ECRONICON OPEN ACCESS

EC PAEDIATRICS

Research Article

Development and Evaluation of AI (Artificial Intelligence) Models for Predicting Preterm Birth

Kishore Kumar R[1,2]*, Arvind Kasargod[3], Rashmi Chaudhary[4], Shalini Agarwal[5] and Yazdan Khan[6]

[1]Cloudnine Hospital, Jayanagar Bangalore, Karnataka, India

[2]Notre Dame University, Perth, Australia

[3]Group Medical Director and Pediatric Intensivist, Cloudnine Hospital, HRBR Layout, Bangalore, Karnataka, India

[4]Consultant Obstetrician and Gynaecologist, Cloudnine Bellandur, Bangalore, Karnataka, India

[5]Consultant ObGyn, Cloudnine Hospital, Patparganj, Delhi, India

[6]Chief Data Scientist, Actify Data Labs Pvt Ltd, Bangalore, Karnataka, India

***Corresponding Author:** Kishore Kumar R, Cloudnine Hospital, Jayanagar, Bangalore, Karnataka, India.

Received: April 03, 2025; **Published:** April 14, 2025

DOI: 10.31080/ECPE.2025.14.01573

Abstract

Objective: The objective of this study was to develop and evaluate various machine learning algorithms for predicting preterm birth, aiming to overcome the limitations of conventional approaches in terms of reliability and predictive power.

Study Design: This study utilized health records from 5900 patients to train and assess different machine learning algorithms. The focus was on identifying patterns and predicting outcomes related to preterm birth, with an emphasis on the interpretability and generalizability of the models.

Result: Among the various algorithms evaluated, CatBoost emerged as the most effective, achieving an accuracy of 84%. This performance surpassed that of predictions based on manual reviews. Key factors contributing to the model's predictive ability were also identified, highlighting its potential in clinical applications.

Conclusion: The findings demonstrate the efficacy of machine learning, particularly the CatBoost algorithm, in predicting preterm birth. This suggests a significant advancement over traditional methods, offering a more reliable tool for healthcare professionals in anticipating and managing preterm birth complications.

Keywords: *Premature Birth; Machine Learning; Infant, Newborn; Pregnancy*

Abbreviations

CHN: Cloudnine Hospital Network; AUROC: Area Under the Receiver Operating Characteristics Curve

Introduction

Despite the advances in public health and medicine, global preterm birth rates have not changed in the past decade [1]. Accounting for one in five of all deaths occurring in children less than 5 years of age, preterm birth is a leading cause of child mortality. In addition to the increased risk of mortality, preterm births also cause early childhood impairments, such as cerebral palsy, intellectual disability, and sensory impairments [2]. Furthermore, high rates of dysfunction in visual processing, cognitive development, academic progress, and executive function have been reported in later childhood. Such cognitive disadvantages may persist until late adolescence and early adulthood. Children born prematurely may also be at an increased risk of developing diseases such as diabetes, cardiovascular and cerebrovascular diseases, hypertension, chronic kidney disease, and asthma later in life [3,4].

The complexity of preterm birth onset and its association with multiple etiologic pathways pose barriers to investigating this phenomenon [5]. The underlying etiology of preterm birth remains unclear, with much variability depending upon race, ethnicity, and geographical boundaries [6,7]. Several maternal characteristics, ultrasound markers, and biomarkers have been studied in this context [7]. Risk factors such as race, age (< 18 years or > 40 years), low socioeconomic status, lack of prenatal care, smoking and narcotic use, urinary or lower genital tract infection, high levels of stress, and anemia have been implicated as causes of preterm birth [8]. Higher rates of preterm birth have been reported among mothers with comorbidities, such as hypertension and diabetes, as compared to those without comorbidities [9]. Understanding the mechanisms by which such factors lead to premature delivery can facilitate the development of effective preventative and treatment strategies.

Effective interventions for the prevention of preterm birth may depend upon several characteristics, making it essential for care providers to identify individual risks [7,10]. There is a need for more research to develop approaches aimed at identifying candidates who may benefit the most from the available prophylactic interventions [7]. Clinical risk prediction models aim to assist in medical decisions by providing objective measures of potential health outcomes based on data [11].

