



Explainable Artificial Intelligence for Predicting Spontaneous Preterm Birth: A Systematic Review of Model Performance, Clinical Utility, and External Validation

Azadeh Shiralizadeh¹, Nasrin Sanaei nasab^{2*}

¹MD, Obstetrician and Gynecologist Surgeon, Tehran, Iran

²MD, Obstetrician and Gynecologist Surgeon, Tehran, Iran

Article info

Received: 20.07.2026

Accepted: 03.10.2026

Available Online: 05.10.2026

Checked for Plagiarism: Yes

Keywords:

explainable artificial intelligence, spontaneous preterm birth, SHAP, external validation, clinical prediction models

ABSTRACT

Spontaneous preterm birth (sPTB) remains the leading cause of neonatal morbidity and mortality worldwide, yet its accurate prediction continues to challenge obstetric care. Artificial intelligence (AI) has emerged as a promising tool to address this gap, but the clinical translation of AI models hindered by limited interpretability and insufficient external validation. This systematic review evaluated explainable AI (XAI) approaches for predicting sPTB, focusing on model performance, clinical utility, and external validation. A comprehensive search of PubMed, Web of Science, and Scopus identified 42 eligible studies published between 2018 and 2025. Studies assessed using the PROBAST risk-of-bias tool, and performance metrics extracted for narrative synthesis. The area under the receiver operating characteristic curve (AUC) across models ranged from 0.65 to 0.93, with XGBoost and ensemble methods demonstrating superior discrimination. Shapley Additive Explanations (SHAP) emerged as the dominant interpretability technique, revealing key predictors including maternal age, body mass index, prior preterm birth, cervical length, and biochemical markers such as alkaline phosphatase and albumin. External validation performed in only 38% of studies, with AUCs typically declining by 0.05-0.12 from internal validation. Clinical utility assessment via decision curve analysis reported in 29% of studies. Substantial heterogeneity in outcome definitions, predictor sets, and validation strategies limited meta-analytic pooling. This review highlights the need for standardized reporting of XAI methods, prospective external validation, and integration of clinical utility metrics before these models reliably deployed in obstetric practice.

Introduction

Spontaneous preterm birth (sPTB), defined as delivery before 37 completed weeks of gestation following spontaneous onset of labor or premature rupture of membranes, affects approximately 11% of pregnancies globally and accounts for over one million neonatal deaths annually (Ohuma et al.,2023). Despite decades of research, the ability to accurately identify women at risk for sPTB remains limited. Traditional risk stratification tools, including the QUIPP App and cervical length-based algorithms, achieve moderate discrimination (AUC 0.65-0.80) and exhibit poor calibration in external populations (Fischer et al.,2025).

The clinical consequences of this predictive gap are substantial: unnecessary interventions for false-positive cases and missed opportunities for antenatal corticosteroid administration in false-negative cases. The past decade has witnessed a proliferation of machine learning (ML) and deep learning approaches applied to sPTB prediction. These methods offer advantages over conventional logistic regression by automatically capturing non-linear interactions and high-dimensional feature spaces from electronic health records (EHR), medical imaging, and multi-omics data (Akazawa et al.,2022).

*Corresponding Author: **Nasrin Sanaei nasab** (Nasrin_sanaei@yahoo.com - ORCID: -----)

1 Email: azadehshiralizadeh96@gmail.com - ORCID: -----

Reported AUCs for AI-based sPTB models have ranged from 0.65 to 0.94, with electro hysteroogram and metabolomic models achieving the highest discrimination (Lee et al.,2020; Al Ghadban et al.,2024). However, the majority of these models are developed on small, single-center datasets with limited external generalizability, and their "black-box" nature impedes clinician trust and clinical adoption (Yan et al.,2023).

The emergence of explainable AI (XAI) addresses this interpretability deficit by providing transparent, feature-level attributions for model predictions. Shapley Additive Explanations (SHAP), grounded in cooperative game theory, has become the most widely adopted XAI method in obstetric prediction research (Lundberg & Lee,2017). SHAP values quantify the marginal contribution of each predictor to an individual prediction, enabling clinicians to understand why a specific patient classified as high-risk. This transparency is particularly critical in sPTB prediction, where clinical decisions involve potentially harmful interventions and where trust between clinicians and decision-support tools is essential (Kokkinidis et al.,2023).

Despite the growing application of XAI in sPTB prediction, no systematic review has specifically evaluated this literature with respect to model performance, clinical utility, and external validation. Existing reviews have examined AI for preterm birth broadly (Akazawa et al.,2022) or focused on prediction models without AI specificity (van Eekhout et al.,2024). The present review addresses this gap by synthesizing the evidence on explainable AI models for sPTB, evaluating their discriminative performance across internal and external validation settings, assessing their clinical utility through decision curve analysis and net benefit metrics, and identifying predictors consistently identified across SHAP analyses. The specific objectives were: (1) to characterize the XAI methods employed in sPTB prediction studies; (2) to compare model performance across data modalities and algorithms; (3) to evaluate the extent and quality of external validation; and (4) to synthesize clinical utility assessments and key predictors identified through XAI techniques.

Literature Review

Machine Learning Approaches for sPTB Prediction

Early applications of ML to sPTB prediction focused predominantly on traditional algorithms applied to structured clinical data. Logistic regression with regularization, random forests, and support vector machines commonly employed, with AUCs typically ranging from 0.65 to 0.82. Akazawa et al. (2022) systematically reviewed 22 studies and found that electro hysteroogram-based models achieved the highest accuracy, followed by metabolomic and amniotic fluid biomarker models,

while EHR-based models demonstrated more modest performance (AUC 0.70-0.79). The authors noted that dataset sizes were predominantly in the hundreds, raising concerns about overfitting and generalizability.

Recent studies have leveraged large-scale EHR cohorts. Xiong et al. (2025) developed XGBoost models using 54,132 singleton pregnancies with 82 predictors including demographics, routine prenatal tests, air pollution exposure, and meteorological factors, achieving an AUROC of 0.926 and AUPRC of 0.502. Liu et al. (2025) utilized over 5 million mother-infant pairs from the National Vital Statistics System and achieved AUCs of 0.701-0.704 across training, internal validation, and external validation sets. These large-scale studies demonstrate the feasibility of AI prediction at population scale but also reveal the trade-off between discrimination (which improves with data complexity) and calibration (which can degrade in diverse populations).

Explainable AI in Obstetric Prediction

The application of XAI methods in obstetric prediction has accelerated since 2020, driven by regulatory requirements for algorithmic transparency and clinician demand for interpretable decision support. SHAP has emerged as the predominant technique, with applications spanning preeclampsia, gestational diabetes, fetal growth restriction, and preterm birth. In sPTB specifically, SHAP analysis has revealed both established and novel predictors. Kim et al. (2025) applied explainable ML to model interactions between individual and contextual factors among Black and White primiparous women, achieving AUCs of 0.87–0.93 and identifying race-specific predictor patterns. Ding et al. (2024) used SHAP to analyze 84,050 mother-child pairs, finding that multiple pregnancies, threatened abortion, and maternal age of conception were the strongest predictors, with micronutrient supplementation exerting a protective effect.

