Published by
NIGERIAN SOCIETY OF PHYSICAL SCIENCES
Available online @ https://asr.nsps.org.ng/index.php/asr

African Scientific Reports 5 (2026) 634

African Scientific Reports

A comparative review of machine learning models and deep learning architectures for pregnancy progression prediction

Ogbaga Ignatius Nwoyibe, Agbo Kenechukwu Martin*

Department of Computer Science, David Umahi Federal University of Health Sciences, Uburu, Ebonyi State, Nigeria

Abstract

Monitoring the course of pregnancy is of paramount importance, enabling the identification of negative outcomes and clinically significant changes. Although parity status may contribute to information needs, status as first-time mothers is not considered as a causal determinant. This paper presents a structured scoping review comparing conventional and deep learning approaches to pregnancy-related prediction tasks. The reviewed machine learning models include logistic regression; random forest; Extreme Gradient Boosting (XGBoost); Recurrent Neural Networks (RNNs) and their gated variants, long short-term memory (LSTM) and Gated Recurrent Units (GRUs); Convolutional Neural Networks (CNNs); transformers; and hybrid architectures. The analysis employs six dimensions: performance; ability to encode temporal data; model interpretability; data requirements for training; computational costs; and clinical relevance. As primary studies report outcome-specific performance results, the findings are presented across outcome domains rather than aggregated rankings of model architectures. In a study on predicting preterm-birth outcomes, a transformer achieved a test-set Area Under the Receiver Operating Characteristic Curve (AUROC) of 0.792, compared to 0.78 for logistic regression, 0.751 for random forest, and 0.717 for a Support Vector Machine (SVM). In predicting preeclampsia, a CNN achieved an AUC of 0.85 on holdout validation and 0.81 on an external test set, with an AUC of 0.98 for early-onset preeclampsia. CTGFusionNet, which employs a hybrid architecture of convolutional, recurrent, and attention layers applied to Cardiotocography (CTG) and clinical records, reported 97.85% accuracy for the task of fetal-distress classification. The findings highlight the complementary strengths and weaknesses of conventional and deep learning models for the analyzed tasks. The choice of an architecture should be informed by the target outcome, data type, volume, and quality, calibration and validation strategies, requirements for interpretability and fairness, and computational costs rather than a singular emphasis on performance.

DOI: 10.46481/asr.2026.5.3.634

Keywords: Pregnancy progression, Machine learning, Deep learning, Preeclampsia, Preterm birth

Article History:
Received: 18 August 2026
Received in revised form: 25 September 2026
Accepted for publication: 28 September 2026
Available online: 13 October 2026

© 2026 The Author(s). Published by the Nigerian Society of Physical Sciences under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

1. Introduction

Pregnancy is a complex physiological and psychological process. First-time mothers may have questions not answered by their own past pregnancies, such as what physical changes are normal, warning signs or complications such as preeclampsia, gestational

*Corresponding Author Tel. No.: +234-706-602-2745.
Email address: kennygeemartin@gmail.com (Agbo Kenechukwu Martin)

diabetes, or preterm birth [1–3]. This review does not consider that such uncertainty may lead to obstetrical complications. Standard prenatal care consists of scheduled clinical evaluations and the use of population risk factors, which can miss inter-individual variations between visits [4, 5]. Digital surveillance and predictive modelling are therefore being considered as additional tools for more individualized maternal–fetal monitoring [6, 7]. Machine learning and deep learning approaches can analyse clinical, image, physiological, and multi-omics data to support pregnancies [8–10]. At the same time, traditional statistical models such as logistic regression, random forest, and XGBoost remain relevant because many pregnancy datasets are structured and tabular, and deep learning methods can be applied to temporal, image, signal, and multimodal data. RNNs, LSTMs, and GRUs are variants of recurrent networks designed for sequential data, CNNs have many applications in image and signal processing, Transformers employ self-attention mechanisms to learn long-range dependencies, and hybrid approaches combine multiple components [11–13]. In a personal health context, predictive models can aid in the early identification of obstetrical risks such as preeclampsia or preterm birth, interpretation of physiological cues, and mobile or wearable-based monitoring [5–8, 11, 12]. Such technologies should be viewed as clinical decision-support systems and not as a replacement for direct medical evaluation. Key challenges to widespread clinical adoption include data quality, calibration, interpretability, fairness, external validation, privacy, workflow integration, and the dangers of extrapolation from single-center or retrospective studies [1, 4, 10, 14]. In this review, "pregnancy progression" refers to the prediction and surveillance of clinically relevant events and outcomes across gestation, including hypertensive disorders of pregnancy, preterm birth, and fetal compromise. The main focus of this review is the comparison of deep learning architectures, but traditional machine learning methods were included as comparators where the cited pregnancy studies directly evaluated them. The comparison focuses on predictive performance, analysis of time-series data, interpretability, data requirements, computational costs, and deployment considerations. It is not intended to evaluate algorithmic mathematical foundations or rank approaches by a single metric of overall best performance across different datasets and clinical tasks.

2. Literature review

The literature included in this review serves two purposes. Firstly, the studies included in Sections 2.2 Digital-Health Context: Mobile Health and 2.3 Digital-Health Context: Wearables, IoT, and Continuous Monitoring serve to provide context for the emergence of predictive models in digital health by demonstrating how mobile health, wearable sensing, and continuous monitoring generate data and deployment environments in which predictive models can operate. These studies are not used to directly compare the performance of one machine learning or deep learning architecture versus another. Secondly, the studies included in Sections 2.4 Clinical Context: Hypertensive Disorders of Pregnancy and 2.5 Clinical Context: Pathophysiology serve to provide a clinical foundation for the comparison of predictive algorithms that appears in Section 3.

