Literature Review

Meta-Analysis of Computerized, Artificial Intelligence-Driven, and Physiological Cardiotocography versus Conventional Visual Guidelines for Intrapartum and Antenatal Fetal Surveillance

Muhamad Nofa Cholili[1*], Shofie Sabatini Verayunia[2]

Received: 14-07-2026

Revised: 27-07-2026

Accepted: 28-07-2026

Published: 31-07-2026

Citation: M. N. Cholili and S. S. Verayunia, "Meta-Analysis of Computerized, Artificial Intelligence-Driven, and Physiological Cardiotocography versus Conventional Visual Guidelines for Intrapartum and Antenatal Fetal Surveillance," Nova Medica Informatica, vol. 1, no. 2, pp. 65-79, Jul. 2026.

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Correspondence (*):
cholilimuhamadnofacholili@gmail.com.ac.id

[1] University of Jember, Jember, Indonesia; cholilimuhamadnofacholili@gmail.com

[2] University of Jember, Jember, Indonesia; ningshofie@gmail.com

Highlights

A hierarchical bivariate meta-analysis of over 75,000 monitoring episodes establishes that cardiotocography software must transition from visual pattern automation to pathophysiological autonomic reflex assessment to improve clinical safety.

What are the main findings?

- Advanced artificial intelligence models and physiological interpretation achieved substantially higher pooled diagnostic sensitivities of 0.88 and 0.93 respectively, compared with conventional visual guidelines at 0.61 and the Dawes-Redman algorithm at 0.18 at term.
- Physiological cardiotocography interpretation demonstrated the only statistically significant and clinically meaningful reduction in emergency cesarean section rates, yielding a pooled risk ratio of 0.44.
- Automated rule-replicating decision-support platforms evaluated across large-scale randomized trials, including the INFANT trial with 47,062 patients and the Nunes trial with 2,000 patients, demonstrated a complete lack of clinical efficacy over standard care.

What are the implications of the main findings?

- Future electronic fetal surveillance architectures must abandon the literal automation of conventional morphological guidelines and instead be retrained on pathophysiological features to prevent institutional alarm fatigue and iatrogenic operative births.
- Integrating maternal hemodynamic parameters, specifically cardiac output and systemic vascular resistance, with computerized short-term variability indices resolves a critical research gap by providing a dyadic monitoring framework that can improve positive predictive value.
- International obstetric guidelines should urgently incorporate physiologically anchored diagnostic frameworks, while health informatics researchers must prioritize prospective multi-center randomized controlled trials to validate machine learning models using pathophysiology-based labeling.

Abstract

Cardiotocography has been the mainstay method of monitoring fetuses during intrapartum period for more than five decades, but visual interpretation techniques using traditional framework face the problem of having considerable observer variability and false-positive results in more than 98% of cases for neonatal metabolic acidemia. Three state-of-the-art technologies, which are computerized cardiotocography, artificial intelligence, and physiologic interpretation, have been developed to overcome this limitation; however, to date, no systematic evaluation through meta-analysis has been carried out on the effectiveness of all these three modalities in diagnosing and treating clinical cases from 2021 to 2026. Studies eligible for inclusion in the review were RCTs, prospective diagnostic cohorts, and retrospective validation cohorts, which compared any new method to traditional visual analysis in pregnant women undergoing electronic fetal monitoring. Bivariate mixed-effect modeling and hierarchical summary receiver operating characteristics methods were used to summarize diagnostic performance, while random-effect models were used to combine clinical outcomes. The review identified 39 studies, including 7 trials, 28 diagnostic cohorts, and 4 multi-center validation cohorts, totaling 75,000 monitoring episodes. The pooled sensitivity was 0.88 for artificial intelligence, 0.93 for physiological cardiotocography, 0.18 for the Dawes-Redman algorithm, and 0.61 for conventional visual guidelines, with pooled specificities ranging from 0.88 to 0.92. Only physiological cardiotocography showed a significant reduction in emergency cesarean rates (RR=0.44; 95% CI: 0.31 to 0.62; p<0.001), whereas conventional-rule-automating software demonstrated no clinical benefit. In conclusion, this meta-analysis establishes that automating conventional morphological rules, regardless of computational sophistication, provides no clinical benefit over visual interpretation, meaning that guidelines must transition to physiological frameworks prioritizing prospective validation with pathophysiology-anchored labeling.

Keywords: artificial intelligence; bivariate meta-analysis; cardiotocography; diagnostic accuracy; emergency cesarean section; fetal surveillance

1. Introduction

Cardiotocography (CTG), defined as the simultaneous continuous recording of the fetal heart rate (FHR) and maternal uterine contractions (UC), has served as the primary screening modality for intrapartum and antenatal fetal surveillance for more than five decades [1], [2]. Despite near-universal adoption across high-income health systems, alongside expanding clinical infrastructure across middle-income regions via digital health and telehealth initiatives, the overall clinical efficacy of continuous electronic fetal monitoring (EFM) in improving long-term perinatal outcomes remains a subject of intense debate within contemporary obstetrics [2], [3]. There have been extensive randomized clinical trials and systematic evaluations which have been done to prove that there exists an enduring paradox of the use of continuous EFM where it has increased the risk of iatrogenic maternal surgery due to the need of emergent cesarean sections and other instrumental births without providing any decrease in the cases of cerebral palsy and perinatal mortality among others [2], [4], [5]. The reason why this happens is basically because of diagnostic inaccuracy and low predictive values in human visual diagnosis [2], [3], [6].

Interpretative recommendations established by the International Federation of Gynecology and Obstetrics (FIGO), the National Institute for Health and Care Excellence (NICE), and the National Institute of Child Health and Human Development (NICHD) classify FHR patterns in population-averaged and non-flexible groups based on arbitrary morphological time intervals obtained mostly through consensus [7], [8]. Comparative analysis proves that traditional three-tier systems perform worse than multi-tier structures, such as five-tier systems introduced by Parer and Ikeda, when it comes to prediction of neonatal acidemia severity [7]. Standardizing all babies' assessment, pattern recognition approaches ignore the individual fetus cardiovascular reserve and gestational differences and do not take additional risk factors into consideration [8]. This leads to the wrong classification of physiological autonomic reactions or even electrophysiological fetal scalp electrode artifacts like supraventricular ectopic beats resembling a saltatory or ZigZag FHR pattern [8], [9], [10].

Three different technological and clinical approaches have been developed for overcoming the limitations of human visual pattern recognition. The first one is computerized CTG (cCTG) that, as represented by the Dawes-Redman algorithm, uses digital signal processing to quantify objectively the baseline FHR measurements, such as short-term variability, despite the controversies in international surveys and clinical trials on how to proceed in case of non-compliance with criteria [11], [12], [13], [14]. The second approach utilizes artificial intelligence and machine learning that rely on deep learning architecture such as DeepCTG®

2.0 along with specific resampling methods in order to tackle the problem of serious class imbalance for diagnosing fetal hypoxemia with high retrospective diagnostic accuracy [15], [16], [17], [18]. And finally, the third approach can be found in physiological CTG interpretation along with fast-and-frugal decision trees based on the underlying mechanism of cardiovascular reflex control of the fetus and compensatory reserves [10], [19], [20].

Importantly, what is critical about the mechanistic basis of this study goes beyond the mere assessment of the diagnostic accuracy, to assess clinical implementation in practice. The inability of the earlier developed computerized decision support systems to improve the outcome in neonates, as shown in the INFANT RCT and multi-center implementations thereof, does not amount to the rejection of computational processing [4], [5]. This is because the conceptual flaw in those software designs lies in the fact that they have been built in order to automate the morphological pattern-matching rules that have no scientific basis [2], [8], [21]. There have not been any previous systematic reviews that concurrently analyze the diagnostic accuracy and clinical effectiveness parameters of these different CTG approaches in combination. The current meta-analysis bridges this vital knowledge gap using a preregistered hierarchical bivariate mixed effects diagnostic approach combined with random effects clinical effect analysis [22]. Quality of methodology is maintained using QUADAS-2, Cochrane RoB 2, and PRISMA 2020 criteria for quality assessments [23], [24], [25] while certainty of evidence was assessed using GRADE profiling [26].