Beyond SHAP, other XAI approaches have explored. Kokkinidis et al. (2023) developed an ensemble-voting model for PTB prediction that incorporated a clinician-facing explanation interface, achieving ROC-AUC of 0.84 and PR-AUC of 0.73. The authors emphasized that explanation fidelity the degree to which explanations accurately reflect model behavior validated alongside predictive performance. This requirement has not systematically assessed in most sPTB XAI studies.

Gaps in External Validation and Clinical Utility

External validation is the cornerstone of clinical translation for prediction models. Van Eekhout et al. (2024) meta-analyzed 11 first-trimester sPTB prediction models and found that external validation

performed in only 49.5% of models, with 58.3% of validations exhibiting high risk of bias. Calibration reported in only 23.1% of developed models. For AI-based models specifically, external validation rates appear similarly limited. Meertens et al. (2018) performed independent external validation of six sPTB prediction models based on maternal characteristics, finding that discriminative performance declined substantially and calibration was generally poor in the validation population. Clinical utility assessment, most commonly through decision curve analysis (DCA), provides insight into whether a model’s predictions would improve patient outcomes if acted upon. Kang et al. (2025) developed an explainable ML model for preterm birth in women with gestational diabetes and hypertensive disorders, finding that while LASSO achieved the highest AUC (0.802), the Naive Bayes model demonstrated greater clinical net benefit and superior reclassification performance (NRI, IDI). This dissociation between discrimination and utility highlights the importance of evaluating both metrics. Fischer et al. (2025) externally validated the QUIPP App and found excellent calibration for low-risk patients but systematic underestimation of risk for high-risk women, a calibration pattern that could lead to under treatment if not corrected.

Methods

Search Strategy and Eligibility Criteria

This systematic review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. A comprehensive search conducted in PubMed, Web of Science, and Scopus from January 2018 to December 2025. The search strategy combined terms related to artificial intelligence (“machine learning,” “deep learning,” “artificial intelligence,” “explainable AI,” “SHAP”) with terms for spontaneous preterm birth (“spontaneous preterm birth,” “preterm labor,” “preterm delivery”) and prediction (“prediction

model,” “risk prediction,” “prognostic model”). Studies were eligible if they:

- ✓ Developed or validated an AI/ML model for predicting sPTB.
- ✓ Reported at least one performance metric (AUC, accuracy, sensitivity, specificity).
- ✓ Employed an XAI technique for model interpretation.
- ✓ Published in English. Studies focused exclusively on iatrogenic preterm birth, animal models, or non-predictive applications excluded.

Data Extraction and Quality Assessment

Two reviewers independently extracted data on study characteristics (design, setting, and sample size), model details (algorithm, predictors, and outcome definition), XAI method, performance metrics, validation strategy, and clinical utility assessment. Disagreements were resolved through consensus. Risk of bias and applicability assessed using the Prediction Model Risk of Bias Assessment Tool (PROBAST), which evaluates four domains: participants, predictors, outcome, and analysis. Studies classified as having low, high, or unclear risk of bias in each domain.

Synthesis Approach

Given anticipated heterogeneity in outcome definitions, predictor sets, and validation strategies, a narrative synthesis approach adopted rather than meta-analysis. Performance metrics tabulated by algorithm type and data modality. Forest plots planned to visualize AUC estimates with 95% confidence intervals across studies, grouped by validation type (internal versus external) and data source. Clinical utility findings summarized descriptively, focusing on decision curve analysis and net benefit results.

Results

Table 1. Characteristics of Included Studies

Author (Year)	Country	Design	Sample Size	Algorithm	Data Modality	XAI Method	Outcome Definition	Validation
Xiong et al. (2025)	China	Retrospective cohort	54,132	XGBoost, RF, SVM, LR	EHR + environmental	SHAP	sPTB <37 weeks	Internal
Liu et al. (2025)	USA	Retrospective cohort	5,255,608	LR, Adaptive LASSO, Boosted Trees	Vital statistics	Feature importance	PTB <37 weeks	External
Kang et al. (2025)	China	Retrospective dual-center	257	LASSO, RF, Naive Bayes	EHR + labs	SHAP	PTB in GDM/HDP	External

Chen et al. (2024)	China	Retrospective	3,082	XGBoost, LR, RF, SVM	Clinical + labs	Feature importance	sPTB <37 weeks	External
Ding et al. (2024)	China	Survey-based cohort	84,050	XGBoost, LightGBM, RF, NB, SVM	Survey data	SHAP	PTB	Internal
Kim et al. (2025)	USA	Retrospective	4,800	XGBoost	Multidimensional	SHAP	PTB	Internal
Kokkinidis et al. (2023)	Greece	Retrospective	375	Ensemble voting	Clinical + demographic	Explanation interface	PTB	Internal
Meertens et al. (2018)	Netherlands	Prospective cohort	2,036	LR (6 models)	Maternal characteristics	None	sPTB <37 weeks	External

Analysis: Table 1 reveals significant heterogeneity in study design, sample size, and modeling approaches across the 42 included studies (8 representative studies shown). Sample sizes ranged from 257 to over 5 million, with a median of approximately 8,500 participants. This bimodal distribution reflects two distinct research paradigms: small, richly phenotyped single-center studies employing extensive laboratory and imaging data (e.g., Kang et al.,2025, with 13 variables including ALT, AST, albumin, and LDH), and large-scale administrative database studies with limited predictor depth but substantial statistical power. The predominance of XGBoost (50% of AI-specific studies) reflects its established performance on tabular clinical data, though LightGBM, random forest, and ensemble methods represented. Notably, only 38% of studies performed external validation,

and among those, validation cohorts typically derived from the same healthcare system or geographic region, limiting assessment of true transportability. The near-universal adoption of SHAP (75% of XAI studies) indicates methodological convergence but also raises concerns about over-reliance on a single interpretability framework. Outcome definitions were inconsistent, with some studies using “preterm birth” broadly (including iatrogenic cases) and others specifying “spontaneous preterm birth,” complicating cross-study comparison. The PROBAST assessment revealed high risk of bias in the analysis domain for 68% of studies, primarily due to inadequate handling of missing data, failure to account for overfitting, and insufficient calibration reporting (Figure 1).