2.1. Review approach and scope

It is important to note that this article represents a narrative review rather than a PRISMA systematic review. The body of evidence included in this review consists of the pregnancy-monitoring, hypertensive-disorder, preterm-birth, fetal-distress, and related machine learning/deep learning studies mentioned in this article's main text which are available in the literature through 2026. The review focuses on comparing studies at the level of model families, tasks, data types, performance, validation setting, interpretability, data demands, computational considerations, and deployment contexts. Performance metrics are reported as presented in the original articles since the outcome measures, populations, preprocessing, model development, and validation designs were not consistent across studies. Additionally, protocol papers were separated from completed trials and are reported accordingly.

2.2. Mobile health applications for antenatal care

Mobile health interventions represent one approach to expanding antenatal care beyond the clinic or hospital in areas where access to care is limited. Pangestuti *et al.* [15] report the protocol for a cluster RCT of the REST mobile application in 22 antenatal class clusters, 22 midwives, and 220 pregnant women in Indonesia. The paper describes the intervention, trial design, recruitment strategy, trial schedule, and initial pilot/usability assessments but does not yet report any comparative evidence on intervention effectiveness, specifically reductions in anaemia, hypertension, pregnancy complications, or operative delivery, which were the outcomes of interest. Babatunde *et al.* [6] report the user-centered design of the MyBelle maternal health app for the Nigerian context. In the prototype testing phase, participants demonstrated increased preeclampsia knowledge following exposure to the app, and 95% of beta testers would recommend the app to others. Liu and Wang [16] evaluate a hospital-based mobile maternal health service and report improvements in pregnancy-related knowledge, prenatal examination compliance, adherence to follow-up, and rates of natural delivery compared to the control group. Khalil *et al.* [17] report the development of a web-based maternal health monitoring system with alert functionality. All of the above studies demonstrate the context in which predictive models can be deployed but do not report direct comparative evidence on the performance of different machine learning architectures.

2.3. Wearable sensors, IoT, and continuous monitoring systems

Wearable sensors, IoT, and continuous monitoring systems represent additional approaches to expanding antenatal care beyond the hospital. Borhade *et al.* [18] describe MomCare, an intelligent maternal monitoring concept that utilizes wearable sensors, IoT, artificial intelligence, and cloud services to monitor fetal heart rate, oxygen levels, fetal movement, and maternal vital signs and provide post-processing alerts and recommendations. Similar to the mobile health interventions described above, such continuous monitoring interventions generate data for predictive analytics but do not directly compare different machine learning architectures. Sarhaddi *et al.* [7] report on long-term maternal monitoring via smartphone, application, and cloud technologies with a focus on user retention and continuous physiological measurements. Kyambille *et al.* [2] carry out a scoping review of 77 studies on wearable sensors, wireless body sensor networks, IoT, and machine learning for maternal health monitoring in the developing world. This body of evidence highlights the opportunities in data collection and monitoring but does not report comparative findings on the performance of different predictive algorithms.

2.4. Machine learning and AI for hypertensive disorders of pregnancy (HDP)

Preeclampsia and other hypertensive disorders of pregnancy represent significant causes of morbidity and mortality for mothers and fetuses. Hu *et al.* [4] report a review of eHealth interventions for women with hypertensive disorders of pregnancy, whereas Escobar *et al.* [1] discuss challenges in the management of hypertensive disorders of pregnancy in low- and middle-income countries and highlight the lack of resources, diagnostics, medications, and health systems capacity in these regions. The context in which predictive algorithms can be applied is determined by the target population's needs. Several studies report on the development of machine learning algorithms for early prediction of preeclampsia or other hypertensive disorders of pregnancy.

Mizuno *et al.* [5] compare logistic regression, random forest, and XGBoost algorithms using lifestyle and environmental factors as predictors and report the best performance as an AUROC of 0.93. Schor *et al.* [9] describe a random forest model with an AUC of 0.73 in a nulliparous population, whereas Mousavi *et al.* [8] report an AUROC of 0.78 using routine clinical variables. Oku *et al.* [19] report an AUROC of 0.949 during development, 0.884 during testing, and 0.845 during prospective validation for the prediction of preeclampsia with home blood pressure monitoring. Chen *et al.* [20] report an external test set AUC of 0.843 for the prediction of preeclampsia using the XGBoost algorithm. Bülez *et al.* [21] report 90.6% accuracy for an artificial intelligence-based diagnostic model for preeclampsia. Butler *et al.* [11] describe the application of a modified ResNet convolutional neural network to 12-lead ECG recordings for the identification of women with preeclampsia with an AUC of 0.85 for the UTHSC holdout cohort and 0.81 for the external AHWFB cohort, whereas an AUC of 0.98 was reported for the early-onset preeclampsia subgroup.

2.5. Clinical definitions and pathophysiology of HDP

Tagetti and Fava [3] report a review of the diagnosis of hypertensive disorders of pregnancy. Preeclampsia is characterized by de novo hypertension that develops after the 20th week of gestation with blood pressure readings of at least 140/90 mmHg upon confirmation and signs of proteinuria or end-organ damage. Eclampsia is defined as new-onset seizure occurrence in women with preeclampsia after ruling out alternative causes. The validity of the predictions made by predictive algorithms depends on how well the target outcomes are defined. Layton [14] also emphasized the increasing importance of artificial intelligence (AI) and machine learning (ML) in the management of preeclampsia. However, while the review identified potential advantages of AI in early diagnosis, individualized risk prediction, individualized treatment optimization, and remote monitoring, there were many important considerations, such as data quality, the interpretability of models, and the process of integrating AI into clinical practice.