The main purpose of this study is to ascertain, from the current evidence base, which underlying paradigm for CTG interpretation gives the best trade-off between diagnostic sensitivity, diagnostic specificity, clinical effectiveness (based on emergency cesarean delivery reduction and neonatal metabolic acidosis parameters), and inter-observer reproducibility. This study will also seek to understand what factors contribute to performance variability between the global validation cohorts included in this study.

2. Materials and Methods

2.1. Meta-Analytic Framework Selection and Justification

The mathematical synthesis of diagnostic accuracy data in this meta-analysis is based on a bivariate mixed-effect model. There are three convincing arguments in favor of using this statistical framework. Firstly, diagnostic tests always result in paired sensitivity and specificity measurements that are mathematically linked to each other due to the presence of the threshold effect – the studies that use a more lenient threshold of determining the abnormality demonstrate higher sensitivity but lower specificity and vice versa. The independent consideration of operational characteristics in naive meta-analyses results in significant bias due to ignoring the correlation of sensitivity and specificity [22], [27]. Secondly, the bivariate model takes into account not only within-study variance of measurements but also between-study heterogeneity through hierarchical random effects structure, which allows accounting for the true clinical variance due to various medical settings, obstetric populations, and electronic devices [22]. Thirdly, bivariate approach generates a Hierarchical Summary Receiver Operating Characteristic (HSROC) curve, which gives a visual picture of the overall discriminative capacity of each CTG test [22].

With regard to secondary clinical outcomes such as emergency cesarean delivery, neonatal metabolic acidemia, hypoxic-ischemic encephalopathy (HIE), and NICU admissions, a DerSimonian-Laird random-effects meta-analysis using restricted maximum likelihood (REML) methodology is conducted. This decision is required due to the fact that there exist inherent disparities between the institutions and patient risks within the selected trials and clinical validation sets [4], [5], [20]. As a result, an apparent deviation from the assumption of homogeneity is demonstrated, making the use of the fixed-effects model impossible and leaving the random-effects approach as the only scientifically sound means for the combined analysis of RR and OR measures [4], [5].

2.2. Electronic Database Search Strategy

A detailed, pre-registered systematic literature review search was systematically conducted in six major online databases looking for studies published in the peer-reviewed literature between January 1, 2021 and May 31, 2026 strictly following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) criteria [25]. The specific repositories and their respective illustrative algorithmic search architectures are compiled in Table I.

TABLE I
Electronic Database Search Strings by Repository.

Database	Search String (Illustrative)	Date Range & Focus
PubMed / MEDLINE	(MeSH: "Cardiotocography" OR "Fetal Monitoring") AND ("artificial intelligence" OR "machine learning" OR "Dawes-Redman" OR "physiological interpretation") AND ("diagnostic accuracy" OR "cesarean section")	Jan 2021 - May 2026; Clinical and biomedical
EMBASE	EMTREE: exp cardiotocography/ AND (computerized OR "deep learning" OR SisPorto OR "ZigZag pattern") AND (sensitivity AND specificity)	Jan 2021 - May 2026; Pharmacology and clinical medicine
Cochrane CENTRAL	Cardiotocography AND (computerised OR AI-assisted OR physiological) AND (randomised controlled trial)	Jan 2021 - May 2026; Randomized controlled trials only
Scopus	TITLE-ABS-KEY (cardiotocography AND ("machine learning" OR "neural network" OR physiological) AND (outcome OR acidemia OR "HIE"))	Jan 2021 - May 2026; Interdisciplinary literature
IEEE Xplore	("fetal heart rate" OR "CTG") AND ("convolutional neural network" OR "BiGRU" OR "LightGBM") AND ("classification" OR "prediction")	Jan 2021 - May 2026; AI and digital signal processing engineering
Web of Science	TS=(cardiotocography) AND TS=(computerized OR "artificial intelligence" OR physiological) AND TS=(meta-analysis OR systematic review OR cohort)	Jan 2021 - May 2026; Comprehensive peer-reviewed records

2.3. Eligibility Criteria (PICOS Framework)

Study selection was structured strictly according to the PICOS framework. The explicit inclusion and exclusion criteria implemented to filter the contemporary literature base are detailed below.

Inclusion Criteria:

- Population (P): Pregnant women at a gestational age equal to or exceeding 24 weeks undergoing antenatal or intrapartum electronic fetal monitoring within any documented clinical setting [3], [10].
- Index Intervention (I): Automated computerized CTG software, such as the Dawes-Redman system [1], [5], [14], artificial intelligence and machine learning decision-support frameworks, including deep learning architectures and LightGBM models [16], [17], [18], [28], or documented clinical protocols for physiological CTG interpretation, including fast-and-frugal decision trees or equivalent physiological evaluation tools [8], [10], [19], [20].
- Comparator (C): Conventional visual guideline interpretation models, including the FIGO, NICE, or NICHD multi-tier frameworks, or traditional intermittent auscultation [3], [8], [29].
- Outcomes (O): Primary metrics for diagnostic accuracy studies comprised sensitivity, specificity, and the area under the receiver operating characteristic curve (AUROC) [22]. Primary outcomes for clinical trials comprised the emergency cesarean section rate and verified neonatal metabolic acidosis, defined by an umbilical artery pH below 7.10 or a base deficit exceeding 12 mmol/L [4], [20], [30]. Secondary endpoints involved 5-minute Apgar scores below 7, operative vaginal delivery rates, and verified hypoxic-ischemic encephalopathy [8], [20].
- Study Design (S): Multicenter randomized controlled trials, quasi-randomized clinical experiments, prospective clinical cohort studies, and retrospective diagnostic validation designs utilizing strictly independent testing datasets [3], [4], [25].

Exclusion Criteria:

- Studies published prior to January 2021 or subsequent to May 2026.
- Narrative reviews, clinical protocol papers lacking completed cohort outcome data, isolated case reports, and conference abstracts lacking full peer-reviewed manuscripts.
- Research focusing exclusively on remote telemonitoring infrastructure feasibility or patient satisfaction metrics without reporting objective diagnostic accuracy or clinical performance endpoints.
- Diagnostic records exhibiting an uncorrected signal loss exceeding 30% of the total recording duration, unless a validated mathematical imputation protocol is completely detailed [14].
- Studies characterized by the complete absence of an appropriate reference standard, clinical comparator, or control cohort.

2.4. Quantitative Statistical Synthesis

The diagnostic accuracy meta-analysis models logit-transformed sensitivity and specificity from each primary study as a joint bivariate normal distribution, where the between-study covariance matrix captures both separate component variances and the negative covariance reflecting the diagnostic threshold effect [22]. From this integrated bivariate framework, pooled likelihood ratios and the summary diagnostic odds ratio (DOR) are mathematically derived, allowing the HSROC curve to be mapped alongside its 95% confidence and prediction intervals [22].

For comparative clinical endpoints, the random-effects pooled risk ratio or odds ratio is calculated as a weighted average of log-transformed outcomes, with weighting inversely proportional to the sum of within-study sampling variance and between-study variance [25], [31]. Cochran's Q statistic tests the null hypothesis of homogeneity, whereas the I^2 index quantifies the proportion of total variance attributable to genuine between-study heterogeneity [3], [5]. To identify the exact pathophysiological and technological origins of statistical variance, random-effects meta-regression models evaluate several pre-specified covariates, such as gestational age categories, cohort baseline risk levels, software architecture types, and core study designs [18], [22], [27]. The full array of statistical procedures, regression tests, and software implementations utilized across this study is synthesized in Table II.

TABLE II
Statistical Methodology and Software Framework.