Maternal health records, which are valuable sources of routinely collected obstetric and medical information, can be used to gain predictive insights on preterm birth. Although a few predictive systems have used demographic, medical, and obstetric data for this purpose, they were limited by their low predictive power [12]. Such limitations may be due to the complexity of preterm birth, where the use of simple linear statistical models may not provide optimal results.

Machine learning approaches are promising tools for risk prediction beyond traditional models. Such algorithms are widely used in healthcare research and applications [13]. Given the observed variability and complexity of preterm birth etiology, it may be helpful to leverage machine learning methods to model the complex relationships between various maternal factors and the risk of preterm birth. This study aimed to investigate the applicability of machine learning techniques for the prediction of preterm birth using patient records. Data were collected from patient records spanning ultrasound scan reports, laboratory reports, and vital signs. Different tree-based algorithms were evaluated and compared for their utility in predicting preterm birth. Of these, CatBoost outperformed all other algorithms and resulted in a prediction accuracy of 84%.

Methods

Data collection: The data used in this study were gathered from a diverse cohort of 5900 patients who had undergone childbirth over a span of 16 months at the Cloudnine Hospital Network (CHN). The data was collected from the hospital database and medical reports and included information on ultrasound scan reports, laboratory reports on blood and urine samples, and measurements of the patient's vital signs (Table 1). The reports from the ultrasound scans provided anatomical and physiological aspects of the pregnancy. The laboratory reports provided information on biochemical parameters in blood and urine samples. The vital signs included information such as blood pressure, heart rate, and body temperature, among other measurements. Following the extraction of data, the records were verified for correctness and completeness. The collected data included features encompassing physiological, biochemical, and clinical parameters. The parameters were selected based on their potential relevance, as reported in the literature [14-16]. Data was collected for the indicated parameters for all trimesters.

Physiological and physical parameters	
Abdominal circumference	This measure provides an estimate of fetus size and can indicate whether the fetus is growing at a normal rate.
Cervix length	The length of the cervix can change during pregnancy and may be an indicator of preterm labor.
Current pregnancy week	The gestational age at the time of data collection can provide context for other measurements and is a critical factor in determining the likelihood of preterm birth.
Estimated fetal weight	This is another measure of fetal growth and development.
Number of fetuses	Multiple pregnancies have a higher risk of preterm birth, making this an important feature.
Fetal humerus and trans cerebellar diameter	These measurements can provide information about fetal growth and development.
Amniotic fluid index	This measure of the amount of amniotic fluid can indicate conditions like polyhydramnios or oligohydramnios, which can increase the risk of preterm birth.
Placenta	The position and health of the placenta can impact the risk of preterm birth.
Height and weight	The mother's height and weight can influence the risk of preterm birth. The derived feature, body mass index, was also included.
Systolic and diastolic blood pressure	High blood pressure can indicate conditions like preeclampsia, a risk factor for preterm birth.
Mean pulsatility index	Used in Doppler ultrasound to assess the resistance to blood flow in vessels for evaluating fetal and placental health
Biochemical and hematological parameters	
Mean corpuscular hemoglobin concentration, mean corpuscular volume, and hemoglobin	These are measures of red blood cell health and can indicate conditions like anemia.
Albumin/globulin ratio, alkaline phosphatase, bilirubin (direct, indirect, and total), and total serum protein	These are measures of liver function, which can be affected during pregnancy.
Monocytes, basophils, neutrophils, lymphocytes, and white blood cell count	These measures of immune function can be altered during pregnancy and impact the risk of preterm birth.
HbA_{1c} and random blood sugar	These measures of blood sugar control can indicate gestational diabetes, a risk factor for preterm birth.
T3, T4, and TSH	These thyroid hormones can impact pregnancy outcomes, with both hypo- and hyperthyroidism associated with an increased risk of preterm birth.
Urine pH, pus cells in the urine, and urine culture	These measures of urinary health can indicate urinary tract infections, which can increase the risk of preterm birth.
Urine specific gravity	Urine specific gravity can provide insights into the hydration status and kidney function of the mother.
HbA_{1c}: Hemoglobin A_{1c}; T3: Triiodothyronine; T4: Thyroxine; TSH: Thyroid Stimulating Hormone.	