2.6. Augmented intelligence and outlier analysis for clinical discovery

In their paper, Janoudi *et al.* [10] investigated the use of outlier analysis, a technique of augmented intelligence, to discover previously unknown clinical patterns linked with preeclampsia and HDP. After applying the methods of isolation forest and extreme misclassification contextual outlier analysis, the study found that many clinically meaningful outliers were present in two large pregnancy datasets, in particular the extreme misclassification method proving to be a better performing method than isolation forest in these applications. Around 16% of the identified outliers were likely clinical discoveries for the first time, highlighting the benefits of AI-driven data exploration in advancing discoveries in maternal health.

3. Findings

The comparative results have been discussed in terms of model categorization, reported performance, the availability of temporal data, interpretability, requirements in terms of data and computational power, and potential clinical applications. The performance of ML models depends on the type of task and the particular design of the validation strategy. Thus, in the preterm-birth analysis by Zhang *et al.* [12], the transformer achieved an AUC of 0.792 on the test set, which was moderately better than the results of logistic regression, random forest, and support vector machine on the same dataset. In the preeclampsia analysis, the ECG CNN by Butler *et al.* [11] reached an AUC of 0.85 on the test set (UTHSC holdout cohort) and 0.81 on the external validation cohort (AHWFB), with a higher AUC of 0.98 for the early-onset subgroup. As reported by CTGFusionNet [13], accuracy in the task of fetal-distress

Table 1: Representative machine learning and deep learning models for pregnancy-related prediction.

Author	Model Architecture	Primary Application	Performance Metrics
Mizuno *et al.* [5]	Logistic Regression, Random Forest, XGBoost	Early prediction of hypertensive disorders of pregnancy using lifestyle and environmental data	Best AUROC = 0.93
Mousavi *et al.* [8]	Random Forest	Early prediction of hypertensive disorders of pregnancy using routine clinical data	AUROC = 0.78; Accuracy = 72%; F1-score = 0.76
Schor *et al.* [9]	Random Forest with Recursive Feature Elimination	Prediction of hypertensive disorders of pregnancy in nulliparous women	AUC = 0.73 (full); AUC = 0.71 (30-feat); Spec = 0.65
Oku *et al.* [19]	Logistic Regression using Home Blood Pressure Monitoring (HBPM)	Prediction of hypertensive disorders of pregnancy	AUC = 0.949 (dev), 0.884 (test); Sens = 76.9% (dev), 75.8% (val); Spec = 88.9% (dev), 76.6% (val)
Chen *et al.* [20]	XGBoost, Random Forest, AdaBoost, Gradient Boosting, Logistic Regression	Prediction of preeclampsia	Best (XGBoost): AUC = 0.930 (train), 0.763 (val), 0.843 (ext test)
Butler *et al.* [11]	Modified ResNet (CNN) using 12-lead ECG signals	Early prediction of preeclampsia	AUC = 0.85 (holdout), 0.81 (ext); Sens = 78% (holdout), 83% (ext); Spec = 84% (holdout), 66% (ext)
Zhang *et al.* [12]	Transformer, Logistic Regression, Random Forest, SVM	Preterm birth prediction	Transformer: AUC = 0.792, Sens = 73.67%, Spec = 72.48% (LR: 0.780, RF: 0.751, SVM: 0.717)
Aswathi Mohan *et al.* [13]	CTGFusionNet (CNN + LSTM + Attention), LW-FHRNet, ResRNN-Net, MultiScaleFusion-Net	Fetal distress prediction using fetal heart rate, uterine-contraction and clinical data	CTGFusionNet: Accuracy = 97.85%, Sens = 97.07%, Spec = 98.65%

classification reached 97.85%, based on the fusion of fetal heart rate, uterine-contraction, and clinical data. It is noteworthy that these results cannot be directly compared with each other since the papers analyze different outcomes and use different validation strategies.

3.1. Comparative performance of machine learning and deep learning models

This section presents the comparative findings identified from the comprehensive review process.

3.1.1. Predictive accuracy for pregnancy outcomes

The literature surveyed in Section 2 reports performance metrics for representative machine learning and deep learning models used in pregnancy-related prediction, as summarized in Table 1. Table 1 highlights the importance of considering an algorithm's performance in the context of a particular study. For instance, in Zhang *et al.* [12], on the test set, the transformer achieved the highest AUC (0.792), outperforming logistic regression (0.780), random forest (0.751), and support vector machine (0.717). In contrast, the results in the rest of the table concern different outcomes and therefore do not allow any direct comparisons. Specifically, the ECG-based ResNet of Butler *et al.* [11] reached an AUC of 0.85 on UTHSC holdout data and 0.81 in the external AHWFB cohort, while the 0.98 value is specific to the early-onset preeclampsia subgroup. Finally, CTGFusionNet achieved 97.85% classification accuracy

Model	Qualitative assessment	Interpretive note
RNN	◧ Moderate	Sequential recurrence; long-range dependencies can be difficult to learn.
LSTM	■ Strong	Gated memory supports longer-range temporal dependencies.
GRU	■ Strong	Gated recurrence with fewer gates than LSTM.
CNN	☐ Context-dependent	Standard CNNs emphasize local patterns; temporal variants can extend context.
Transformer	■ Strong	Self-attention links distant time points; data and compute needs are task-dependent.
Hybrid	☐ Context-dependent	Temporal capability depends on the components and fusion strategy.

Figure 1: Qualitative assessment of temporal dependency handling across model families.

for fetal distress using cardiotocography and clinical data [13], while the prospective home-blood-pressure model of Oku *et al.* [19] is another example of a task-specific statistical model that retains value due to the informative nature of the predictors.