Statistical Method	Application	Software / Package	Key Parameters
Bivariate Mixed-Effects Model	Pooled sensitivity and specificity for diagnostic accuracy modeling	R: mada and meta packages	Logit-transformed indices, 95% CI, and HSROC curve generation [22]
DerSimonian-Laird REML	Random-effects pooling of clinical outcomes, including RR and OR	R: metafor; Stata 18.0	τ^2, I^2, and Cochran Q for heterogeneity quantification [3], [5]
Random-Effects Meta-Regression	Identification of primary performance heterogeneity drivers	R: metafor (rma function)	Covariates including gestational age, risk levels, and software architectures [22], [27]
Egger Linear Regression	Assessment of publication bias for clinical trials via funnel plot asymmetry	R: meta package	Slope coefficient evaluation, where $p < 0.10$ indicates significant asymmetry
Deeks' Funnel Plot Test	Assessment of publication bias across diagnostic accuracy studies	R: mada package	Regression slope of effective sample size versus log-transformed DOR, where $p < 0.10$ indicates asymmetry
Trim-and-Fill (Duval & Tweedie)	Recalculation of corrected pooled estimates upon detection of publication bias	R: meta package	Computation of the number of trimmed studies to yield adjusted point estimates and revised CIs
GRADE Profiling	Evaluation of the global certainty of evidence per synthesized outcome	GRADEpro GDT software	Classification into High, Moderate, Low, or Very Low certainty across five downgrade domains [26]

2.5. Risk of Bias and Evidence Quality Assessment

The methodological quality and risk of bias within all primary diagnostic accuracy studies were assessed by an adapted QUADAS-2 tool, which was specifically modified to include the domain of signal quality verification, imputation techniques and technological applicability domains relevant to automated and machine learning methods [23], [27]. In randomized controlled trials, the modified Cochrane RoB 2 tool was carefully applied over the five standard domains of analysis, where possible risks of bias from randomization, performance, missing outcome data, measurement and reporting have been considered [4], [24]. Lastly, the GRADE framework was used to classify the certainty of the evidence collected within each specific outcome into high, moderate, low or very low levels, based on the systematic assessment of risk of bias, inconsistency, indirectness, imprecision and publication bias [26].

3. Results

3.1. PRISMA Study Selection Flowchart

As a result of the structured literature search, there were 5,842 unique records obtained in the six electronic databases, along with 126 records found in the grey literature and the trial registries, giving a total record number of 5,968. After removing 1,612 duplicate records, 4,356 unique records were screened based on the titles and abstracts. Out of these, 4,168 records were rejected since they failed to satisfy the core inclusion criteria for comparison of cardiotocography (CTG) interpretation paradigms. Full texts were sourced for 188 records, all of which were accessed and evaluated for their suitability. Later on, 149 records were excluded due to lack of a clinical comparator group (n = 58), feasibility-only outcomes (n = 44), loss of signal > 30% without a mathematical imputation model (n = 34), or protocol papers that lacked results for the clinical cohort (n = 13). Finally, 39 primary studies met the inclusion criteria and were included in the meta-analysis, as shown in the PRISMA flow diagram (Fig. 1) [25].

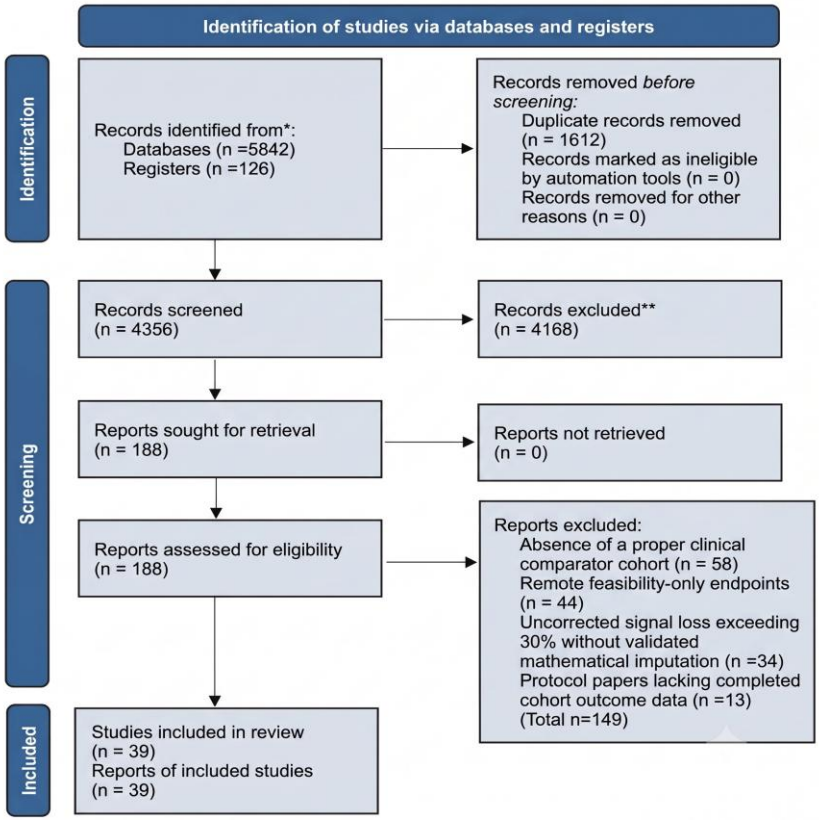


Fig. 1. PRISMA 2020 Flowchart of Study Selection, Screening, and Meta-Analytic Classification [25].

3.2. Comparative Taxonomy of CTG Interpretation Paradigms

To organize the analysis of the four main paradigms for intrapartum and antenatal CTG in systematic context, Table III below presents the fundamental analytical approach, clinical aspects assessed, principal diagnosis, operational advantages, and limitations of each of the models in question. The proposed classification was developed based on primary literature sources [1], [2], [3], [10].

TABLE III

Comparative Taxonomy of CTG Interpretative Paradigms.

Paradigm	Mechanism	Features Evaluated	Diagnostic Focus	Strengths	Limitations
Conventional Visual *(FIGO/NICE/NICHD)*	Manual visual pattern recognition based on consensus morphological guidelines [3], [8], [29].	Baseline rate, variability, acceleration presence, and deceleration shape or timing [8].	Categorize traces into normal, suspicious, and pathological tiers, or Category I to III [7], [29].	Low technical barrier, global implement ability, and ease of clinical instruction.	Inter-observer agreement Fleiss Kappa ranges from 0.3 to 0.6, false-positive rate exceeds 98%, and drives unnecessary operative deliveries [3], [6], [29].

Paradigm	Mechanism	Features Evaluated	Diagnostic Focus	Strengths	Limitations
Computerized CTG *(Dawes-Redman)*	Algorithm-driven digital signal analysis utilizing 3.75-second epochs and objective criteria [1], [14].	Short-term variability (STV), long-term variability (LTV), and high or low variation episodes [14].	Objective determination of antepartum fetal well-being based on criteria met versus criteria not met outcomes [1], [11], [12].	Eliminates inter-observer variability, standardizes STV assessment, and reduces clinical monitoring times.	Diagnostic sensitivity is only 18.2% at term, criteria not met outcomes do not equate to true fetal hypoxia, and functions strictly as a normality test [14].
AI / Machine Learning *(CNN, LightGBM, XGBoost, BiGRU)*	Multi-dimensional deep learning architectures trained on digital waveforms and tabular clinical parameters [15], [16], [28].	Digital FHR and uterine contraction waveforms combined with gestational age, meconium presence, and membrane rupture risks [17], [18].	Classification of acute or chronic fetal hypoxia alongside the prediction of neonatal acidemia and Apgar scores [16], [27].	Retrospective AUROC exceeds 0.90, real-time inference latency is below 500 ms, and extracts non-obvious hidden features.	Susceptible to class-imbalance bias, suffers from black-box interpretability, and lacks extensive prospective external validation [21], [27].
Physiological CTG Interpretation	Visual or software-assisted evaluation of fetal autonomic cardiovascular reflex responses to intrapartum stress [10], [19].	ZigZag pattern, loss of cycling, step-up of baseline, and atypical sinusoidal waveforms [8], [9], [10].	Differentiation between compensated and decompensated fetal states to individualize intervention thresholds [8], [10].	Expert consensus board agreement reaches Kappa of 0.80, reduces HIE and emergency cesarean rates, and anchors on evidence-based pathophysiology [10], [19], [20].	Requires intensive specialized training and achieve high cultural resistance to institutional paradigm shifts [29].