Table 1: *Parameters considered for the prediction model.*

This study was approved by the Institutional Ethics Committee of the CHN. Written informed consent was obtained from all subjects before the study was initiated. The identities of the subjects were kept confidential, and anonymized data was used for all analyses.

Data preprocessing and model selection: The raw data was preprocessed to create a clean and normalized dataset suitable for training the machine learning model. Issues related to missing values and outliers were addressed. For numerical features, mean- or median-based imputation was typically used, while for categorical features, mode imputation was the method of choice. Outliers, which could potentially skew the model's performance, were identified using statistical methods, such as the Z-score or the interquartile range. Outliers were either corrected or removed depending on the nature of the outlier and the specific feature. Following this step, data normalization, encoding of categorical variables, and feature engineering were carried out. Tree-based machine learning algorithms were chosen due to their inherent ability to handle missing values and sparse data and capture complex non-linear relationships [17-19]. The feasibility of several tree-based algorithms for predicting preterm birth was explored. These algorithms included Random Forest, Gradient Boost, AdaBoost, XGBoost, and CatBoost [19-23]. The dataset was randomly divided into training and test sets. Five-fold cross-validation was used for tuning hyperparameters for optimizing model performance. The comparison of model performance was based on the ability to distinguish between preterm and full-term births. Standard performance metrics such as accuracy, precision, recall, F1 score, and area under the receiver operating characteristics curve (AUROC) were calculated for each model.

Evaluation: Based on the performance metrics, the CatBoost algorithm was chosen. The trained model was evaluated on an independent data set of 40 patients, with data collected up to the 30th week of pregnancy. At the time of analysis, these patients had still not given birth. The patients were followed up to determine if any of the deliveries were preterm. The model predictions and those made by two physicians were compared to the actual outcomes of preterm births. The model predictions were also compared with those made by the two physicians.

Results

Study population: The dataset used in this study was collected from a cohort of 5900 participants over a period of 16 months.

Comparison of model performances: Based on performance metrics, CatBoost outperformed the other tree-based models. The results are summarized in table 2. CatBoost was selected for further validation on a live dataset with undetermined pregnancy results at the time of analysis.

Model	Accuracy	Precision	Recall	F1-score	AUROC
Random Forest	0.82	0.81	0.80	0.79	0.6
Gradient Boost	0.78	0.79	0.78	0.77	0.59
AdaBoost	0.77	0.75	0.74	0.76	0.57
XGBoost	0.83	0.82	0.83	0.78	0.62
CatBoost	0.84	0.83	0.84	0.79	0.64
AUROC: Area Under the Receiver Operating Characteristics Curve					

Table 2: *Comparison of model performance on the test set.*

Evaluation: Further evaluation of the CatBoost algorithm for preterm birth prediction was carried out using a separate validation set. The model's performance was compared with those of two physicians who independently made predictions on which births would be preterm in the same dataset. The results are summarized in table 3.

Model	Accuracy	Precision	Recall	F1-score
CatBoost	0.84	0.83	0.84	0.79
Physician A	0.46	0.74	0.47	0.57
Physician B	0.38	0.75	0.30	0.43

Table 3: *Comparison of model performance.*

Influence of variables on predictions: The relative importance of variables incorporated in the CatBoost prediction model is shown in figure 1. The top five risk factors for preterm birth, in order of importance, were multiple pregnancies (number of fetuses), followed by fetal humerus length (fetus humerus), trans cerebellar diameter, BMI, and mean pulsatility index.

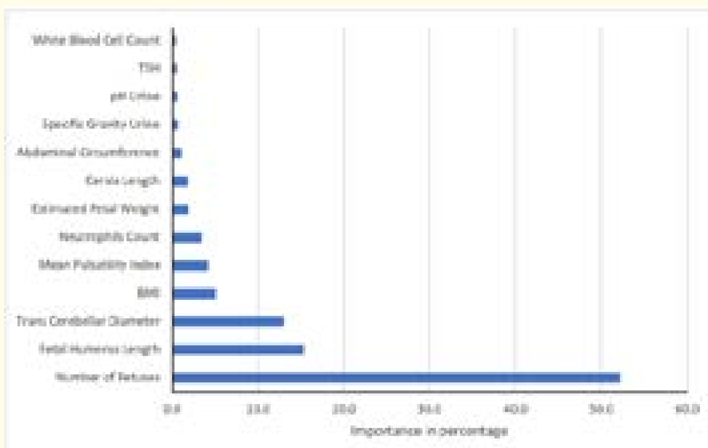

Figure 1: *Ranking of features by importance.*

BMI: Body Mass Index; TSH: Thyroid Stimulating Hormone.