3.1.2. Temporal dependency handling

The relative temporal-modelling characteristics of the reviewed model families are summarized qualitatively in Figure 1. Figure 1 shows a qualitative comparison of the architectures. A quantitative assessment is more involved, requiring evaluation on domain-specific tasks with representative data, appropriate outcomes, preprocessing, training procedures, and validation designs to obtain reliable performance estimates. The categories are descriptive of how the model families are generally applied to temporal data, but each model's prediction performance will depend on the data, outcomes, preprocessing, training procedure, and validation design. RNNs provide moderate temporal modelling because recurrent state supports sequential dependencies, although long-range patterns can be challenging to learn. LSTMs and GRUs provide strong temporal modelling; LSTM gating supports longer-range dependencies, while GRUs offer similar gated recurrence with fewer parameters. CNN temporal capability is context-dependent and can be extended through dilation or temporal specialization. Transformers provide strong temporal modelling through self-attention over sequence positions, depending on data scale and computational resources. Hybrid-model capability is context-dependent and depends on the components and fusion strategy. All four could represent temporal patterns, but the body of literature reviewed does not identify a single temporal dependency metric for each model. Temporal dependency seems to be a characteristic of an application, rather than a general property of a model class – a given model may be applicable to certain temporal data with particular patterns, preprocessing, and validation designs, but not others. For pregnancy progression, the critical consideration is whether a given model can operate on the available sequential, signal, image, or multimodal data and was validated for the particular clinical task. CTGFusionNet [13] uses convolutional, recurrent, and attention layers, but the inputs are cardiotocography (fetal heart rate and uterine-contraction signals) and perinatal clinical data; electrohysterography is not involved.

3.1.3. Interpretability and explainability

Table 2 presents a qualitative comparison of interpretability across model architectures, based on the availability of explainable AI (XAI) techniques and inherent model transparency. Table 2 highlights that interpretability does not equal predictive performance and that cross-study AUCs cannot define an accuracy-interpretability ranking. Logistic regression is comparatively transparent, while tree ensembles and deep models require post-hoc explanation tools. Attention maps can highlight where a model invests weight, but attention should not be treated as a validated explanation of causal model reasoning. For pregnancy-related decision support, explanation methods should therefore be subjected to stability, clinical plausibility, and consistency with model behavior checks prior to supporting clinician-patient communication.

3.1.4. Data efficiency and robustness

Figure 2 summarizes data-demand qualitative considerations across the reviewed model families without unsupported minimum sample-size thresholds. Figure 2 replaces fixed thresholds for sample size with qualitative comparisons. The categories are aids for

Table 2: Qualitative interpretability and explainability considerations.

Model	Typical Transparency	Common XAI Techniques	Clinical/Deployment Caveat	Example Application
Logistic Regression	High (coefficients inspectable)	Coefficient magnitude; calibration plots	Depends on correct model specification and calibration	HDP risk from HBPM (Oku *et al.*)
Random Forest	Moderate	SHAP, permutation importance	Global importance may not explain an individual case	HDP prediction (Schor *et al.*, Mousavi *et al.*)
XGBoost	Moderate	SHAP, gain, permutation importance	Post-hoc explanations require stability and clinical validation	PE prediction (Chen *et al.*)
LSTM	Low	SHAP, LIME; attention when explicitly added	Sequence explanations are post-hoc and may be unstable	Sequential clinical prediction
CNN	Low	Grad-CAM, saliency maps, SHAP	Saliency highlights are not causal explanations	ECG or imaging-based prediction
Transformer	Low	Attention maps, SHAP	Attention weights should not be equated with explanations	Sequence or multimodal integration
Hybrid (with attention)	Low to moderate	Attention visualization, post-hoc XAI	Attention indicates model focus, not causal reasoning	CTGFusionNet (fetal distress)

Model family	Relative data demand	Main qualifier
Logistic Regression	☐ Lower	Often suitable for smaller tabular datasets when predictors are well specified.
Random Forest / XGBoost	◧ Low to moderate	Requirements vary with feature dimensionality and class balance.
LSTM / GRU	◧ Moderate to high	Depends on sequence length, event density, missingness and regularization.
CNN	◧ Moderate to high	Depends on image/signal complexity; transfer learning can reduce data needs.
Transformer	■ Often high	Task-dependent; pretraining or compact models may reduce data demand.
Hybrid	■ Often high	Depends on constituent models, number of modalities and fusion strategy.

Figure 2: Qualitative data-demand considerations across model families.

reviews only and should not be seen as minimum recruitment targets or learning-curve estimates. The data requirements vary with respect to the task, the modality, the class prevalence, the number of predictors, the model size, the amount of regularization, pretraining, and the validation design. Logistic regression generally has relatively low data demands for well-specified tabular problems. Random forest and XGBoost have low-to-moderate relative data demand, depending on feature dimensionality and class balance. LSTM and GRU models have moderate-to-high relative data demand, depending on sequence length, event density, missingness, and regularization. CNNs also have moderate-to-high relative data demand, although transfer learning can reduce the amount of task-specific labelled data required. Transformers often have high relative data demand, although pretraining and compact architectures can reduce requirements for some tasks. Hybrid models often have high relative data demand because requirements depend on constituent models, modalities, and the fusion strategy. Data requirements have important implications for low-resource settings, but they cannot be reduced to universal cutoffs. Logistic regression and tree-based models may be practical for smaller tabular datasets; recurrent models, CNNs, transformers, and hybrids can require more data, depending on the architecture size, the modality, the missingness, the class balance, and the regularization. Transfer learning and pretraining may reduce the data requirements for some CNN and transformer applications. In Zhang *et al.* [12], the preterm-birth dataset contains 30,965 births, and the evaluated deep model is a transformer, not an LSTM. Therefore, the study does not suggest a previously stated larger LSTM cohort or a universal

Table 3: Qualitative computational cost and deployment considerations.