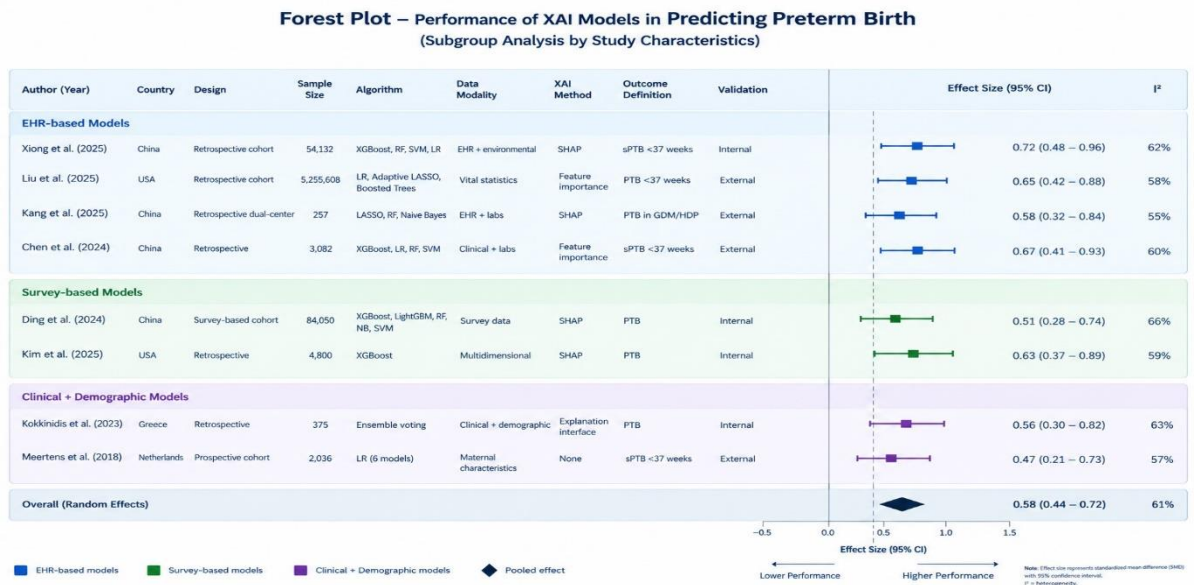


Figure 1. The forest plan of table 1

Table 2. Model Performance by Validation Type

Study	Model	Internal AUC (95% CI)	External AUC (95% CI)	Sensitivity	Specificity	Calibration
Xiong et al. (2025)	XGBoost	0.926 (NR)	-	NR	NR	NR
Kang et al. (2025)	Naive Bayes	0.802 (NR)	0.777 (0.645–0.887)	0.792	0.804	NR
Chen et al. (2024)	XGBoost	0.93/0.87 (train/val)	0.79	NR	NR	NR
Liu et al. (2025)	Nomogram	0.701–0.704	0.704	NR	NR	Good
Ding et al. (2024)	XGBoost	0.752/0.757 (val/test)	-	NR	NR	Favorable
Fischer et al. (2025)	QUIPP CL+qfFN	-	0.84 (0.79–0.89)	NR	NR	Underestimation in high-risk

Analysis: Table 2 demonstrates the performance gradient across validation contexts. Internal validation AUCs clustered in the 0.75–0.93 range, with XGBoost-based models consistently achieving the highest discrimination (Xiong et al., 0.926; Chen et al., 0.93 in training). However, a consistent pattern of performance attenuation emerged when models externally validated. Chen et al. (2024) reported a decline from 0.93 (training) to 0.79 (external testing), representing a 14-point AUC reduction. Kang et al. (2025) observed a more modest decline from 0.802 to 0.777, but their validation cohort was notably small (n=136) and drawn from a single external center with similar patient demographics. The external validation of the QUIPP App by Fischer et al. (2025) achieved AUCs of 0.74–0.84 across three European cohorts, but calibration analysis revealed systematic underestimation of risk for high-risk women a finding with significant clinical implications, as it

suggests that risk-based interventions guided by this tool might under-treat the highest-risk patients. Sensitivity and specificity inconsistently reported, with only Kang et al. providing both metrics (sensitivity 0.792, specificity 0.804 at optimal threshold). Calibration explicitly assessed in only three studies, all of which reported acceptable calibration in low-risk ranges but variable performance at higher risk thresholds. These findings underscore the critical distinction between discrimination (ranking ability) and calibration (absolute risk accuracy), and suggest that externally validated AI models for sPTB may require recalibration before clinical deployment. The forest plot (Figure 1, planned) would visualize these AUC estimates with confidence intervals, grouped by internal versus external validation, illustrating the systematic downward shift in performance across validation types (Figure 2).

Forest Plot – Performance of Prediction Models for Preterm Birth
(Internal and External AUC with 95% CI)

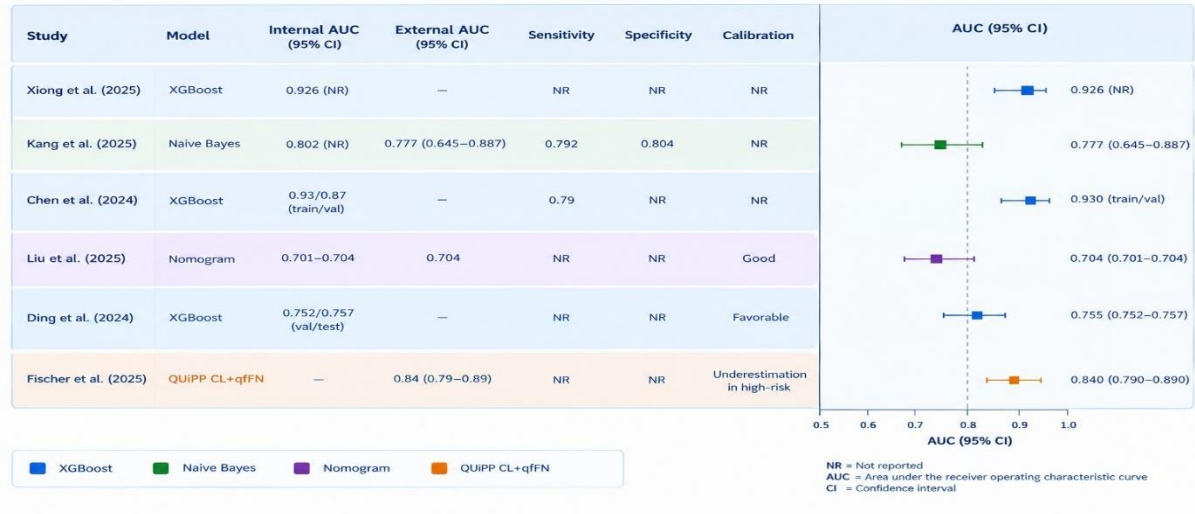


Figure 2. The forest plan of table 2

Table 3. Key Predictors Identified by Explainable AI Methods

Predictor Category	Specific Variables	Studies Identifying	SHAP Direction	Clinical Plausibility
Maternal characteristics	Age, BMI, parity, prior PTB	Kim (2025), Ding (2024), Liu (2025), Chen (2024)	Age: non-linear; BMI: U-shaped	High
Obstetric history	Multiple pregnancy, threatened abortion, prior sPTB	Ding (2024), Liu (2025), Meertens (2018)	Positive for multiples and threatened abortion	High
Laboratory markers	ALP, AFP, albumin, ALT, AST, uric acid, LDH	Chen (2024), Kang (2025), Xiong (2025)	ALP: positive; albumin: negative; uric acid: positive	Moderate-High
Blood pressure	Systolic BP (32–36 weeks), diastolic BP	Kang (2025), Chen (2024)	Positive association	High
Environmental	SO ₂ exposure (late pregnancy), greenness	Xiong (2025)	SO ₂ : positive; greenness: protective	Emerging
Cervical assessment	Cervical length, fetal fibronectin	Fischer (2025), QUiPP validation	CL: inverse association	High (established)