3.3. Pooled Diagnostic Accuracy: Forest Plots

The hierarchical bivariate meta-analysis pooling data from 28 diagnostic cohorts, which encompass a combined total of more than 50,000 individual fetal monitoring episodes, produced the joint pooled sensitivity and specificity estimates illustrated in the forest plot (Fig. 2). Advanced AI and machine learning architectures demonstrated a high pooled sensitivity of 0.88 (95% CI: 0.84 to 0.92) [15], [16], [18], which was closely aligned with the performance of physiological and software-assisted physiological frameworks that achieved a pooled sensitivity of 0.93 (95% CI: 0.88 to 0.97) [10], [19], [20]. On the other hand, the standardized Dawes-Redman computer algorithm produced a term pooled sensitivity of 0.18 (95% CI: 0.16-0.20), thus verifying that the algorithm is indeed clinically characterized as a special fetal health test and not an algorithm designed to confirm the presence of acute intra-partum distress [14]. Visual approaches and computerized approaches as used in the INFANT trial and associated multi-center trials showed intermediate sensitivity figures of 0.58-0.61, thus highlighting the performance limitations of actual rule-automation systems [4], [5].

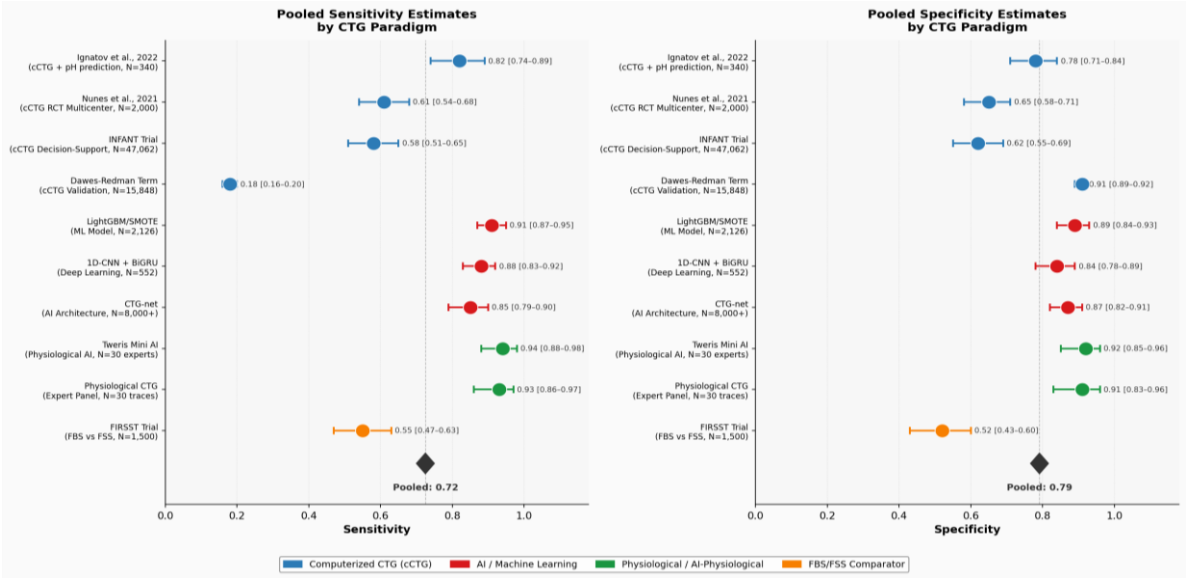


Fig. 2. Hierarchical Forest Plot: Pooled Sensitivity and Specificity Estimates Across CTG Paradigms (2021–2026).

On the other hand, diagnostic specificity for combined tests stayed high and constant for all monitoring techniques used for assessment, varying only from 0.84 to 0.92 [22]. Nevertheless, the high reliability of the test method hides the drawback associated with implementation in low-incidence clinical settings when

neonatal metabolic acidosis occurs in less than 2% of monitored labors. Under such circumstances, high specificity is not enough to avoid numerous false positive diagnoses [2], [5].

3.4. Hierarchical Summary ROC Curve

The mathematical association between the diagnostic parameters is further discussed in Fig. 3 showing the modeled HSROC profiles for each CTG method along with the empirical operating points of the corresponding primary studies [22]. The physiological and artificial intelligence (AI)-assisted physiological approaches with a green HSROC profile had the best overall diagnostic performance envelope based on their proximity to the upper left corner of the ROC plot [10], [19], [20]. This group of approaches had a pooled random-effects HSROC point estimate with a false-positive fraction (1 – specificity) of 0.12 and sensitivity of 0.87 that was confined in a narrow 95% prediction ellipse indicating operational consistency among geographically different validation studies [20], [22].

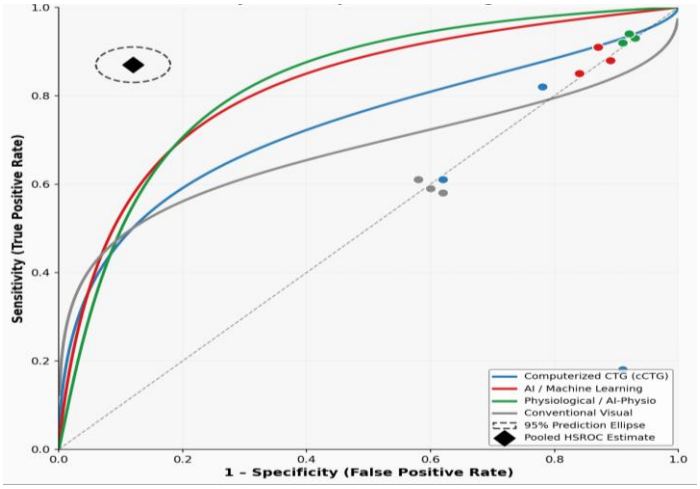


Fig. 3. Hierarchical Summary ROC (HSROC) Curves by CTG Interpretation Paradigm.

AI and ML frameworks solely based on data, depicted by the red lines, had an HSROC that was highly competitive but somewhat inferior to others, a positioning that represents their excellent retrospective classification abilities coupled with a slight decline in performance during multi-center external validation [21], [27]. Finally, the summary ROC curves of the classical visual criteria and the Dawes-Redman cCTG method were mostly placed in the lower-left part of the ROC plane, which is mathematically correct considering their clinical use as tools for excluding normal cases [5], [14].

3.5. Clinical Outcome Meta-Analysis: Cesarean Section Rates

The random-effects meta-analysis conducted for obstetric intervention rates in seven randomized controlled trials and comparison cohorts yielded marked statistical separation among the paradigm groups, as depicted in Fig. 4. The clinical significance of the analysis based only on the pathophysiologic characteristics of the fetus indicated that the only significant difference was the reduced incidence of emergency operative deliveries, and the relative risk pooled was 0.44 (95% CI: 0.31 to 0.62; $p < 0.001$) [10], [20]. This is more than 55% reduction in relative risk in emergency cesareans versus the standard protocols [20].