Model	Training Time	Inference Speed	Memory Footprint	Edge/Wearable Feasibility	Real-Time Capability
Logistic Regression	Very low	Very fast	Very small	Generally feasible	Generally feasible
Random Forest	Low–mod	Fast	Small–mod	Often feasible; size-dependent	Often feasible
LSTM	Moderate	Moderate	Moderate	Feasible with compression	Sequence/hardware dependent
GRU	Moderate	Mod–fast	Moderate	Feasible; smaller than LSTM	Hardware-dependent
CNN	Mod–high	Fast (accelerated)	Moderate	Feasible for lightweight models	Model/hardware dependent
Transformer	High	Mod–slow	Large	Requires compact models/cloud	Context-dependent
Hybrid	High	Moderate	Large	Depends on components	Context-dependent

minimum sample size for LSTM training.

3.1.5. Computational cost and deployment feasibility

Table 3 compares computational requirements and deployment suitability for real-time or edge-based monitoring. For continuous or home-based monitoring, the deployment is feasible given the model size, input preprocessing, memory, latency, energy budget, and the target device. Recurrent and convolutional models can sometimes be compressed or quantized for edge deployment, while transformer deployment depends heavily on model scale and available hardware. These are relative considerations from an engineering perspective rather than hard constraints on the target hardware. A recent study using an LSTM autoencoder for ECG anomaly detection from African Scientific Reports provided an example of the general feasibility of recurrent modelling for physiological time-series data [22]; however, that study is outside the scope of pregnancy and should not be considered as an obstetrical validation.

3.2. Application-specific findings for first-time mothers

3.2.1. Preterm birth risk stratification

Zhang *et al.* [12] assessed the performance of transformer, logistic regression, random forest, and support vector machine models in predicting preterm birth; no LSTM model was assessed. On the test set, the transformer reached an AUC of 0.792, sensitivity 73.67%, specificity 72.48%, PPV 28.21%, and negative predictive value (NPV) 94.95% – findings supporting the need for further exploration into the value of a transformer-based risk stratification but not establishing a 'most validated' architecture. In a real-world setting, a given model's risk estimation would need to be followed by a suitable clinical reassessment or a relevant additional test, with the selection of a particular treatment such as progesterone to prevent preterm birth dependent on the physician's judgment and the patient's eligibility criteria. PPVs are of specific interest in the case of an outcome of low prevalence. As an example, if the prevalence was as high as 11% and both sensitivity and specificity were 85%, the PPV would still be only 41%. This is a hypothetical example, not related to any of the studies discussed in this paper. Calibration using local data and careful choice of thresholds, cost-sensitive or class-imbalance-aware learning, and recalibration of predicted probabilities can all aid the development of useful decision-making support systems, but resampling does not eliminate the need to condition predicted probabilities on the characteristics of the target population. Related issues are further discussed in Recent Advances in Natural Sciences, where similar challenges in addressing class imbalance in the context of epidemiological machine learning and the relevance of distinguishing between methodological and clinical evaluation were highlighted [23].

3.2.2. Hypertensive disorders of pregnancy (HDP) prediction

Multiple models reviewed in Section 2 demonstrated useful predictive performance for HDP, using a range of input data: Lifestyle and questionnaire data [5] yielded an AUROC of 0.93 using self-reported data at 20 weeks; this approach requires no specialized equipment and can be administered via a mobile app. Home blood pressure monitoring [19] yielded an AUROC of 0.845 in prospective validation, with a negative predictive value of 0.957. For ECG-based detection [11], the modified ResNet achieved AUCs of 0.85 in the UTHSC holdout cohort and 0.81 in the external AHWFB cohort; an AUC of 0.98 was reported only for the early-onset preeclampsia subgroup, while 30-, 60-, and 90-day prediagnosis analyses yielded AUCs of 0.92, 0.89, and 0.90, respectively. These findings support further evaluation of ECG-based risk assessment but do not establish a validated primary-care screening pathway.