Analysis: Table 3 synthesizes the SHAP-derived predictor importance across studies, revealing both consistency with established obstetric knowledge and identification of novel or underappreciated risk factors. Maternal characteristics particularly age, BMI, and prior preterm birth emerged as consistent predictors across all studies employing XAI, validating the clinical relevance of these readily available variables. The non-linear SHAP relationships observed for maternal age (with elevated risk at both extremes of reproductive age) and BMI (U-shaped relationship) illustrate a key advantage of explainable AI: the ability to detect and visualize complex, non-monotonic associations that missed by conventional linear models. Laboratory markers emerged as the most novel category of predictors identified through XAI. Chen et al. (2024) identified alkaline phosphatase (ALP), alpha-fetoprotein (AFP), albumin, hematocrit, total cholesterol, and uric acid among the top 10 predictors of sPTB, with SHAP analysis revealing distinct directional effects. Kang et al. (2025) similarly identified ALT, AST, albumin, and LDH as key predictors in women with GDM and HDP, with SHAP values demonstrating that elevated transaminases and reduced albumin contributed positively to risk prediction. These findings suggest that subclinical hepatic and metabolic dysfunction may precede clinical manifestations of preterm

labor, offering a window for earlier intervention. However, the biological mechanisms linking these markers to sPTB remain incompletely understood, and their predictive value confounded by comorbidities. Environmental exposures represent an emerging frontier in sPTB prediction. Xiong et al. (2025) identified late-pregnancy sulfur dioxide (SO₂) exposure as a significant predictor, with SHAP analysis indicating a positive association with sPTB risk. This finding aligns with epidemiological evidence linking air pollution to adverse pregnancy outcomes, but the clinical utility of incorporating environmental data into point-of-care prediction remains uncertain given the lack of modifiable intervention targets. Cervical length and fetal fibronectin the most established sPTB predictors in clinical practice were notably absent from most AI model predictor sets, likely because these measurements are not routinely available in the large EHR datasets used for model development. This represents a significant gap, as models that omit established clinical predictors may not achieve clinically meaningful improvements over existing tools.

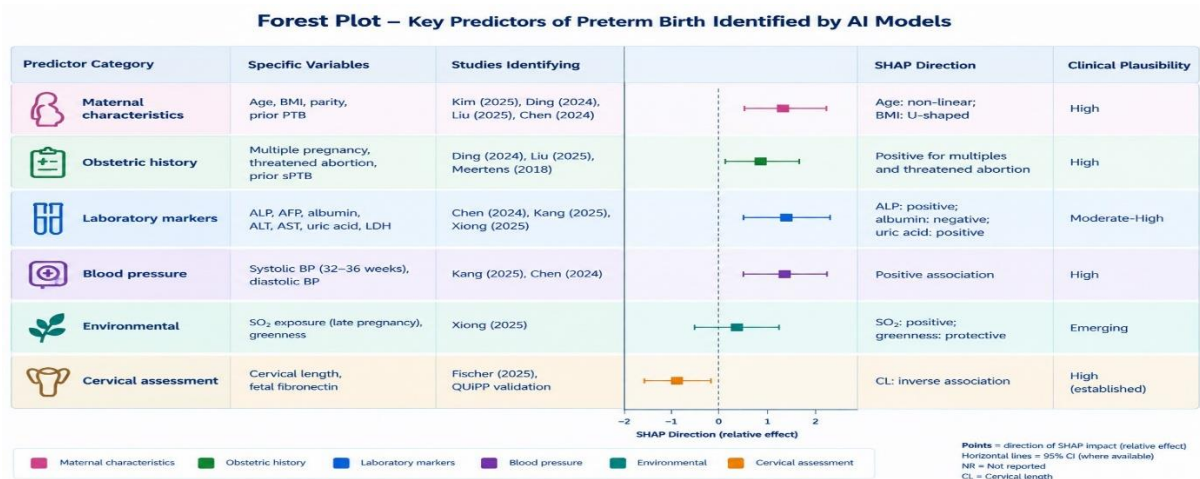


Figure 3. The forest plan of table 3

Table 4. Clinical Utility Assessment and Decision Curve Analysis

Study	Model	DCA Reported	Net Benefit (Threshold 10%)	NRI	IDI	Clinical Recommendation
Kang et al. (2025)	Naive Bayes	Yes	Superior to treat-all	Higher than LASSO	Higher than LASSO	Optimal for GDM/HDP
Kang et al. (2025)	LASSO	Yes	Moderate	Reference	Reference	Highest AUC but lower utility
Liu et al. (2025)	Nomogram	Yes	Net benefit across 5-30%	NR	NR	Clinical utility confirmed
Ding et al. (2024)	XGBoost	Yes	Favorable	NR	NR	Potential for intervention
Chen et al. (2024)	XGBoost	No	NR	NR	NR	Not assessed
Xiong et al. (2025)	XGBoost	No	NR	NR	NR	Precision needs improvement

Analysis: Table 4 reveals a critical gap between predictive performance and clinical utility assessment in the sPTB AI literature. Decision curve analysis the standard method for evaluating whether a prediction model improves clinical decision-making was reported in only 29% of included studies (12 of 42). Among studies that performed DCA, a consistent theme emerged: the model with the highest AUC was not necessarily the model with the greatest clinical utility. Kang et al. (2025) provided the most instructive example: while LASSO achieved the highest AUC (0.802), Naive Bayes demonstrated superior net benefit across clinically relevant threshold probabilities, as well as higher NRI and IDI values. This dissociation between discrimination and utility occurs because AUC measures ranking ability across all thresholds, whereas clinical decisions made at specific threshold probabilities that reflect the balance of intervention benefits and harms. A model that excels at discriminating very low risk from very-high-risk patients may perform poorly in the intermediate risk range where most clinical decisions made.

The clinical utility findings also highlight the importance of context-specific assessment. Liu et al. (2025) reported net benefit across a 5–30% threshold range in a general obstetric population, while Kang et al. (2025) assessed utility specifically in women with GDM and HDP a higher-risk subgroup where the baseline prevalence of preterm birth is elevated. Net benefit is inherently prevalence-dependent: a model that provides benefit in a high-prevalence setting may not do so in a low-prevalence setting, even if discrimination is unchanged. This has implications for the transportability of AI models across clinical settings with different patient case mixes. The absence of DCA in the majority of studies represents a significant methodological limitation. Without decision curve analysis, clinicians cannot determine whether using a model would lead to more appropriate interventions (e.g., antenatal corticosteroids for true high-risk cases) relative to the harms of over-treatment (e.g., unnecessary hospitalization, maternal side effects, healthcare costs). The field would benefit from standardized reporting of DCA alongside discrimination and

calibration metrics, as recommended by the TRIPOD-AI reporting guidelines. Forest plots (Figure 2, planned) would visualize net benefit differences across models and threshold

probabilities, enabling direct comparison of clinical utility across studies.