Discussion

Accurate risk prediction is crucial in clinical settings for the timely administration of interventions. Detecting the risk of preterm birth can allow healthcare providers to implement appropriate interventions, improving the overall outcomes for both mother and child. Furthermore, in resource-limited settings, the identification of mothers at high risk of delivering prematurely can potentially aid in the optimum allocation of healthcare resources, ensuring the delivery of necessary care for those in need. Machine learning algorithms have been used in healthcare settings for diagnosing diseases and predicting patient outcomes [24]. In the context of preterm birth, the use of such approaches can help identify the complex non-linear relationships between various features of the mother and fetus and the risk of preterm birth, thereby improving the accuracy of predictions [12,25].

The choice of a suitable algorithm not only determines the predictive power of the model but also impacts its efficiency, interpretability, and generalization to unseen data. Tree-based algorithms are non-parametric models without distributional assumptions [17]. They can capture non-linear associations between variables and outcomes. These approaches have been used in prognostication due to their high interpretability and suitability for mixed (numerical and categorical) variables [18]. The ease of handling missing values and minimizing the risk of overfitting were also considered when choosing different algorithms. A comparison of several widely used algorithms, such as Random Forest, Gradient Boosting, AdaBoost, XGBoost, and CatBoost, was carried out. With its ensemble approach and feature selection capabilities, Random Forest is robust in its ability to handle noisy data. Gradient Boosting, on the other hand, sequentially builds a robust predictive model by correcting the errors of the preceding model, thus enhancing its predictive accuracy. AdaBoost further refines this process by assigning higher weights to misclassified samples, thereby focusing on challenging instances in the dataset. XGBoost is known for its optimization strategies and regularization techniques. It aims to balance model complexity and predictive performance. By encoding categorical features during the training process, CatBoost reduces the risk of data leakage and ensures robust model generalization. Furthermore, CatBoost provides native support for handling missing values.

In the comparative analysis of model performance, CatBoost demonstrated the highest performance across all metrics. XGBoost was a close contender, with scores that were comparable. The Random Forest algorithm also scored relatively high scores across all metrics and scored the second-highest AUROC score (0.6). This indicates that the Random Forest algorithm performs well in classification tasks, although it is slightly less effective than the boosting algorithms. Gradient Boost and AdaBoost displayed moderate performance, with Gradient Boost having a slight edge over AdaBoost in most metrics. However, both lagged behind the more advanced boosting methods, particularly in terms of their AUROC scores, suggesting that these models may not be able to differentiate between classes as effectively as CatBoost or XGBoost. The superior performance of CatBoost in terms of predictive accuracy, especially in datasets with complex relationships and categorical variables, underscores its applicability in complex medical datasets. The CatBoost algorithm trained on the data from the study cohort was able to identify outcomes (in the evaluation set) with an accuracy of 84%, which outperformed manual review-based predictions by physicians. The external validation using predictions by physicians in this study indicates that CatBoost outperforms AdaBoost which has been previously reported to have an accuracy of 72.73% [26].

The top five risk factors (multiple pregnancies, fetal humerus length, trans cerebellar diameter, BMI, and mean pulsatility index), as highlighted by the analysis in this paper, underscore a combination of both maternal and fetal characteristics that play pivotal roles in preterm birth. Multiple pregnancies emerged as the primary risk factor. This is consistent with previous reports as being one of the most important predictors of preterm birth [26]. The overstretching of the uterus to accommodate multiple fetuses often triggers premature uterine contractions, which can lead to an increased likelihood of complications and an increased risk of preterm birth. Although the relationship between the size of the fetal humerus and risk of preterm birth is indirect, short fetal femur and humeri lengths are known to be associated with fetal growth restriction, which may result in a higher risk of spontaneous preterm birth [27,28]. Similarly, anomalies in the trans cerebellar diameter may point to abnormalities in fetal growth, resulting in the onset of preterm birth. Extremes in maternal BMI are another risk factor. Whether on the lower or higher end of the spectrum, abnormalities in BMI are linked to an increased likelihood of preterm birth [29,30]. The utility of the mean pulsatility index as a predictive factor for preterm birth has been proposed, but in limited capacity [31]. The pulsatility index reflects the resistance to blood flow in the uterine artery, with higher values indicating increased resistance and potential vascular dysfunction.[31] In the context of spontaneous preterm birth, elevated mean uterine artery pulsatility index has been shown to be weakly associated with an increased risk of preterm birth [31].