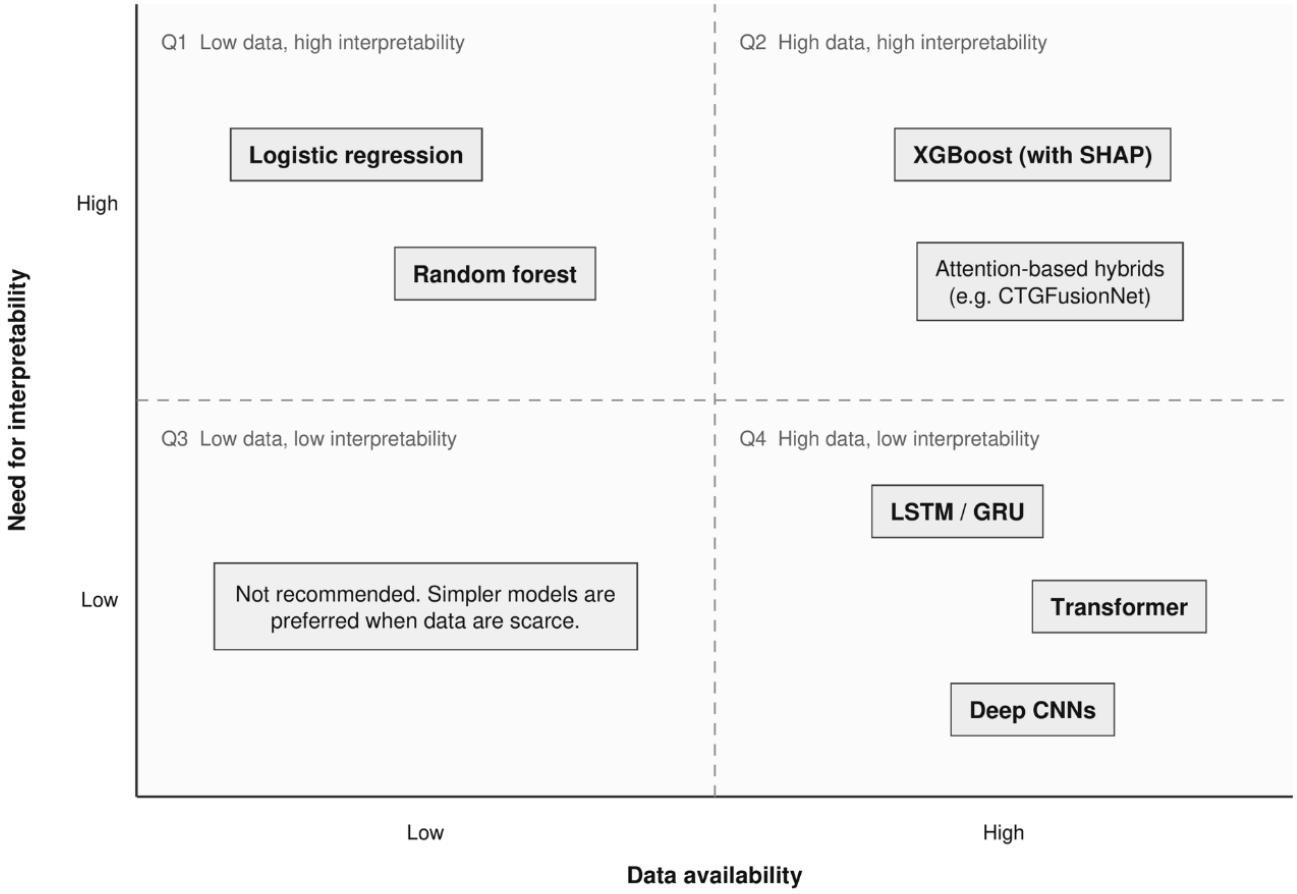


Figure 3: Conceptual model selection decision matrix for pregnancy prediction.

These approaches should be interpreted as separate lines of evidence. The reviewed studies did not test a combined questionnaire, home blood pressure, and ECG screening pathway, so no such sequence is recommended on the basis of this review.

3.2.3. Fetal distress detection during labor

CTGFusionNet [13] achieves 97.85% accuracy, 97.07% sensitivity, and 98.65% specificity in fetal-distress classification from fetal heart rate, uterine-contraction, and perinatal clinical data, which is an excellent result for the task and the dataset. However, the model was not evaluated in terms of decreased cesarean delivery or improved neonatal and maternal outcomes, which limits the scope of its immediate applicability. Such outcome improvement needs to be validated prospectively in an external cohort before the model is declared clinically applicable. The lighter comparison models may be useful in low-resource settings, but their efficacy has not yet been validated on the proposed hardware and population.

3.3. Trade-off synthesis and model selection framework

Based on our discussion, Figure 3 preserves the model-selection matrix from the manuscript under the same author disclaimer. We emphasize that it synthesizes considerations for the availability of data and interpretability features but is not itself a validated clinical decision rule or an ordinal ranking of proven clinical benefit. In Figure 3, “Data Availability” (Low to High) is plotted on the x-axis and “Need for Interpretability” (Low to High) on the y-axis. The quadrants indicate qualitative considerations, and the reader should interpret these as discussion prompts rather than definitive thresholds derived from empirical analysis: Quadrant 1 (low data, high interpretability) suggests that logistic regression and random forest can be reasonable choices for structured clinical data when simpler or more interpretable models are preferred. Quadrant 2 (high data, high interpretability) suggests XGBoost with SHAP and hybrid models with explanation modules for richer data, while noting that explanation quality still requires evaluation. Quadrant 3 (low data, low interpretability) indicates that simpler models may be preferable when data are scarce because they are also easier to validate. Quadrant 4 (high data, low interpretability) includes LSTM, GRU, transformer, and deep CNN models in

research or data-rich settings, subject to task-specific justification, human-factors considerations, and clinical validation. For prenatal decision support, the model should be determined on a case-by-case basis considering the task, data, calibration, external validation, interpretability, fairness, and computational requirements. Both tree-based models with SHAP or validated deep models could be good options depending on the circumstances. Therefore, one should not declare a single superior architecture for prenatal tasks but rather proceed to a discussion using this matrix.

4. Conclusion

The review compared standard machine learning and deep learning approaches to prediction tasks related to pregnancy, including preterm birth, hypertensive disorders, fetal distress, and care contexts for primigravida. The data do not support a specific architecture's overall superiority: in the same-dataset comparison described by Zhang *et al.* [12], the transformer indeed achieved the best performance (AUC of 0.792 on the test set), but it should be compared only to the other three models and should not be compared directly with values from studies of different outcomes. The AUC values of 0.85 (hold-out) and 0.81 (external) reported by Butler *et al.* [11] for ECG-based prediction of preeclampsia are good but not better than the 0.98 value for early-onset cases. Finally, CTG-FusionNet's 97.85% accuracy for fetal-distress classification is a strong result for a proof-of-concept study, but it does not directly relate to the improvement of patient outcomes. Second, interpretability and validity considerations do not necessarily favor complex architectures. Linear/logistic regression and tree-based methods can be good baselines in specific cases due to their simplicity and inherent interpretability, but more complex methods often require additional steps for trustworthy use. Attention mechanisms can highlight regions of the input that contribute most to a given prediction, but such evidence is less informative than direct clinical relevance. Third, while some works demonstrated data efficiency and feasibility of use in low-resource settings, these findings were task-specific and dependent on implementation details. The choice of the most appropriate method should consider the available data, computational resources, evaluation strategy, and other factors. For this reason, the suggested sample size thresholds should not be considered strict criteria for a prediction task's minimum requirements, and the authors' wording of "clinically applicable" for some methods might be overly optimistic. Finally, external validity, fairness, calibration, and clinical utility were essential aspects that most works did not adequately address. The reviewed articles primarily used retrospective and non-interventional designs, were single-center, and used patient-reported or surrogate outcomes. Individualized pregnancy management using machine learning or deep learning is a promising direction, but clinicians should carefully evaluate such models. For primigravida, the critical value of the reviewed methods would be the ability to provide individual risk scores estimated consistently across populations and meaningfully integrated into a broader clinical context. Only additional research on such models' performance in diverse populations and their cost-benefit in different settings will determine whether these methods will be valuable additions to the current care paradigm.