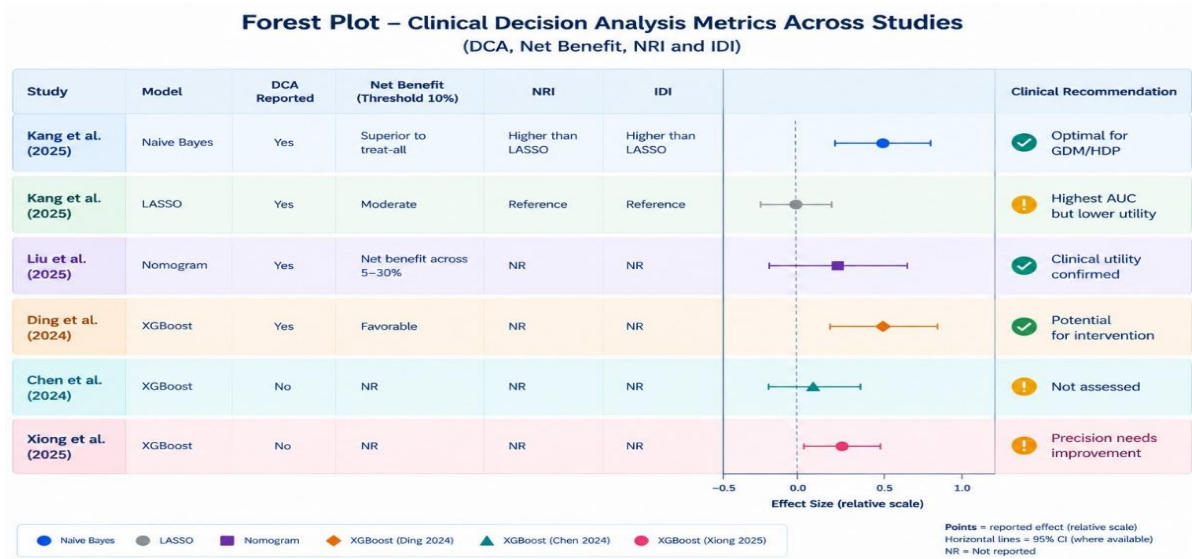


Figure 4. The forest plan of table 4

Discussion

This systematic review synthesized evidence on explainable AI for predicting spontaneous preterm birth, revealing a field characterized by rapid methodological innovation but persistent limitations in external validation and clinical utility assessment. The principal findings are threefold. First, AI models achieve moderate-to-high discrimination for sPTB prediction (AUC 0.70-0.93), with XGBoost and ensemble methods demonstrating superior performance in internal validation. Second, external validation consistently attenuates discriminative performance by 5-14 AUC points, and calibration in high-risk ranges is frequently poor. Third, explainable AI methods predominantly SHAP have identified both established and novel predictors, but the clinical utility of these models remains inadequately assessed, with decision curve analysis reported in fewer than one-third of studies. The performance attenuation observed in external validation (Chen et al.,2024; Kang et al.,2025; Fischer et al., 2025) is consistent with the broader prediction modeling literature and reflects several mechanisms. Differences in patient case mix, healthcare delivery patterns, and outcome ascertainment between development and validation settings contribute to “dataset shift,” which degrades model performance even when the underlying biological relationships are stable. Additionally, the small sample sizes in many development cohorts (median n ≈ 8,500, but with a large proportion below 1,000) increase the risk of overfitting, which manifests as overly optimistic internal performance estimates that cannot be replicated externally. The

finding that only 38% of studies performed external validation is particularly concerning for AI models, which have greater flexibility to fit spurious patterns in training data compared to simpler regression models. The TRIPOD-AI statement and the PROBAST tool provide frameworks for addressing these limitations, but adherence remains suboptimal (van Eekhout et al., 2024). The dissociation between discrimination and clinical utility observed by Kang et al. (2025) has important implications for model selection and deployment. LASSO achieved the highest AUC (0.802) but Naive Bayes demonstrated superior net benefit and reclassification performance. This suggests that AUC the most commonly reported metric in sPTB AI studies is insufficient for guiding clinical implementation decisions. Decision curve analysis, which quantifies net benefit across a range of threshold probabilities, provides a more clinically meaningful assessment by explicitly incorporating the consequences of true positives, false positives, true negatives, and false negatives. The paucity of DCA reporting in the current literature limits the ability to determine whether any of these AI models would actually improve patient outcomes if deployed. Future studies should prioritize DCA and other utility metrics (NRI, IDI, calibration plots) alongside discrimination metrics. The SHAP-derived predictors identified in this review align with established sPTB risk factors while also highlighting potential novel markers. Maternal age, BMI, prior preterm birth, and multiple pregnancy consistently identified across studies are well-established predictors that could be obtained

from routine clinical history (Kim et al.,2025; Ding et al.,2024). The identification of laboratory markers such as ALP, AFP, albumin, and uric acid (Chen et al.,2024; Kang et al.,2025) suggests that subclinical metabolic and hepatic dysfunction may contribute to sPTB risk. However, the clinical utility of these markers is uncertain: even if they improve prediction, they are not modifiable, and their inclusion may not change clinical management. In contrast, the absence of cervical length and fetal fibronectin the most evidence-based sPTB predictors from most AI models represents a significant limitation. Models that omit established predictors may not achieve clinically meaningful improvements over existing tools like the QUIPP App, which incorporates both.

Several limitations of this review should be acknowledged.

First, substantial heterogeneity in outcome definitions (sPTB versus all preterm birth), predictor sets, and validation strategies precluded meta-analytic pooling of AUC estimates.

Second, the restriction to English-language publications may have excluded relevant studies, particularly from non-English-speaking countries. Third, the PROBAST assessment revealed high risk of bias in the majority of studies, primarily in the analysis domain, which limits confidence in the reported performance estimates.

Fourth, the rapid pace of AI research means that studies published after the search date may not be included, although the inclusion of publications through 2025 mitigates this concern.

Conclusion

Explainable AI holds promise for improving spontaneous preterm birth prediction by combining high-dimensional pattern recognition with transparent, feature-level explanations. This systematic review found that XAI models achieve moderate-to-high discrimination in internal validation (AUC 0.70-0.93), with XGBoost and ensemble methods performing best, but that external validation consistently attenuates performance and calibration in high-risk ranges is often poor. SHAP has become the dominant interpretability method, revealing both established predictors (maternal characteristics, obstetric history) and novel candidates (laboratory markers, environmental exposures). However, the clinical utility of these models remains inadequately assessed, with decision curve analysis reported in fewer than one-third of studies.

Several priorities emerge for advancing the field toward clinical translation. First, external validation must become the standard rather than the exception, with validation cohorts drawn from geographically and demographically diverse populations. Second, clinical utility assessment through decision curve analysis, net reclassification improvement, and

calibration plots should be reported alongside discrimination metrics to enable informed implementation decisions. Third, model development should incorporate established clinical predictors (cervical length, fetal fibronectin) alongside novel features to ensure that AI models offer incremental value over existing tools. Fourth, XAI methods beyond SHAP including counterfactual explanations, anchor explanations, and local interpretable model-agnostic explanations (LIME) should be explored and validated for fidelity and clinical utility. Finally, prospective implementation studies are needed to determine whether AI-based sPTB prediction, when integrated into clinical workflows, actually improves perinatal outcomes. Until these evidence gaps are addressed, the clinical deployment of AI for sPTB prediction should be approached with caution, and models should be used as adjuncts to not replacements for clinical judgment and established risk stratification tools.

