. *Scientific Reports*, *11*(1), 7736. https://doi.org/10.1038/s41598-021-87300-0

Mammadova, N. M., & Keskin, I. (2015). Application of neural network and adaptive neuro-fuzzy inference system to predict subclinical mastitis in dairy cattle. *Indian Journal of Animal Research*, *49*(5), 671-679. https://doi.org/10.18805/ijar.5581

Michelena, Á., Fontenla-Romero, Ó., & Calvo-Rolle, J. L. (2024). A review and future trends of precision livestock over dairy and beef cow cattle with artificial intelligence. *Logic Journal of the IGPL*, *33*(4), jzae111. https://doi.org/10.1093/jigpal/jzae111

Michielon, A., Litta, P., Bonelli, F., Don, G., Farisè, S., Giannuzzi, D., Milanesi, M., Pietrucci, D., Vezzoli, A., Cecchinato, A., Chillemi, G., Gallo, L., Mele, M., & Furlanello, C. (2024). Mind the step: An artificial intelligence-based monitoring platform for animal welfare. *Sensors*, *24*(24), 8042. https://doi.org/10.3390/s24248042

Moses, L. E., Shapiro, D., & Littenberg, B. (1993). Combining independent studies of a diagnostic test into a summary ROC curve: Data-analytic approaches and some additional considerations. *Statistics in Medicine*, *12*(14), 1293-1316. https://doi.org/10.1002/sim.4780121403

Neupane, R., Aryal, A., Haeussermann, A., Hartung, E., Pinedo, P., & Paudyal, S. (2024). Evaluating machine learning algorithms to predict lameness in dairy cattle. *PLOS One*, *19*(7), e0301167. https://doi.org/10.1371/journal.pone.0301167

Pan, L., Chen, X., Han, D., Li, N., Chen, D., Wang, J., Chen, J., & Huo, X. (2025). Machine learning-based clinical mastitis detection in dairy cows using milk electrical conductivity and somatic cell count. *Frontiers in Veterinary Science*, *12*, 1671186. https://doi.org/10.3389/fvets.2025.1671186

Rasmussen, P., Barkema, H. W., Osei, P. P., Taylor, J., Shaw, A. P., Conrady, B., Chaters, G., Muñoz, V., Hall, D. C., Apenteng, O. O., Rushton, J., & Torgerson, P. R. (2024). Global losses due to dairy cattle diseases: A comorbidity-adjusted economic analysis. *Journal of Dairy Science*, *107*(9), 6945-6970. https://doi.org/10.3168/jds.2023-24626

Reitsma, J. B., Glas, A. S., Rutjes, A. W. S., Scholten, R. J. P. M., Bossuyt, P. M., & Zwinderman, A. H. (2005). Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology*, *58*(10), 982-990. https://doi.org/10.1016/j.jclinepi.2005.02.022

Rohan, A., Rafaq, M. S., Hasan, M. J., Asghar, F., Bashir, A. K., & Dottorini, T. (2024). Application of deep learning for livestock behaviour recognition: A systematic literature review. *Computers and Electronics in Agriculture*, *224*, 109115. https://doi.org/10.1016/j.compag.2024.109115

Siachos, N., Neary, J. M., Smith, R. F., & Oikonomou, G. (2024). Automated dairy cattle lameness detection utilizing the power of artificial intelligence; current status quo and future research opportunities. *The Veterinary Journal*, *304*, 106091. https://doi.org/10.1016/j.tvjl.2024.106091

Sounderajah, V., Guni, A., Liu, X., Collins, G. S., Karthikesalingam, A., Markar, S. R., Golub, R. M., Denniston, A. K., Shetty, S., Moher, D., Bossuyt, P. M., Darzi, A., & Ashrafian, H. (2025). The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nature Medicine*, *31*(10), 3283-3289. https://doi.org/10.1038/s41591-025-03953-8

Tedeschi, L. O., Guarnido-Lopez, P., Menendez, H. M., III, & Seo, S. (2026). Advancing precision livestock farming: Integrating artificial intelligence and emerging technologies for sustainable livestock management. *Animal Bioscience*, *39*(4), 250-289. https://doi.org/10.5713/ab.25.0289

Van Hertem, T., Bahr, C., Tello, A. S., Viazzi, S., Steensels, M., Romanini, C. E. B., Lokhorst, C., Maltz, E., Halachmi, I., & Berckmans, D. (2016). Lameness detection in dairy cattle: Single predictor *v*. multivariate analysis of image-based posture processing and behaviour and performance sensing. *Animal*, *10*(9), 1525-1532. https://doi.org/10.1017/S1751731115001457

Velasco-Bolaños, J., Ceballes-Serrano, C. C., Velásquez-Mejía, D., Riaño-Rojas, J. C., Giraldo, C. E., Carmona, J. U., & Ceballos-Márquez, A. (2021). Application of udder surface temperature by infrared thermography for diagnosis of subclinical mastitis in Holstein cows located in tropical highlands. *Journal of Dairy Science*, *104*(9), 10310-10323. https://doi.org/10.3168/jds.2020-19894

Volkmann, N., Kulig, B., Hoppe, S., Stracke, J., Hensel, O., & Kemper, N. (2021). On-farm detection of claw lesions in dairy cows based on acoustic analyses and machine learning. *Journal of Dairy Science*, *104*(5), 5921-5931. https://doi.org/10.3168/jds.2020-19206

Wang, Y., Kang, X., He, Z., Feng, Y., & Liu, G. (2022). Accurate detection of dairy cow mastitis with deep learning technology: A new and comprehensive detection method based on infrared thermal images. *Animal*, *16*(9), 100646. https://doi.org/10.1016/j.animal.2022.100646

Zecconi, A., Zaghen, F., Meroni, G., Sommariva, F., Ferrari, S., & Sora, V. (2025). Machine learning approach for early lactation mastitis diagnosis using total and differential somatic cell counts. *Animals*, *15*(8), 1125. https://doi.org/10.3390/ani15081125

Zhang, L., Guo, W., Lv, C., Guo, M., Yang, M., Fu, Q., & Liu, X. (2024). Advancements in artificial intelligence technology for improving animal welfare: Current applications and research progress. *Animal Research and One Health*, *2*(1), 93-109. https://doi.org/10.1002/aro2.44

Zhang, X., Li, Y., Zhang, Y., Yao, Z., Zou, W., Nie, P., & Yang, L. (2024). A new method to detect buffalo mastitis using udder ultrasonography based on deep learning network. *Animals*, *14*(5), 707. https://doi.org/10.3390/ani14050707

Zhou, X., Qu, Y., Xu, C., Wang, H., Lang, D., Jia, B., & Jiang, N. (2026). Machine learning and SHapley Additive exPlanation-Based interpretation for predicting mastitis in dairy cows. *Animals*, *16*(2), 204. https://doi.org/10.3390/ani16020204