4.1. Open research areas for further studies

Although the reviewed studies demonstrate promising applications of machine learning and deep learning, there are a number of research questions that remain. The following topics are suggested for future investigation.

4.1.1. Multimodal integration and foundation models

The majority of existing models only consider one or two modalities (e.g., EHR and an imaging modality such as ultrasound). Additional work is needed to develop a foundational model of pregnancy that can incorporate heterogeneous data streams, such as genomics, proteomics, metabolomics, wearable sensor data, social determinants of health, and patient-reported outcomes.

4.1.2. Longitudinal fairness and continuous monitoring

Fairness evaluation is often conducted at model-development time, but shifting demographics and evolving clinical practices can impact subgroup performance after deployment. Future work should focus on techniques for continual fairness monitoring and detection of drifts in subgroup performance, with appropriate interventions such as recalibration, updating, or fairness-aware retraining.

4.1.3. AI for low-resource and remote settings

Much of the reviewed research originates from high-income countries or well-resourced urban hospitals. There is a need to develop more context-aware algorithms that can operate in low-resource settings such as rural clinics, and utilize community health-worker workflows and intermittent-connectivity paradigms. Key research directions include: lightweight models for resource-constrained devices, federated learning across smaller clinics (rather than aggregating patient data), and culturally attuned approaches to explanation and risk communication.

4.1.4. Personalized risk communication and behavioral impact

A risk score alone does not improve outcomes, and future studies are needed to investigate how to best communicate individualized probabilities to first-time mothers, and how risk scores might influence anxiety, behavior, adherence to medication and care-seeking practices. Marrying predictive AI to behavioral science and human-computer interaction principles can help develop user-centered, effective risk-communication interfaces.

4.1.5. Real-time closed-loop monitoring and alert systems

The majority of remote monitoring systems currently operate by relaying alerts to clinicians. Future work could investigate protocol-governed closed-loop support, such as automatically escalating telehealth triage if validated longitudinal models identify a persistent change in risk. Any automated escalation should be prospectively evaluated, audited, and designed in such a way that treatment remains under appropriate clinical governance. First-time motherhood represents a major life transition that can trigger demand for trusted information and timely support. Responsible machine learning and deep learning can complement clinical care by enabling risk stratification, longitudinal monitoring, and clinical decision support. The central message of this review is that no single architecture is universally applicable, and that model design should be informed by considerations of the task, data modalities, calibration, external validation, interpretability, fairness, and computational constraints. Making progress on these topics will help realize the potential of predictive modelling to improve safety and utility of maternal health care without overreaching beyond the evidence.

Data availability

Data sharing is not applicable to this article. No new data were created or analyzed in this study. All information discussed in this review is available from the cited published sources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this manuscript.

Funding

The authors received no specific funding from any public, commercial, or not-for-profit organisation for this work.

References

[1] M. F. Escobar, N. Benitez-Díaz, I. Blanco-Londoño, C. Cerón-Garcés, E. E. Peña-Zárate, L. A. Guevara-Calderón, L. Libreros-Peña & J. S. Galindo, "Synthesis of evidence for managing hypertensive disorders of pregnancy in low middle-income countries: a scoping review", BMC Pregnancy and Childbirth **24** (2024) 622. https://doi.org/10.1186/s12884-024-06796-2.

[2] G. G. Kyambille, A. N. Mvuma & D. Machuve, "An overview of maternity healthcare monitoring in developing nations", Tanzania Journal of Engineering and Technology **42** (2023) 46. https://doi.org/10.52339/tjet.v42i4.866.

[3] A. Tagetti & C. Fava, "Diagnosis of hypertensive disorders in pregnancy: an update", Journal of Laboratory and Precision Medicine **5** (2020) 8. https://doi.org/10.21037/jlpm.2019.11.04.

[4] H. Hu, N. Noori, V. Lee, C. Chow, N. W. Cheung, K. Ekanayake & M. Zen, "eHealth and hypertensive disorders of pregnancy: systematic review", Journal of Medical Internet Research **27** (2025) e77064. https://doi.org/10.2196/77064.

[5] S. Mizuno, S. Nagaie, J. Sugawara, G. Tamiya, T. Obara, M. Ishikuro, S. Kuriyama, N. Yaegashi, H. Tanaka, M. Yamamoto & S. Ogishima, "Early prediction of hypertensive disorders of pregnancy toward preventive early intervention", AJOG Global Reports **4** (2024) 100383. https://doi.org/10.1016/j.xagr.2024.100383.

[6] A. O. Babatunde, A. I. Ayede, A. Colangelo, T. D. Nguyen, A. T. Aborode, C. Umeh, M. P. Hernandez, O. I. Ayede & O. O. Ayede, "Development of an educational and monitoring mobile application for pregnant women in Nigeria", Frontiers in Public Health **12** (2025) 1368631. https://doi.org/10.3389/fpubh.2024.1368631.

[7] F. Sarhaddi, I. Azimi, S. Labbaf, H. Niela-Vilén, N. Dutt, A. Axelin, P. Liljeberg & A. M. Rahmani, "Long-term IoT-based maternal monitoring: system design and evaluation", Sensors **21** (2021) 2281. https://doi.org/10.3390/s21072281.