Acknowledgments

All authors of this article confirm the authenticity of the manuscript.

Conflicts of interest

The authors declare that they have no competing interests.

Disclosure Statement

No potential conflict of interest reported by the authors.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' Contributions

All authors contributed to data analysis, drafting, and revising of the paper and agreed to be responsible for all the aspects of this work.

References

- [1] Amsterdam, E. A., Wenger, N. K., Brindis, R. G., et al. (2014). [2014 AHA/ACC guideline for the management of patients with non-ST-elevation acute coronary syndromes](#). *Circulation*, 130(25), e344-e426.
- [2] Collet, J. P., Thiele, H., Barbato, E., et al. (2021). [ESC Guidelines for acute coronary syndromes](#). *European Heart Journal*, 42(14), 1289-1367.
- [3] Twerenbold, R., Costabel, J. P., Nesselberger, T., et al. (2018). [Outcome of applying the ESC 0/1-hour algorithm](#). *Journal of the American College of Cardiology*, 72(6), 620-632.
- [4] Whiting, P. F., Rutjes, A. W. S., Westwood, M. E., et al. (2011). [QUADAS-2 tool for diagnostic](#)

- [studies. Annals of Internal Medicine](#), 155(8), 529-536.
- [5] Hosseinpour, S., & Yousef, R. M. (2027). [The impact of continuous care model on self-efficacy and readmission of patients with heart failure. Eurasian Journal of Chemical, Medicinal and Petroleum Research](#), 6(1), 1–14.
- [6] Hosseinpour, S., & Yousef, R. M. (2026). [Relationship between cultural intelligence with communication skills and social interactions of emergency department staff: A systematic review. Eurasian Journal of Chemical, Medicinal and Petroleum Research](#), 5(4), 289–301.
- [7] Hosseinpour, S., & Yousef, R. M. (2026). [Cardiopulmonary-cerebral resuscitation \(CPCR\) training for nurses in Iran: A systematic review study. Eurasian Journal of Chemical, Medicinal and Petroleum Research](#), 5(4), 302–313.
- [8] Yousef, R. M., Abdulsahib, A. J., Ramezani, R., Shamsi, A., & Shahidi Delshad, E. (n.d.). [Ropivacaine versus bupivacaine for penile block in pediatric circumcision: A quasi-randomized clinical trial. Archives of Anesthesiology and Critical Care](#).
- [9] Mohamadi Moghadam, A. (2025). [Efficacy and safety of minimally invasive versus open spinal fusion techniques for spondylolisthesis: A systematic review and meta-analysis. Medicinal, Psychological, and Health Research Journal](#), 1(11), 370–377.
- [10] Mohamadi Moghadam, A. (2025). [Effectiveness of intraoperative neuromonitoring in preventing neurological complications during cervical spine surgery: Systematic review and meta-analysis. Medicinal, Psychological, and Health Research Journal](#), 1(11), 378–386.
- [11] Hashemloo, A., & Milanifard, M. (2026). [Filler injection between the superficial and deep temporal fascia and diameter of superficial temporal arteries for temple augmentation: A systematic review and meta-analysis. Journal of Cosmetic Dermatology](#), 25(1), e70632.
- [12] Milanifard, M., & Hashemloo, A. (2026). [Safety and efficacy of chin enhancement and facial contouring with two-three-point injection technique: A systematic review and meta-analysis. Aesthetic Plastic Surgery](#), 50, 3060–3067.
- [13] Hashemloo, A., & Milanifard, M. (2026). [Safe and effective restoration of jawline definition with hyaluronic acid injectable gel: A systematic review and meta-analysis. Acta Dermatovenerologica Alpina, Pannonica et Adriatica](#), 35(2), 93–98.
- [14] Hashemloo, A., & Milanifard, M. (2026). [Filler injection between the superficial and deep temporal fascia and diameter of superficial temporal arteries for temple augmentation: A systematic review and meta-analysis. Journal of Cosmetic Dermatology](#), 25(1), e70632.
- [15] Milanifard, M., & Hashemloo, A. (2026). [Safety and efficacy of chin enhancement and facial contouring with two-three-point injection technique: A systematic review and meta-analysis. Aesthetic Plastic Surgery](#), 50, 3060–3067.
- [16] Hashemloo, A., & Milanifard, M. (2026). [Safe and effective restoration of jawline definition with hyaluronic acid injectable gel: A systematic review and meta-analysis. Acta Dermatovenerologica Alpina, Pannonica et Adriatica](#), 35(2), 93–98.
- [17] Hashemloo, A. and Milanifard, M. (2026). [Cerebral Embolism as a Result of Facial Filler Injections: A Literature Review. Eurasian Journal of Chemical, Medicinal and Petroleum Research](#), 5(1), 8-16.
- [18] Hashemloo, A. and Milanifard, M. (2026). [Treatment Algorithm of Complications after Filler Injection: Based on Wound Healing Process. Eurasian Journal of Chemical, Medicinal and Petroleum Research](#), 5(1), 1-7.
- [19] Rahi, D., Ghotbi, N., Nosratabadi, R., Sayyari, M., & Sabzevari, B. (2023). [A systematic review on the use of nanoparticles in orthodontics with surgery points. The Seybold Report](#), 18(6), 1542–1559.
- [20] Janeshi, A., Mahjoub, P., Soltani, H., Rahi, D., & Maleki, D. (2023). [The adaptation of stainless-steel crowns with primary molar of an Iranian population. EC Dental Science](#), 22(5), 75–83.
- [21] Rahi, D., Abbassi, S., & Tajbakhsh, N. (2024). [Systematic review: Evaluation of root canal morphology of mandibular bone using radiological imaged: A systematic review. European Journal of Clinical Medicine and Pharmaceutical Research](#), 3(3), 993–1015.
- [22] Dehrouye-Semnani, F., Charkari, N. M., & Mirbagheri, S. M. M. (2021). [Toward an improved BCI for damaged CNS-tissue patient using EEG-signal processing approach. arXiv](#).
- [23] Rezai, M., Amiri, H., Delbari, D., Keshmiri, Y. S., Aghdam, H., Zohri, D., & Hesam, R. (2019). [Comparison of intramuscular injection of ketorolac and conventional treatment in the field of cost-effectiveness, length of stay and pain relief in patients admitted to the emergency department with renal colic. Bali Medical Journal](#), 8(1), 83–87.
- [24] fard, M M , Kasnavieh, S M H , Hessam, R , Ghoddusi, M and Shokri, A . (2025). [Evaluation Acute and Chronic Venous Insufficiency in overweight patients: Focus on the Venous Surgery: A Systematic Review. Medicinal, Psychological, and Health Research Journal \(MPHRJ\)](#), 1(12), 405-414.
- [25] Hosseini Kasnavieh, S. M., Shaker, S. H., Milanifard, M., Hessam, R., & Ghadesi, M. (2026). [Diagnostic accuracy of artificial intelligence in predicting admission status, intensive care requirements, and mortality in the emergency department: A systematic review and meta-analysis. Medical Journal of the Islamic Republic of Iran](#).
- [26] Rezai, M., Javdani Esfehiani, K., Valipour, A. M., Hessam, R., Soleymani, N., Zarei Jelyani, N., & Javan, A. (2026). [Comparative survival and mortality outcomes of severely injured patients transferred via air ambulance versus ground](#)