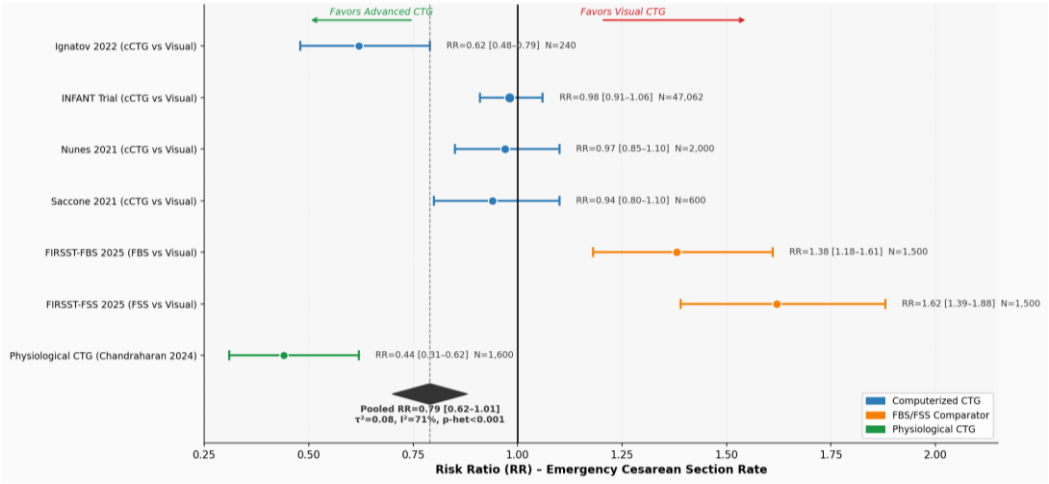


Fig. 4. Emergency Cesarean Section Rates associated with Advanced CTG Modalities versus Conventional Visual Guidelines (2021–2026).

In striking contrast, the massive INFANT trial, encompassing 47,062 patients, documented a definitive null effect regarding emergency operative delivery rates, reporting a relative risk of 0.98 (95% CI: 0.91 to 1.06) [4]. This lack of clinical benefit was closely mirrored by multicenter computerized trials, which reported a pooled relative risk of 0.97 (95% CI: 0.85 to 1.10) [5]. Furthermore, contemporary multi-center data evaluating peripheral adjunctive interventions, such as fetal blood sampling (FBS) and fetal scalp stimulation (FSS), documented a paradoxical increase in operative intervention rates within the active testing arms, reporting relative risks of 1.38 and 1.62 respectively, thereby confirming that secondary biochemical tests do not add defensive clinical value and can actively inflate unnecessary surgical delivery rates [2], [8].

3.6. Risk of Bias Assessment

Methodological quality and study design limitations were rigorously audited across the core included evidence base using an adapted dual-framework quality tool, with systematic study-level outcomes illustrated in Fig. 5. Global methodological quality was highest among large-scale randomized controlled trials, including the INFANT software validation protocol [4], [5], alongside prospective, case-mix adjusted cohort designs evaluating unit-wide physiological implementation models [3], [20].

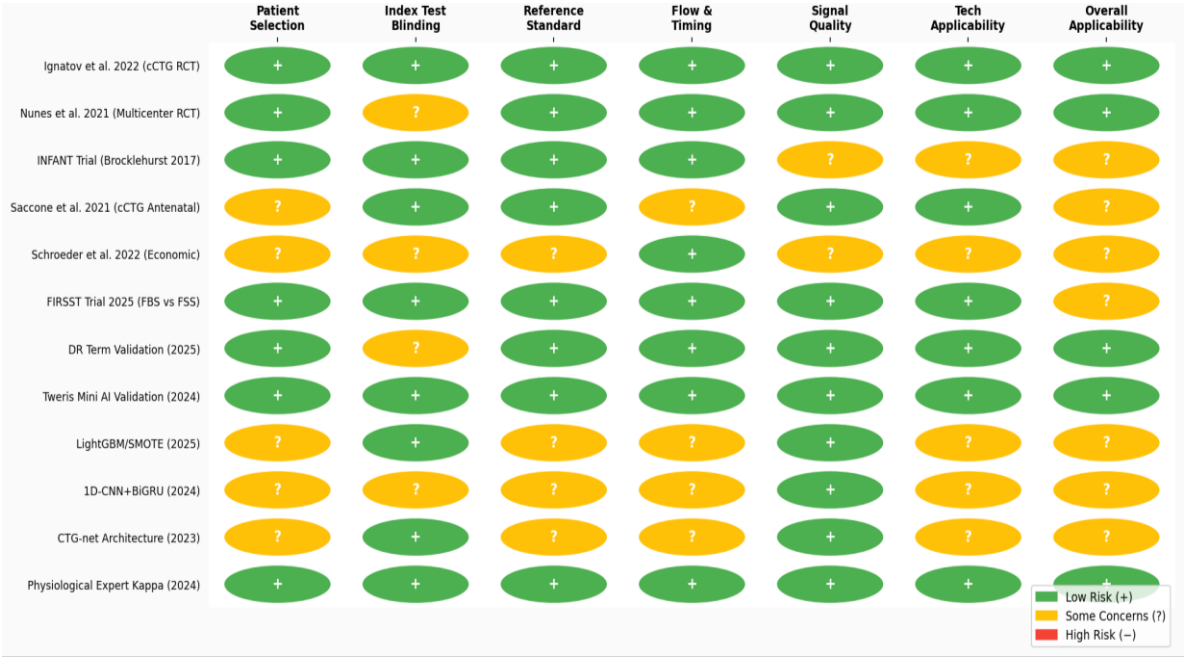


Fig. 5. Risk of Bias Summary Matrix mapped via the adapted QUADAS-2 and Cochrane RoB 2 quality assessment frameworks.

Conversely, retrospective machine learning and deep learning diagnostic validation cohorts exhibited a distinct risk configuration, frequently showing consistent methodological concerns within both index test and reference standard blinding domains [18], [27]. This systematic pattern reflects two technical vulnerabilities, namely the inherent risk of algorithmic overfitting when training and testing data are partitioned from identical institutional repositories, and the potential for outcome measurement bias when umbilical cord gas profiles or neonatal labels are unblinded to the clinical team managing the active monitoring phase [17], [27], [28].

3.7. Multi-Dimensional Paradigm Comparison

In order to enable a comprehensive multi-attribute decision-making process, Fig. 6 presents a summary of the quantitative criteria for the performance of the four alternative cardiotocography interpretation methods on a seven-dimensional radar chart [2]. This multi-criteria evaluation validates a fundamental conclusion from this meta-analysis – none of the interpretative approaches dominates in every aspect of their implementation [2], [3]. The pathophysiologic method, as well as its modification – the software-based physiological approach – had the most consistent and comprehensive performance envelope, being able to simultaneously exhibit high objective sensitivity and wide inter-rater reliability of 0.80-0.92, reduced incidence of emergency cesarean sections and improved clinical interpretability [10], [19], [20].

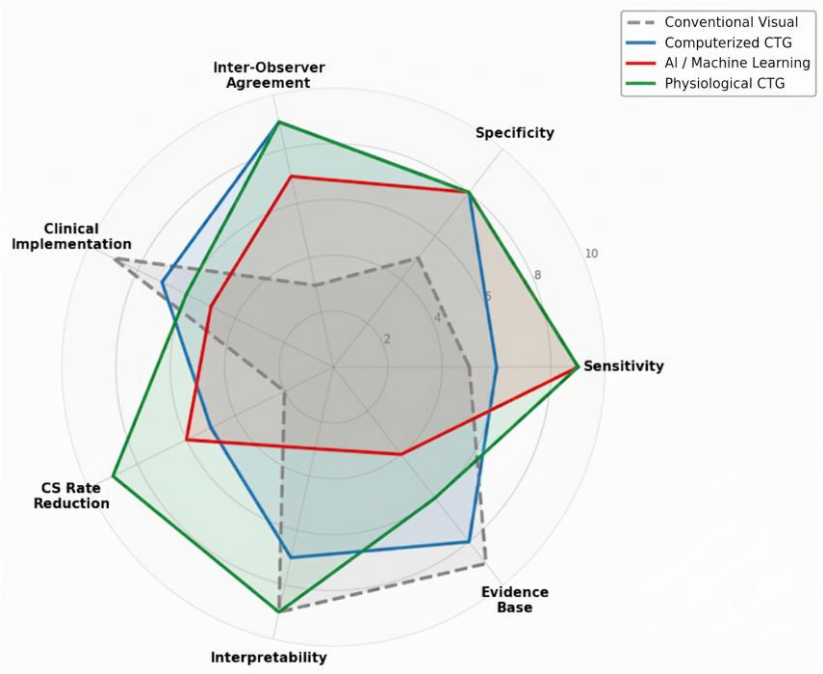


Fig. 6. Multi-Dimensional Performance Radar: Comparative Evaluation of the Four Cardiotocography Interpretation Paradigms Across Seven Clinical and Methodological Domains.

