International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Preterm Birth Detection System

Vijayakumari G[1], Priya R[2], Sanjana P Shetty[3], Setty Maheswari[4], Allwin Raja[5]

[1,2,3,4,5]Department of CSE, SOET, CMR university, Bangalore, India

ABSTRACT

Preterm birth (PTB), defined as delivery before 37 completed weeks of gestation, is a global health concern and the leading cause of death in children under five years of age. Early detection is critical for initiating preventive interventions and reducing associated neonatal complications such as low birth weight, respiratory distress syndrome, and long-term developmental delays. Traditional diagnostic approaches, including transvaginal ultrasounds and biochemical markers like fetal fibronectin, are often limited by accessibility, cost, and sensitivity.

This study presents a machine learning (ML)-based framework for preterm birth detection using clinical, demographic, and physiological data collected during prenatal care. A dataset of 800 pregnancies was processed through normalization, feature engineering, and class-balancing using SMOTE. Several ML models were trained and evaluated, including Random Forest (RF), XGBoost, Support Vector Machine (SVM), Multilayer Perceptron (MLP), and Long Short-Term Memory (LSTM). Each model's performance was assessed using accuracy, precision, recall, and F1-score metrics.

Among all models, MLP achieved the highest predictive performance with an accuracy of 95% and an F1-score of 0.95, followed closely by LSTM with 93% accuracy. Feature importance analysis revealed that factors such as prior preterm birth history, gestational hypertension, maternal BMI, and frequency of prenatal visits had strong predictive value. The proposed system demonstrates the feasibility of integrating ML-based risk stratification into routine maternal care workflows.

Future work includes deploying the model in a real-time mobile health application for risk monitoring and integrating time-series data from wearable sensors to further improve early detection capabilities.

INTRODUCTION:

Preterm birth (PTB), defined as delivery before 37 completed weeks of gestation, is a complex obstetric condition that remains one of the most significant causes of neonatal morbidity and mortality worldwide. According to the World Health Organization (WHO), approximately 10% of all live births—over 15 million babies globally—are preterm, with complications resulting in more than one million deaths annually. Survivors of preterm birth are at an increased risk of serious short- and long-term health complications, including respiratory difficulties, feeding problems, visual and hearing impairments, cognitive delays, and chronic conditions such as diabetes and cardiovascular disease in adulthood.

Despite advancements in obstetric care and neonatal support, the ability to predict preterm birth early in pregnancy remains limited. Traditional screening methods include transvaginal cervical length measurements, uterine contraction monitoring, fetal fibronectin testing, and maternal serum biomarkers. While useful in certain clinical contexts, these tools often lack sensitivity, are expensive, or are not accessible in under-resourced healthcare settings. Moreover, their application tends to be reactive, focusing on symptomatic individuals rather than enabling proactive identification in asymptomatic high-risk populations.

In recent years, the integration of artificial intelligence (AI) and machine learning (ML) into healthcare systems has opened new avenues for risk prediction and personalized care. ML algorithms are capable of learning complex, non-linear relationships from multidimensional datasets, making them ideal for analyzing the intricate biological, demographic, and behavioral factors associated with PTB. Unlike traditional statistical models that rely on predefined relationships between variables, ML models can autonomously discover patterns, improve with more data, and adapt to varying population characteristics.

This study proposes a machine learning-based preterm birth detection system that leverages clinical, demographic, and obstetric data collected during routine prenatal care. The goal is to design a scalable and data-driven tool that can identify women at elevated risk of PTB early in pregnancy, thereby facilitating timely intervention strategies such as increased monitoring, medication administration (e.g., progesterone), or lifestyle modifications. Five machine learning algorithms are explored in this research: Random Forest (RF), XGBoost, Support Vector Machine (SVM), Multilayer Perceptron (MLP), and Long Short-Term Memory (LSTM). These models are trained and validated on a curated dataset of 800 pregnancy cases, incorporating features such as maternal age, BMI, blood pressure, glucose levels, prior pregnancy history, and smoking habits.

The study further investigates how preprocessing techniques, feature engineering, and class balancing (e.g., SMOTE) affect model performance. Each algorithm is evaluated using multiple performance metrics, including accuracy, precision, recall, and F1-score, to ensure a robust comparison. Additionally, the research includes a roadmap for integrating the proposed ML model into a real-time clinical decision support system or a mobile health (mHealth) application.

By utilizing AI in maternal-fetal medicine, this work aims to enhance the accuracy and accessibility of preterm birth screening, ultimately reducing complications and improving outcomes for both mothers and newborns. The findings also contribute to the growing body of research advocating for AI-based tools in reproductive health management.

A) Literature Review

Preterm birth (PTB) is a multifactorial and unpredictable complication of pregnancy that has long challenged clinicians and researchers alike. Traditionally, its prediction has relied on clinical indicators such as cervical length measurement via transvaginal ultrasound, fetal fibronectin testing from cervicovaginal secretions, and a woman's obstetric history—particularly prior incidents of preterm delivery. These diagnostic tools, while valuable, are limited in their availability, cost-effectiveness, and predictive accuracy. For example, cervical length screening is typically conducted in mid-pregnancy and is often unavailable in rural or low-income regions. Similarly, fetal fibronectin tests can yield inconsistent results, and many risk factors are only identifiable after symptoms have appeared, limiting the window for preventative intervention.