However, it is important to note that these risk factors may not act in isolation but may interact with other risk factors and with various environmental and genetic factors in affecting preterm birth. Understanding the interconnected nature of risk factors is essential for developing effective intervention strategies aimed at reducing the rates of preterm birth and improving neonatal outcomes.

While this study focuses on specific risk factors, there are likely other contributing factors that have not been examined. Further research is needed to explore such additional factors and their interactions. Another key area for future exploration is in the use of different computational approaches to predict pregnancy-related complications and maternal and neonatal health outcomes. While the current study utilized traditional machine learning algorithms, there is a growing interest in leveraging deep learning techniques, which have shown promise in handling more complex data patterns and relationships. Deep learning models, particularly those using neural networks, could potentially uncover deeper insights from the data, especially in recognizing subtle patterns that traditional machine learning might miss. However, the interpretability of deep learning models remains a concern, especially in clinical settings where understanding the decision-making process of the model is crucial.

Conclusion

Timely prediction of the risk of preterm birth has the potential to aid clinical decisions for effective prevention and management of preterm birth. Furthermore, it can help optimize resource allocation such that this is geared toward those at high risk. This study used data from hospital records to evaluate the feasibility of predicting preterm birth. Various physical, physiological, biochemical, and hematological parameters were considered. Critical determinants of preterm birth were also identified. The integration of predictive artificial intelligence models for preterm birth risk assessment can potentially improve the care of mothers and newborns. The findings from this study provide a foundation for further evaluation of such approaches for integration into clinical practice.

Conflict of Interest

None.

Funding Support

None.

Availability of Data and Materials

Will be provided on request to the corresponding author.

Ethics Approval and Consent to Participate

This study was approved by the Institutional Ethics Committee of the CHN with the approval number- IEC/C9791/2003/24009. The study was performed in accordance with the Declaration of Helsinki.

Patient Consent Statement

Written informed consent was obtained from all subjects before the study was initiated. The identities of the subjects were kept confidential, and anonymized data was used for all analyses.

Permission to Reproduce Material from Other Sources

Not applicable.

Acknowledgement

The authors of the manuscript would like to thank Neeraj Verma, chief tech team, Cloudnine Corporate Office, Bangalore.

The authors would like to acknowledge the support of Mr. Raviganesh Venkataraman, CEO for his valuable support and encouragement of this study.

Authors Contribution

All authors have conceived and/or designed the work that led to the submission, acquired data, and/or played an important role in interpreting the results. Drafted or revised the manuscript. Approved the final version. Agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Bibliography