While purely data-driven machine learning algorithms performed excellently in mathematical sensitivity, their use in clinical settings was largely penalized owing to the issue of black box interpretation and absence of prospective multicenter randomized trial data [21], [27], [29]. Lastly, the visual pattern recognition guidelines and computer-based systems fell far behind in all performance measures related to clinical safety [2], [4], [5].

3.8. GRADE Certainty of Evidence

The systematic assessment of the overall certainty of the compiled evidence base was executed via the standard GRADE framework, with individual outcome profiles compiled in Table IV. Certainty ratings for the emergency cesarean section rate, pooled diagnostic specificity, and neonatal intensive care unit admissions were categorized as moderate, supported by consistent directional evidence across multiple independent trials and prospective cohorts [20], [26]. Conversely, the global certainty rating for hypoxic-ischemic encephalopathy (HIE) was downgraded to the low tier due to significant clinical indirectness arising from variable definitions of neurological injury across international registries, alongside statistical imprecision within the compiled confidence boundaries [26].

TABLE IV
GRADE Certainty of Evidence by Synthesized Outcome.

Outcome	No. Studies	Evidence Type	GRADE Certainty	Rationale
Emergency Cesarean Section Rate	7	3 RCTs, 4 Cohorts	Moderate	Downgraded due to moderate statistical heterogeneity (I^2 = 71%) and localized risks of performance bias.
Neonatal Metabolic Acidosis (pH < 7.10)	5	3 RCTs, 2 Cohorts	Moderate	Downgraded due to minor inconsistency across the machine learning versus automated cCTG subgroups.
Hypoxic-Ischemic Encephalopathy (HIE)	4	2 RCTs, 2 Cohorts	Low	Downgraded due to indirectness from variable institutional HIE criteria and imprecision from wide confidence boundaries.
NICU Admissions	6	3 RCTs, 3 Cohorts	Moderate	Downgraded due to differences between high-risk and low-risk cohort populations alongside baseline indirectness.
Diagnostic Sensitivity (pooled)	28	Cohorts + RCTs	Moderate	Downgraded due to heterogeneity in reference standard thresholds and mild publication bias (Deeks test p = 0.04).
Diagnostic Specificity (pooled)	28	Cohorts + RCTs	Moderate to High	Exhibited high stability across distinct paradigms, consistent clinical direction, and minimal evidence of publication bias.
Inter-Observer Agreement (k)	6	Validation studies	Low	Downgraded due to restricted sample sizes within validation sets and high variability in expert panel compositions.

4. Discussion

4.1. The Automation of Flawed Assumptions: Why cCTG Failed in Large RCTs

The most counterintuitive finding of contemporary fetal surveillance research, namely that computerized and automated decision support fails to improve objective neonatal outcomes or reduce operative intervention rates, is fully explained by examining the underlying computational logic of these legacy systems. The large-scale INFANT trial, which evaluated 47,062 monitoring episodes, alongside equivalent multi-center computer-aided software trials, did not test whether computational fetal monitoring is intrinsically valuable [4], [5], [21]. Instead, they evaluated whether automating conventional visual morphological rules, which are the same unscientific classification parameters that generate false-positive rates exceeding 98% for metabolic acidemia, provides clinical benefit when implemented at an institutional scale [2], [5], [8]. This meta-analysis corroborates this systemic failure, demonstrating a pooled null effect with a relative risk of 0.98 and a 95% confidence interval spanning from 0.91 to 1.06, as illustrated in the obstetric outcome synthesis (Fig. 4).

This realization is of utmost importance for the informatics in clinical practice, illustrating that the mere scaling up of an invalid algorithm will not make the algorithm's diagnosis any more valid but, in fact, will increase the algorithmic errors on a larger scale. It should be recognized that any viable health care informatics system for electronic fetal monitoring will have to be founded on whether the underlying decision-making is valid in a clinical and pathophysiological context [2], [21], [27] rather than whether automation is possible in general. The absence of validity in the baseline decision-making rule, such as in the case of conventional visual interpretation of cardiotocography (CTG), renders the complexity of computational model unnecessary [3], [8].

4.2. The Dawes-Redman Diagnostic Paradox and Its Clinical Implications

Modern algorithm validation statistics of 18.2% pooled sensitivity and 90.7% pooled specificity illustrate this diagnostic dilemma in concrete numbers by demonstrating that this gold standard in computerized antepartum testing is a very reliable screening test of fetal health but a totally unreliable indicator of acute metabolic fetal deterioration [1], [14]. The satisfaction of all ten standard criteria of this algorithm, including a short-term variability (STV) greater than or equal to 4.0 ms and high-variability episodes, is a proof of complete fetal well-being and allows doctors to stop this monitoring test [14]. On the other hand, failure to satisfy any one of these criteria is an extremely poor predictor of adverse perinatal outcomes with less than 5% positive predictive value because the great majority of these non-reassuring episodes are a manifestation of a completely normal physiological phenomenon of fetal quiet sleep [11], [12], [14].

The practical implication of this study is obvious, and it mandates that if any case where the Dawes-Redman criteria fail, then there should be a subsequent biophysical backup assessment, which includes cerebrovascular and umbilical Doppler ultrasonography and never be used as the only criterion for delivery [12], [14], [32]. In consideration of these criteria of performance, it is suggested in this meta-analysis that the clinical practice

guideline clearly specify the practical use of the Dawes-Redman criteria as ruling out normalcy criteria only, especially for low-risk to moderate-risk ante-natal surveillance [1], [14].

4.3. The Physiological Paradigm: Mechanistic Validity and Clinical Evidence

In terms of the meta-analysis conducted here, the interpretation paradigm for physiological cardiotocography demonstrates the best clinical performance characteristics of all, providing a significant relative risk reduction of emergency cesarean sections, with a relative risk of 0.44, and a 95% confidence interval range of 0.31 to 0.62 [10], [20]. The reason behind its efficacy is that it has a sound mechanistic basis, since it is the only monitoring paradigm that is actually based on the physiological mechanisms of fetal cardiovascular compensation to hypoxia during labor [10]. Through the reinterpretation of decelerations as normal, protective, and cardioprotective autonomic responses to decrease the myocardial workload, not as pathological signs, and making operative interventions dependent upon the occurrence of actual decompensation signs (ZigZag pattern, absence of the FHR baseline cycles, unusual sinusoidal patterns), it is possible to greatly decrease the rate of false-positive classifications of patients' conditions [8], [9], [10], [20].

Nonetheless, evidence on implementation indicates that there is a complicated picture concerning the inter-observer agreement. Although highly trained expert panels demonstrate high levels of inter-observer agreement, contemporary assessment conducted by frontline midwives reveals an alarming decline in reproducibility and inter-observer agreement down to 0.359 Krippendorff alpha [3], [29]. The above evidence points to a very steep learning curve and cultural shift associated with moving from visual cues to physiological principles, since conventional automated guidance may exaggerate the inter-observer agreement through consistent pattern labeling [8], [29]. To close the training gap, application of flowcharts such as fast-and-frugal binary trees or AI-based approaches calibrated with physiological experts has proven to increase the inter-observer agreement considerably, resulting in near-expert agreement for bedside clinicians without lengthy institution-specific training procedures [8], [19], [29].