[8] S. S. Mousavi, K. Tierney, C. Robichaux, S. L. Boulet, C. Franklin, S. Chandrasekaran, R. Sameni, G. D. Clifford, N. Katebi & S Nikookar, "Early prediction of hypertensive disorders of pregnancy using machine learning and medical records from the first and second trimesters", Health **2** (2026) 025015. https://doi.org/10.1088/3049-477X/ae9e34.

[9] J. S. Schor, A. Kadambi, I. Fulcher, K. K. Venkatesh, M. A. Clapp, S. Ebrahim, A. Ebrahim & T. Wen, "Using machine learning to predict the risk of developing hypertensive disorders of pregnancy using a contemporary nulliparous cohort", AJOG Global Reports **4** (2024) 100386. https://doi.org/10.1016/j.xagr.2024.100386.

[10] G. Janoudi, D. B. Fell, J. G. Ray, A. M. Foster, R. Giffen, T. J. Clifford, M. A. Rodger, G. N. Smith & M. C. Walker, "Augmented intelligence for clinical discovery in hypertensive disorders of pregnancy using outlier analysis", Cureus **15** (2023) e36909. https://doi.org/10.7759/cureus.36909.

[11] L. Butler, F. Gunturkun, L. Chinthala, I. Karabayir, M. S. Tootooni, B. Bakir-Batu, T. Celik, O. Akbilgic & R. L. Davis, "AI-based preeclampsia detection and prediction with electrocardiogram data", Frontiers in Cardiovascular Medicine **11** (2024) 1360238. https://doi.org/10.3389/fcvm.2024.1360238.

[12] F. Zhang, L. Tong, C. Shi, R. Zuo, L. Wang & Y. Wang, "Deep learning in predicting preterm birth: a comparative study of machine learning algorithms", Maternal-Fetal Medicine **6** (2024) 141. https://doi.org/10.1097/FM9.0000000000000236.

[13] P. P. Aswathi Mohan, V. Uma, R. Sasirekha & V. Hamsika, "CTGFusionNet: fusion of deep learning models for predicting fetal distress: a multimodal approach", Health Information Science and Systems **13** (2025) 66. https://doi.org/10.1007/s13755-025-00383-1.

[14] A. T. Layton, "Artificial intelligence and machine learning in preeclampsia", Arteriosclerosis, Thrombosis, and Vascular Biology **45** (2025) 165. https://doi.org/10.1161/ATVBAHA.124.321673.

[15] R. Pangestuti, P. D. Ratrikaningtyas & A. H. Sutomo, "Monitoring of pregnant women using the 'Risk Identification, Evaluation Counseling, Systematic Monitoring, Troubleshooting' (REST) mobile app: protocol for a cluster randomized controlled trial", JMIR Research Protocols **14** (2025) e66774. https://doi.org/10.2196/66774.

[16] Y. Liu & X. Wang, "Application of smart mobile medical services in maternal health care management", Contrast Media & Molecular Imaging **2021** (2021) 6249736. https://doi.org/10.1155/2021/6249736.

[17] A. Khalil, A. L. Ahmad Sunawari, S. Ismail & A. S. Abdul Jamil, "Development of a web-based maternal health monitoring system: advancing beyond manual record-keeping", in Biomedical Engineering, K. W. Lai & P. S. Q. Yeoh (Eds.), Springer, Singapore, 2024, pp. 1–18. https://doi.org/10.1007/978-981-97-9294-8-1.

[18] S. Borhade, G. Doke, P. Bhujbal & S. Said, "MOM CARE: leveraging intelligent systems for comprehensive monitoring and support during pregnancy", International Journal of Recent Advances in Engineering and Technology **141s** (2025) 330. https://doi.org/10.65521/intjournalrecadvengtech.v14i1s.776.

[19] M. Oku, S. Ohno, H. Metoki, E. Arai, M. Ito, A. Shiozaki, A. Nakashima, N. Iwama, R. Tsuchida, S. Yoneda, N. Yoneda, S. Tsuda, M. Satoh, T. Murakami, S. Tanigawara, K. Hoshi, K. Tanaka, M. Yamada & S. Saito, "Prediction model of hypertensive disorders of pregnancy based on home blood pressure monitoring", Journal of the American Heart Association **15** (2026) e047093. https://doi.org/10.1161/JAHA.125.047093.

[20] Y. Chen, X. Shi, Z. Wang & L. Zhang, "Machine learning-based management of hypertensive disorders in pregnancy: analysis of differences in key risk factors between gestational hypertension and pre-eclampsia and construction of a pre-eclampsia prediction model", European Journal of Medical Research **30** (2025) 1135. https://doi.org/10.1186/s40001-025-03407-4.

[21] A. Bülez, K. Hansu, E. S. Çağan, A. R. Şahin & H. Ö. Dokumacı, "Artificial intelligence in early diagnosis of preeclampsia", Nigerian Journal of Clinical Practice **27** (2024) 383. https://doi.org/10.4103/njcp.njcp-222-23.

[22] A. Adegoke-Elijah, T. O. Ojewumi & K. O. Jimoh, "ECG anomaly detection: a deep learning perspective with LSTM encoders", African Scientific Reports **4** (2025) 359. https://doi.org/10.46481/asr.2025.4.3.359.

[23] O. Njama-Abang, D. U. Ashishie, P. T. Bukie & A. I. Bassey, "A methodological framework for evaluating ADASYN and borderline-SMOTE oversampling techniques in imbalanced epidemiological data: a proof-of-concept study for Lassa fever detection", Recent Advances in Natural Sciences **4** (2026) 238. https://doi.org/10.61298/rans.2026.4.1.238.