1. Anonymous. "152 Million Babies Born Preterm in the Last Decade". n.d.
2. Saigal S and Doyle LW. "An overview of mortality and sequelae of preterm birth from infancy to adulthood". *Lancet* 371.9608 (2008): 261-269.
3. Raju TNK., *et al*. "Long-term healthcare outcomes of preterm birth: An executive summary of a conference sponsored by the national institutes of health". *Journal of Pediatrics* 181 (2017): 309-318.e1.
4. Crump C., *et al*. "Preterm birth and risk of type 1 and type 2 diabetes: a national cohort study". *Diabetologia* 63.3 (2020): 508-518.
5. Frey HA and Klebanoff MA. "The epidemiology, etiology, and costs of preterm birth". *Seminars in Fetal and Neonatal Medicine* 21.2 (2016): 68-73.
6. Purisch SE and Gyamfi-Bannerman C. "Epidemiology of preterm birth". *Seminars in Perinatology* 41.7 (2017): 387-391.
7. Oskovi Kaplan ZA and Ozgu-Erdinc AS. "Prediction of preterm birth: Maternal characteristics, ultrasound markers, and biomarkers: An updated overview". *Journal of Pregnancy* (2018): 8367571.
8. Zierden HC., *et al*. "Next generation strategies for preventing preterm birth". *Advanced Drug Delivery Reviews* 174 (2021): 190-209.
9. Auger N., *et al*. "Association between maternal comorbidity and preterm birth by severity and clinical subtype: retrospective cohort study". *BMC Pregnancy Childbirth* 11 (2011): 67.
10. Koullali B., *et al*. "Risk assessment and management to prevent preterm birth". *Seminars in Fetal and Neonatal Medicine* 21.2 (2016): 80-88.
11. Sharma V., *et al*. "Clinical risk prediction models: the canary in the coalmine for artificial intelligence in healthcare?" *BMJ Health and Care Informatics* 28.1 (2021): e100421.
12. Sharifi-Heris Z., *et al*. "Machine learning approach for preterm birth prediction using health records: Systematic review". *JMIR Medical Informatics* 10.4 (2022): e33875.
13. Sun H., *et al*. "Machine learning-based prediction models for different clinical risks in different hospitals: Evaluation of live performance". *Journal of Medical Internet Research* 24.6 (2022): e34295.
14. Huang J., *et al*. "Analysis of factors related to preterm birth: a retrospective study at Nanjing Maternity and Child Health Care Hospital in China". *Medicine (Baltimore)* 99.28 (2020): e21172.
15. Nieto-Del-Amor F., *et al*. "Combination of feature selection and resampling methods to predict preterm birth based on electrohysterographic signals from imbalance data". *Sensors (Basel)* 22.14 (2022): 5098.
16. Goldenberg RL., *et al*. "Biochemical markers for the prediction of preterm birth". *American Journal of Obstetrics and Gynecology* 192.5 (2005): S36-S46.

17. Song Y-Y and Lu Y. "Decision tree methods: applications for classification and prediction". *Shanghai Archives of Psychiatry* 27.2 (2015): 130-135.

18. Gonçalves DM., *et al*. "Predicting postoperative complications in cancer patients: A survey bridging classical and machine learning contributions to postsurgical risk analysis". *Cancers (Basel)* 13.13 (2021): 3217.

19. Chen T and Guestrin C. "XGBoost". In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining ACM: New York, NY, USA (2016).

20. Breiman L. "Random Forests". *Machine Learning* 45 (2001): 5-32.

21. Friedman JH. "Greedy function approximation: A gradient boosting machine". *Annals of Statistics* 29.5 (2001): 1189-1132.

22. Freund Y and Schapire RE. "A decision-theoretic generalization of on-line learning and an application to boosting". *Journal of Computer and System Sciences* 55.1 (1997): 119-139.

23. Dorogush AV., *et al*. "CatBoost: Gradient boosting with categorical features support". arXiv [csLG] (2018).

24. Deo RC. "Machine learning in medicine". *Circulation* 132.20 (2015): 1920-1930.

25. AlSaad R., *et al*. "PredictPTB: an interpretable preterm birth prediction model using attention-based recurrent neural networks". *BioData Mining* 15.1 (2022): 6.

26. Zhang Y., *et al*. "Establishment of a model for predicting preterm birth based on the machine learning algorithm". *BMC Pregnancy Childbirth* 23.1 (2023): 779.

27. de Carvalho AAV., *et al*. "Association of midtrimester short femur and short humerus with fetal growth restriction". *Prenatal Diagnosis* 33.2 (2013): 130-133.

28. Partap U., *et al*. "Fetal growth and the risk of spontaneous preterm birth in a prospective cohort study of nulliparous women". *American Journal of Epidemiology* 184.2 (2016): 110-119.

29. Han Z., *et al*. "Maternal underweight and the risk of preterm birth and low birth weight: a systematic review and meta-analyses". *International Journal of Epidemiology* 40.1 (2011): 65-71.

30. Liu K., *et al*. "Association of maternal obesity with preterm birth phenotype and mediation effects of gestational diabetes mellitus and preeclampsia: a prospective cohort study". *BMC Pregnancy Childbirth* 22.1 (2022).

31. van Zijl MD., *et al*. "The predictive capacity of uterine artery Doppler for preterm birth—A cohort study". *Acta Obstetricia et Gynecologica Scandinavica* 99.4 (2020): 494-502.

Volume 14 Issue 5 May 2025