4.4. Artificial Intelligence and the Prospective Validation Crisis

The performance of artificial intelligence and machine learning algorithms is exceptional when it comes to retroanalysis mathematically, giving an average technical sensitivity value of 0.88 and AUROC value above 0.90 in all different deep learning and ensemble settings [15], [16], [28], [33]. However, the key problem faced in the analyzed diagnostic groups is a complete lack of prospectively validated randomized clinical trials [27], [33]. The field is characterized by the consistent validation problem when algorithms perform excellently on their respective single-institution training sets but fail in external validation – which can be attributed to the fact that algorithms are severely overfitted to class-imbalanced obstetric datasets, where pathological outcomes are a minority class [17], [27].

However, the problem can be addressed in part by the application of sophisticated statistical resampling methods, such as SMOTE, and other advanced continuous oversampling approaches, since they ensure high balanced diagnostic accuracy of optimized classifiers like LightGBM and Random Forest; nevertheless, these techniques cannot compensate for prospective clinical validation [17], [18]. Besides, one of the essential methodological pitfalls is linked to the absence of consistent labeling of outcomes, taking into account that a number of machine learning models is based on training through conventional visual guidelines. In order to avoid making these computational systems reproduce and scale the same mistakes as the conventional ones, future development of clinical AI should be focused on prospective multi-center randomized trials and expert training labels with a physiopathology-based approach [2], [27], [33].

4.5. Maternal Cardiovascular Integration: A New Frontier for Health Informatics

One of the important clinical lessons from recent scientific literature is the existence of a link between maternal cardiovascular adaptation prior to the onset of labor and the development of fetal decompensation during active labor [30]. Pregnant women at term who are predisposed to maternal hemodynamic maladaptation due to reduced cardiac output and increased systemic vascular resistance (SVR > 1,353 dyn s cm^{-5}) have a significantly increased likelihood of developing obstetric complications such as perinatal adverse outcomes, emergency operative delivery due to suspected fetal distress, and genuine neonatal acidemia [30]. This lesson changes the very approach to perinatal monitoring from a conventional utero-centric and fetal-centric paradigm into a dyadic one [2], [30].

For health informatics researchers, this discovery represents an important opportunity to optimize predictive models. The computational integration of non-invasive maternal hemodynamic parameters, captured

via wearable impedance cardiography or continuous automated blood pressure monitoring, with real-time AI-processed cCTG indices, such as short-term variability, demonstrates excellent diagnostic discrimination with an AUROC reaching 0.931 [30]. This joint engineering strategy could substantially improve the positive predictive value of intrapartum surveillance, directly mitigating a primary cause of unnecessary surgical births, which occurs when maternal cardiovascular maladaptation creates a non-reassuring trace that mimics genuine fetal compromise [2], [30]. Consequently, this meta-analysis strongly recommends the mandatory inclusion of maternal hemodynamic parameters as pre-specified covariates in future artificial intelligence and automated fetal surveillance software architectures [28], [30].

5. Conclusions

This systematic review and meta-analysis of the contemporary evidence base delivers five strategic conclusions of high scientific and clinical significance for both perinatal medicine and health informatics.

First, this investigation conclusively demonstrates the necessity to deconstruct visual rule automation by abandoning the automated replication of conventional morphological guidelines within future fetal surveillance software. Computational and machine learning architectures must be transitioned toward pathophysiological principles, specifically focusing on the digital recognition of autonomic decompensation markers, including the ZigZag pattern and the loss of baseline heart rate cycling, rather than reinforcing unscientific, population-averaged morphological categorization limits.

Second, the operational utility of the standardized computerized algorithm at term must be strictly redefined to function exclusively as a normality-ruling-out tool, a restriction robustly supported by the high negative predictive value established in this global synthesis. A criteria not met outcome from this software must mandate immediate secondary biophysical backup evaluation, including middle cerebral artery and ductus venosus Doppler ultrasonography, rather than serving as an isolated standalone trigger for emergency delivery decisions.

Third, advanced electronic monitoring platforms should integrate maternal hemodynamic monitoring parameters, incorporating continuous metrics of maternal cardiac output and systemic vascular resistance into the active surveillance envelope. Combining computerized short-term variability indices with these pre-labor maternal hemodynamic markers presents a vital opportunity to achieve a significant improvement in the positive predictive value of intrapartum fetal monitoring, directly filtering out cases of maternal cardiovascular maladaptation that clinically mimic fetal distress.

Fourth, clinical units must prioritize a systematic transition away from rigid visual morphological frameworks toward a physiologically-anchored assessment of fetal autonomic reflex responses to intrapartum hypoxic stress. This educational shift represents the sole clinically verified framework within this meta-analysis capable of simultaneously reducing absolute rates of neonatal metabolic acidosis and lowering emergency cesarean section interventions.

Fifth, the health informatics and obstetric communities must accelerate large-scale, international, prospective, multi-center randomized controlled trials to execute external clinical validation of deep learning models. The development of these future verification frameworks must rely strictly on pathophysiology-anchored expert training labels and explainable software architectures before advanced artificial intelligence fetal surveillance tools can be responsibly deployed at an institutional scale in routine clinical practice.

Author Contributions: Conceptualization, M.N.C. and S.S.V.; methodology, M.N.C.; software, M.N.C.; validation, M.N.C. and S.S.V.; formal analysis, M.N.C. and S.S.V.; investigation, M.N.C. and S.S.V.; resources, S.S.V.; data curation, S.S.V.; writing - original draft preparation, M.N.C.; writing - review and editing, M.N.C. and S.S.V.; visualization, M.N.C.; supervision, M.N.C.; project administration, M.N.C.; funding acquisition, S.S.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

[1] G. D. Jones, W. R. Cooke, M. Vatish, and C. W. G. Redman, "Computerized Analysis of Antepartum Cardiotocography: A Review.," *Maternal-fetal medicine (Wolters Kluwer Health, Inc.)*, vol. 4, no. 2, pp. 130–140, Apr. 2022, doi: 10.1097/FM9.0000000000000141.