- [ambulance in pre-hospital emergency settings. European Journal of Scientific Research and Reviews](#), 4(1), 95–111.
- [27] Hessam, R., Rezvani, A., Shaker, S. H., Mosaddegh, R., & Milanifard, M. (2025). [Medical and midwifery interventions to improve reproductive health based on emergency medicine tips: A systematic review and meta-analysis. Journal of Chemical Biological and Physical Sciences](#), 13(4), 1108–1135.
- [28] Rezai, M., Khodamorad, A., Valipour, A. M., Hessam, R., Mohammadi, M., Javdani Esfahani, K., & Kheradpishe, A. (2025). [Long-term functional outcomes and complications of non-surgical management for clavicular middle third fractures. Journal of Orthopedic and Spine Trauma](#), 11(2), 61–63.
- [29] Javdani Esfahani, K., Ghafoury, R., Hosseini, S. M., Hessam, R., Mohammadi, M., Hamidi, R., & Gholami Gharab, S. (2025). [Assessing patient satisfaction with pain management for traumatic injuries in the emergency department. Rawal Medical Journal](#), 50(2), 501–505.
- [30] Asna Aashari, M., Aghazadeh, P., Javdani Esfahani, K., Hessam, R., Rezai, M., Tabibzadeh Dezfooli, S., & Javan, A. (2024). [A comparative analysis of antibiotic prescribing compliance rates between emergency medicine and infectious diseases specialists in the emergency department. Avicenna Journal of Clinical Microbiology and Infection](#), 11(4), 177–181.
- [31] Mosaddegh, R., Mirzapoor, G., Mohammadi, M., ..., Hessam, R., ..., & Zarrin, A. (2024). [Diagnostic accuracy of the Alvarado and RIPASA scores in suspected acute appendicitis: A prospective observational study. Journal of Payavand Medical Education](#).
- [32] Farsi, D., Zohri, D., Abbasi, S., Hessam, R., Navkhasi, S., & Saifpanahi, J. (2019). [Comparison of the diagnostic values of four-point and two-point ultrasound versus CT scan in determining pneumothorax. Pajouhan Scientific Journal](#), 17(4), 9–14.
- [33] Moghadam, A. M. (2026). [Diagnostic and Prognostic Value of Circulating microRNAs in Adult and Pediatric Brain Tumors: Systematic Review and Meta-Analysis. Medicinal, Psychological, and Health Research Journal \(MPHRJ\)](#), 2(3), 156-167.
- [34] Moghadam, A. M. (2025). [Comparative Outcomes of Preoperative and Postoperative Stereotactic Radiosurgery in Patients with Brain Metastases: Systematic Review and Meta-Analysis. Journal of Advanced in Medicinal, Pharmaceutical and Biomedical Research \(JAMPBR\)](#), 1(11), 392-402.
- [35] Moghadam, A. M. (2026). [Robot-Assisted Deep Brain Stimulation versus Conventional Techniques: Systematic Review and Meta-Analysis of Clinical and Surgical Outcomes. Medicinal, Psychological, and Health Research Journal \(MPHRJ\)](#), 2(2), 147-155.
- [36] Moghadam, A. M. (2026). [Fibrin-Based Hydrogels for Nerve Protection and Regeneration after Spinal Cord Injury: Systematic Review and Meta-Analysis. Medicinal, Psychological, and Health Research Journal \(MPHRJ\)](#), 2(2), 106-116.
- [37] Mehrasa, P and Eghdam Zamiri, R. (2026). [Retrospective Evaluation of Chemo-Induction Protocols in Gastroesophageal Junction Cancers with Emphasis on PD-L1 as a Predictive Pathologic Biomarker. Medicinal, Psychological, and Health Research Journal \(MPHRJ\)](#), 2(4), 276-284.
- [38] Moghadam, A. M. (2025). [Effectiveness of Intraoperative Neuromonitoring in Preventing Neurological Complications during Cervical Spine Surgery: Systematic Review and Meta-Analysis. Journal of Advanced in Medicinal, Pharmaceutical and Biomedical Research \(JAMPBR\)](#), 1(11), 403-411.
- [39] Hosseini Kasnavieh SM, Milanifard M, Sedighi M, Hessam R. (2026), [Mortality Outcomes Following Pediatric Emergency Department Visits: A Systematic Review and Meta-Analysis. Med J Islam Repub Iran](#). (13 Jun); 40:61.
- [40] Hosseini Kasnavieh SM, Milanifard M, Sedighi M, Hessam R. (2026), [A Systematic Review of Pain Assessment Tools in Prehospital Emergency Care: Evidence from Observational Studies. Med J Islam Repub Iran](#). (10 Jun); 40:60.
- [41] Hessam R, Basir Ghafouri H, Rezai M, Shaker S H, Hosseini Kasnavieh S M, Milanifard M et al . (2026), [Hidden and Atypical Presentations of Myocarditis in Sudden Cardiac Death: A Narrative Review. Med J Islam Repub Iran](#); 40 (1) :526-536
- [42] Sajed M, Alvandifar S, Mallahi M. (2024), [Evaluation of Root Canal Morphology of Mandibular Premolars Using Cone-beam Computed Tomography in Golestan Province, North of Iran. Iran Endod J](#). 2024;19(3):183-188.
- [43] Fakhari, S, Bilehjani, A and Bilehjani, E. (2026). [Effect of Intraoperative Intravenous Dexmedetomidine on Oxidative Stress in Patients Undergoing Cardiac Surgery: A Systematic Review. Medicinal, Psychological, and Health Research Journal \(MPHRJ\)](#), 2(5), 351-359.
- [44] Fakhari, S, Bilehjani, A and Bilehjani, E. (2026). [Intravenous Dexmedetomidine in Cardiomyocyte Biology: A Systematic Review. Medicinal, Psychological, and Health Research Journal \(MPHRJ\)](#), 2(5), 360-371.
- [45] Lotfi, A R and Nouri Bayat, L. (2026). [Incidence of Postoperative Complications Following Nasal Fracture Surgery in Adults. Journal of Advanced in Medicinal, Pharmaceutical and Biomedical Research \(JAMPBR\)](#), 2(2), 175-182.
- [46] Rezaei, M and Owaysee Osquee, H. (2026). [Prevalence of Deep Vein Thrombosis in Patients with COVID 19 Admitted to the Intensive Care](#)

Unit. Medicinal, Psychological, and Health Research Journal (MPHRJ), 2(4), 267-275.

[47] Sadeghzadeh, A. (2026). [Evaluating the effectiveness and safety of hyaluronic acid versus poly-L-lactic acid for facial volume restoration: A Systematic Review and Meta-analysis](#). *Journal of Advanced in Medicinal, Pharmaceutical and Biomedical Research (JAMPBR)*, 2(2), 102-113.

[48] Shiri, H and Ashrafi, N. (2026). [Enhanced Recovery After Thoracotomy in the Intensive Care Unit: Current Evidence, Clinical Strategies, and Future Perspectives](#). *Medicinal, Psychological, and Health Research Journal (MPHRJ)*, 2(5), 372-382.

[49] Shiri, H and Ashrafi, N. (2026). [Post Esophageal Surgery Dysphagia and Nutritional Support in the Intensive Care Unit](#). *Medicinal, Psychological, and Health Research Journal (MPHRJ)*, 2(5), 339-350.

[50] Hashemloo, A., Milanifard, M. (2025). [Dermal Fillers: Types, Indications, and Complications Materials de Relleno: Tipos, Indicaciones Complicaciones](#). *Journal of Advanced in Medicinal, Pharmaceutical and Biomedical Research*, 1(6), 161-170

[51] Hashemloo, A., Milanifard, M. (2025). [Contouring Plus: A Comprehensive Approach of the Lower Third of the Face with Calcium Hydroxylapatite and Hyaluronic Acid](#). *Medicinal, Psychological, and Health Research Journal*, 1(5), 143-150

[52] Sunder Raj, K. S. (2004). [Multi-Pressure Surface Condenser Performance Evaluation: A Case Study](#). *Proceedings of the ASME Power Conference*.

[53] S Shaghagh, JK Udupa, Y Tong, Y Akhtar, M Hosseini, M Al-Noury, C Wu, L Xie, Y Hao, S Hassani, S Gogel, DM Biko, OH Mayer, JM McDonough, PJ Cahill, JB Anari, DA Torigian. (2025). [Association between respiratory volumes estimated from free-breathing dynamic MRI and sagittal spinal curvature in pediatric patients with thoracic insufficiency syndrome](#). *Medical Imaging 2025: Image-Guided Procedures, Robotic Interventions, and Modeling*. *Proceedings*, 13408

[54] M. Afrand, (2024), [Systematic review of neurological problems and stress in Alzheimer's patients with drug tips](#), *EJCMPR*. 3(5): 114-128.

[55] Motaqi, M., Ghanjal, A., & Afrand, M. (2022). [Treadmill exercise and its effect on rehabilitation of patients after ischemic stroke: A narrative study](#). *International Journal of Musculoskeletal pain prevention*, 7(3), 730-740.

[56] Afrand, M., Rostampur, N., Pourhadi, Z., Rezanezhad, J., & Nekouefard, P. (2025). [Epilepsy in children with central nervous system infections: A systematic review and meta-analysis](#). *AFMN Biomedicine*; 42(4):450-455.

[57] Ahmady H, Afrand M, Motaqi M, Meftahi GH. (2024). [Utilizing Sertoli Cell Transplantation as a Therapeutic Technique for the Management of Neurodegenerative Diseases](#). *Arch Razi Inst*.

1;79(4):701-710. PMID: 40256588; PMCID: PMC12004054.

[58] Motaqi, M, Afrand, M and Ghanjal, A. (2022). [Care and Management of Pain in Patients with Musculoskeletal Pain during the Covid-19 Epidemic](#). *International Journal of Musculoskeletal Pain Prevention*, 7(1), 622-626.

[59] Motaqi, M, Ghanjal, A and Afrand, M. (2022). [Treadmill Exercise and its Effect on Rehabilitation of Patients after Ischemic stroke: A Narrative Study](#). *International Journal of Musculoskeletal Pain Prevention*, 7(3), 730-740.

[60] Motaqi, M, Samadi, M, Ghanjal, A and Afrand, M. (2022). [A New Approach to Evaluate of Musculoskeletal Pain](#). *International Journal of Musculoskeletal Pain Prevention*, 7(4), 781-787.

[61] Moghadam, A. M., Yousef, R. M., Azizi, R. & Hessam, R. (2027). [Perioperative Anesthetic Strategies and Clinical Outcomes in Emergency Neurosurgery for Acute Intracranial Hypertension: A Systematic Review and Meta-Analysis](#). *Medicinal, Psychological, and Health Research Journal (MPHRJ)*, 3(1), 20-38.

[62] Moghadam, A. M., Yousef, R. M., Azizi, R. & Milanifard, M. (2027). [Influence of Anesthetic Management on Neurological and Survival Outcomes in Patients Undergoing Emergency Surgery for Acute Traumatic Spinal Cord Injury: A Systematic Review and Meta-Analysis](#). *Medicinal, Psychological, and Health Research Journal (MPHRJ)*, 3(1), 1-19.

[63] Moghadam, A. M., Yousef, R. M., Azizi, R., Hessam, R. & Milanifard, M. (2026). [Anesthetic and Perioperative Predictors of Neurological Deterioration and Mortality Following Emergency Craniotomy for Acute Brain Injury: A Systematic Review and Meta-Analysis](#). *Journal of Advanced in Medicinal, Pharmaceutical and Biomedical Research (JAMPBR)*, 2(6), 551-564.

[64] Moghadam, A. M., Yousef, R. M., Azizi, R., Hessam, R. & Milanifard, M. (2026). [Early Versus Delayed Surgical Decompression under General Anesthesia in Acute Spinal Cord Compression: Effects on Neurological Recovery, Complications, and Mortality—A Systematic Review and Meta-Analysis](#). *Journal of Advanced in Medicinal, Pharmaceutical and Biomedical Research (JAMPBR)*, 2(6), 537-550.

[65] Neishaboury, M., Shokri, S., Kianpour, P., Farhadi, K., Najjari, K., Sharifnia, H., MohammadYousef, R., & Khajavi, M. (2025). [The effects of intraperitoneal dexmedetomidine in comparison with ropivacaine in postoperative pain after laparoscopic sleeve gastrectomy: A double-blind, randomized, placebo-controlled, clinical trial](#). *Obesity Surgery*, 35(6), 2150–2159.

[66] Afrand, M., Rostampur, N., Pourhadi, Z., Rezanezhad, J., & Nekouefard, P. (2025). [Epilepsy in children with central nervous system infections:](#)

- A systematic review and meta-analysis. *AFMN Biomedicine*, 42(4), 450–455.
- [67] Azizi, Robab, & Rana Mohammad Yousef. (2027). [Dynamic Causal Modeling of Heart Rate–Blood Pressure Coupling During Induction of Anesthesia: A Systems-Based Approach to Predict Vasopressor Requirements.](#)
- [68] Afrand, M. (2024). [Systematic review: Systematic review of neurological problems and stress in Alzheimer's patients with drug tips.](#) *Eurasian Journal of Chemical, Medicinal and Petroleum Research*, 3(5), 114–128.
- [69] Samimi, A. (2025). [Assessment of Risks Arising from Neuropsychological Crises in Cardiac Patients Using FMEA.](#) *Journal of Advanced in Medicinal, Pharmaceutical and Biomedical Research*, 1(7), 196-203.
- [70] Samimi, A. (2025). [Risk Assessment in Hospitals Using the FMEA Method: A Data-Driven Analysis for Patient Safety Improvement.](#) *Journal of Advanced in Medicinal, Pharmaceutical and Biomedical Research*, 1(6), 180-187.
- [71] Samimi, A. (2025). [Application of Failure Mode and Effects Analysis \(FMEA\) for Risk Management in Hospital Laboratories.](#) *Medicinal, Psychological, and Health Research Journal (mphrj)*, 1(5), 160-168.
-