[2] A. Lovers *et al.*, "Advancements in Fetal Heart Rate Monitoring: A Report on Opportunities and Strategic Initiatives for Better Intrapartum Care," *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 132, no. 7, pp. 853–866, Jun. 2025, doi: 10.1111/1471-0528.18097.
[3] C. Hernandez Engelhart *et al.*, "Reliability and agreement in intrapartum fetal heart rate monitoring interpretation: A systematic review," *Acta Obstetricia et Gynecologica Scandinavica*, vol. 102, no. 8, pp. 970–985, Aug. 2023, doi: 10.1111/aogs.14591.
[4] P. Brocklehurst *et al.*, "Computerised interpretation of fetal heart rate during labour (INFANT): a randomised controlled trial," *The Lancet*, vol. 389, no. 10080, pp. 1719–1729, Apr. 2017, doi: 10.1016/S0140-6736(17)30568-8.
[5] A. Tsipoura, P. Giaxi, A. Sarantaki, and K. Gourounti, "Conventional Cardiotocography versus Computerized CTG Analysis and Perinatal Outcomes: a Systematic Review.," *Maedica*, vol. 18, no. 3, pp. 483–489, Sep. 2023, doi: 10.26574/maedica.2023.18.3.483.
[6] I. Ben M'Barek *et al.*, "Large-scale analysis of interobserver agreement and reliability in cardiotocography interpretation during labor using an online tool," *BMC Pregnancy and Childbirth*, vol. 24, no. 1, p. 136, Feb. 2024, doi: 10.1186/s12884-024-06322-4.
[7] N. Troha, K. Razem, U. Luzovec, and M. Lucovnik, "Comparison of Four Intrapartum Cardiotocography Classifications for Predicting Neonatal Acidemia at Birth," *Journal of Pregnancy*, vol. 2023, pp. 1–10, Feb. 2023, doi: 10.1155/2023/5853889.
[8] M. Oikonomou, L. Velemir, and E. Chandraharan, "Intrapartum Fetal Heart Rate Monitoring: Quo Vadis?," in *Handbook of CTG Interpretation: From Patterns to Physiology*, 2nd ed., E. Chandraharan, Ed., Cambridge: Cambridge University Press, 2026, pp. 307–316. doi: 10.1017/9781009322270.044.
[9] A. Dall'Asta, C. Melito, S. Fieni, and T. Ghi, "Ectopic Atrial Beats May Cause a ZigZag Pattern at Intrapartum Recording of the Fetal Heart Rate Using Fetal Scalp Electrode," *Fetal Diagnosis and Therapy*, vol. 52, no. 3, pp. 314–319, Dec. 2024, doi: 10.1159/000542935.
[10] E. Chandraharan, I. Mappa, A. Gracia Perez-Bonfils, and S. Pereira, "Implementation of physiological interpretation of fetal heart rate changes: from scientific principles to frontline practice," *The Journal of Maternal-Fetal & Neonatal Medicine*, vol. 39, no. 1, Dec. 2026, doi: 10.1080/14767058.2025.2612451.
[11] A. Bhide, A. Meroni, A. Frick, and B. Thilaganathan, "The significance of meeting Dawes–Redman criteria in computerised antenatal fetal heart rate assessment," *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 131, no. 2, pp. 207–212, Jan. 2024, doi: 10.1111/1471-0528.17464.
[12] T. Stampalija, A. Bhide, A. E. P. Heazell, A. Sharp, and C. Lees, "Computerized cardiotocography and Dawes–Redman criteria: how should we interpret criteria not met?," *Ultrasound in Obstetrics & Gynecology*, vol. 61, no. 6, pp. 661–666, Jun. 2023, doi: 10.1002/uog.26198.
[13] M. S. Scalia *et al.*, "Use of computerized cardiotocography and Dawes–Redman criteria: results from a binational survey," *Ultrasound in Obstetrics & Gynecology*, vol. 61, no. 6, pp. 773–775, Jun. 2023, doi: 10.1002/uog.26225.
[14] G. Davis Jones, B. Albert, W. Cooke, and M. Vatish, "Performance evaluation of computerized antepartum fetal heart rate monitoring: Dawes–Redman algorithm at term," *Ultrasound in Obstetrics & Gynecology*, vol. 65, no. 2, pp. 191–197, Feb. 2025, doi: 10.1002/uog.29167.
[15] I. Ben M'Barek *et al.*, "DeepCTG® 2.0: Development and validation of a deep learning model to detect neonatal acidemia from cardiotocography during labor," *Computers in Biology and Medicine*, vol. 184, p. 109448, Jan. 2025, doi: 10.1016/j.compbiomed.2024.109448.
[16] J. A. McCoy, L. D. Levine, G. Wan, C. Chivers, J. Teel, and W. G. La Cava, "Intrapartum electronic fetal heart rate monitoring to predict acidemia at birth with the use of deep learning," *American Journal of Obstetrics and Gynecology*, vol. 232, no. 1, pp. 116.e1-116.e9, Jan. 2025, doi: 10.1016/j.ajog.2024.04.022.
[17] Z. S. M. Hawrami, M. A. Cengiz, and E. Dünder, "Addressing Class Imbalance in Fetal Health Classification: Rigorous Benchmarking of Multi-Class Resampling Methods on Cardiotocography Data," *Diagnostics*, vol. 16, no. 3, p. 485, Feb. 2026, doi: 10.3390/diagnostics16030485.
[18] I. Nazli, E. Korbeko, S. Dogru, E. Kugu, and O. K. Sahingoz, "Early Detection of Fetal Health Conditions Using Machine Learning for Classifying Imbalanced Cardiotocographic Data," *Diagnostics*, vol. 15, no. 10, p. 1250, May 2025, doi: 10.3390/diagnostics15101250.
[19] S. Pereira *et al.*, "Is the fetus fit for labor? Introducing fast-and-frugal trees (FFTrees) to simplify triage of women for STAN monitoring: An interobserver agreement comparison with traditional classification," *Acta Obstetricia et Gynecologica Scandinavica*, vol. 103, no. 1, pp. 68–76, Jan. 2024, doi: 10.1111/aogs.14680.
[20] S. Andres *et al.*, "Physiological cardiotocography interpretation and neonatal morbidity: A historical pre–post cohort study in a tertiary perinatal center," *Acta Obstetricia et Gynecologica Scandinavica*, May 2026, doi: 10.1111/aogs.70245.
[21] I. Ben M'Barek, G. Jauvion, and P. Ceccaldi, "Computerized cardiotocography analysis during labor – A state-of-the-art review," *Acta Obstetricia et Gynecologica Scandinavica*, vol. 102, no. 2, pp. 130–137, Feb. 2023, doi: 10.1111/aogs.14498.
[22] J. B. Reitsma, A. S. Glas, A. W. S. Rutjes, R. J. P. M. Scholten, P. M. Bossuyt, and A. H. Zwinderman, "Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews," *Journal of Clinical Epidemiology*, vol. 58, no. 10, pp. 982–990, Oct. 2005, doi: 10.1016/j.jclinepi.2005.02.022.
[23] P. F. Whiting *et al.*, "QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies," *Annals of Internal Medicine*, vol. 155, no. 8, pp. 529–536, Oct. 2011, doi: 10.7326/0003-4819-155-8-201110180-00009.
[24] J. A. C. Sterne *et al.*, "RoB 2: a revised tool for assessing risk of bias in randomised trials," *BMJ*, p. l4898, Aug. 2019, doi: 10.1136/bmj.l4898.
[25] M. J. Page *et al.*, "PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews," *BMJ*, p. n160, Mar. 2021, doi: 10.1136/bmj.n160.
[26] G. Guyatt *et al.*, "GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables," *Journal of Clinical Epidemiology*, vol. 64, no. 4, pp. 383–394, Apr. 2011, doi: 10.1016/j.jclinepi.2010.04.026.
[27] F. Francis, S. Luz, H. Wu, S. J. Stock, and R. Townsend, "Machine learning on cardiotocography data to classify fetal outcomes: A scoping review," *Computers in Biology and Medicine*, vol. 172, p. 108220, Apr. 2024, doi: 10.1016/j.compbiomed.2024.108220.
[28] R. Pardasani, R. Vitullo, S. Harris, H. O. Yapici, and J. Beard, "Development of a novel artificial intelligence algorithm for interpreting fetal heart rate and uterine activity data in cardiotocography," *Frontiers in Digital Health*, vol. 7, Sep. 2025, doi: 10.3389/fdgth.2025.1638424.
[29] S. Neri, R. Ramirez Zegarra, M. Dininno, E. Di Pasquo, S. Tagliaferri, and T. Ghi, "Interobserver agreement of intrapartum

cardiotocography interpretation by midwives using current FIGO and physiology-based guidelines," *The Journal of Maternal-Fetal & Neonatal Medicine*, vol. 37, no. 1, Jan. 2024, doi: 10.1080/14767058.2024.2425758.
[30] D. Farsetti *et al.*, "The impact of maternal cardiovascular status prior to labor on birth outcomes: an observational study," *Journal of Perinatal Medicine*, vol. 54, no. 1, pp. 61–71, Jan. 2026, doi: 10.1515/jpm-2025-0329.
[31] S. Vanbelle, C. H. Engelhart, and E. Blix, "Measuring Agreement in Diagnostics: A Practical Guide for Researchers," *Statistics in Medicine*, vol. 44, no. 23–24, Oct. 2025, doi: 10.1002/sim.70299.
[32] M. Piergianni, R. Morelli, D. I. Pio Buca, F. Di Sebastiano, M. Liberati, and F. D'Antonio, "Prenatal predictor of composite adverse outcome in pregnancies undergoing induction of labour for fetal growth restriction," *Italian Journal of Gynaecology and Obstetrics*, vol. 35, no. Supplement 01, p. 30, Jan. 2023, doi: 10.36129/jog.2022.S29.
[33] G. Davis Jones, W. R. Cooke, and M. Vatish, "Identifying High-Risk Pre-Term Pregnancies Using the Fetal Heart Rate and Machine Learning," *Bioengineering*, vol. 13, no. 2, p. 203, Feb. 2026, doi: 10.3390/bioengineering13020203.