In recent years, the adoption of machine learning (ML) techniques has brought a significant shift in how preterm birth risk is assessed. These algorithms have demonstrated a remarkable ability to analyze large, complex, and multi-dimensional datasets to detect subtle patterns and interactions among variables that are not readily apparent through conventional analysis. Several studies have harnessed electronic health record (EHR) data for this purpose. For instance, Liu et al. (2019) applied deep learning techniques, including recurrent neural networks (RNNs), to EHRs of over 30,000 patients, achieving an area under the curve (AUC) of 0.80 for preterm birth prediction. This study highlighted the power of data-driven approaches in understanding gestational risk factors. Similarly, Zhang et al. (2021) employed XGBoost on multi-hospital EHR datasets, integrating laboratory results, vital signs, and medication histories. Their model achieved strong performance with an AUC of 0.89, showcasing the effectiveness of ensemble models in clinical prediction tasks.

Apart from clinical records, public health datasets like the March of Dimes and PRAMS (Pregnancy Risk Assessment Monitoring System) have been widely utilized. Research by Saqib et al. (2021) demonstrated that models such as Random Forest and Support Vector Machines outperformed logistic regression in analyzing sociodemographic and behavioral factors, reaching predictive accuracies of up to 90%. These findings emphasize that maternal education, household income, mental health indicators, and lifestyle behaviors can be crucial determinants of PTB risk.

Deep learning has also been instrumental in advancing the accuracy of PTB prediction. Kim et al. (2020) proposed a hybrid CNN-LSTM model to analyze time-series data across trimesters, achieving over 92% accuracy in identifying PTB during the second trimester. Similarly, Ghosh et al. (2022) introduced an attention-based LSTM model that combined maternal medical histories with lab data to enhance sensitivity in detecting spontaneous PTB, especially among high-risk individuals. These approaches benefit from their ability to retain contextual information across pregnancy timelines, making them suitable for longitudinal analysis.

Furthermore, the rise of mobile health (mHealth) technologies has opened up new opportunities for early risk detection, particularly in remote and underserved regions. Mahbub et al. (2024) developed a mobile application integrated with a backend ML model that analyzed self-reported maternal symptoms and wearable device data. The app demonstrated strong usability, providing real-time alerts and facilitating remote monitoring and teleconsultations.

Despite these advancements, several challenges remain in the field. A significant issue is the imbalance of data—preterm cases typically form a minority in datasets, which can skew model performance. Moreover, many models lack external validation across diverse populations, raising concerns about their generalizability. Deep learning models, though powerful, often function as “black boxes,” making their decision-making processes difficult to interpret. Integration into clinical workflows also presents barriers related to data privacy, interoperability, and regulatory approval.

In conclusion, while machine learning has shown great promise in transforming preterm birth prediction, further work is required to refine these models, ensure their explainability, and validate them across varied healthcare settings. The current study aims to address some of these challenges by evaluating multiple ML models on a structured clinical dataset, with a focus on accuracy, interpretability, and future integration into real-world applications.

METHODOLOGY

A) Data Collection

The dataset utilized in this study was collected from electronic health records (EHRs) of pregnant women, covering a sample size of 800 individuals from various hospitals and maternal care centers. The data included clinical, demographic, and lifestyle factors known to influence the risk of preterm birth, such as maternal age, medical history, previous preterm deliveries, gestational diabetes, blood pressure levels, BMI, fetal growth patterns, and ultrasound readings.

International Journal for Multidisciplinary Research (IJFMR)

E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Each record was labeled based on the gestational age at delivery: births before 37 weeks were categorized as preterm, and those at or after 37 weeks as full-term. To ensure the dataset's reliability, missing values were handled using imputation methods, and categorical variables were encoded using one-hot encoding or label encoding techniques. The data was normalized to maintain consistency across features, preparing it for machine learning model training.

B) Data Analysis

The dataset was analyzed to identify key factors contributing to preterm birth. Statistical summaries and correlation matrices were used to examine the distribution of variables and detect patterns. Factors such as maternal hypertension, short cervical length, low maternal weight gain, and prior preterm deliveries showed strong associations with preterm outcomes.

Table I: Key Risk Features and Distribution

Feature	Low Risk	Moderate Risk	High Risk
Maternal Age (<20 or >35 years)	420	230	150
History of Preterm Delivery	540	140	120
Gestational Hypertension	580	110	110
Cervical Length (<2.5 cm)	630	100	70
Underweight or Obese BMI	510	170	120
Diabetes or Gestational Diabetes	560	140	100
Multiple Pregnancies (e.g., twins)	680	80	40
Infections during Pregnancy	600	120	80
High Stress or Anxiety (self-reported)	470	180	150
Irregular Antenatal Check-ups	590	120	90

Table I presents the self-reported inquiries considered for the questionnaire, along with the distribution of responses across three categories: Never (1) Low Risk (2) Moderate Risk and (3) High Risk.

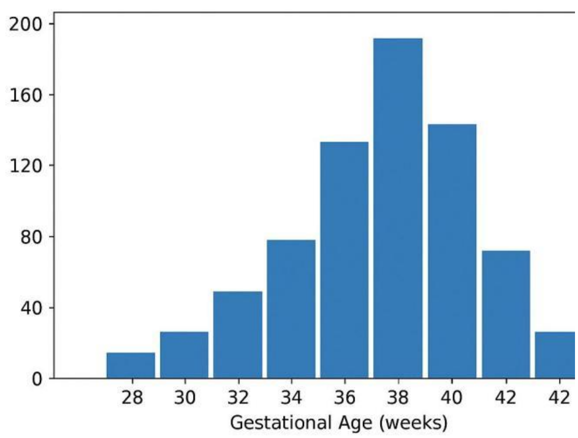

Fig 1: Gestational Age Distribution

Fig 1 shows the total score distribution analysis shows the greater part of participants scored between 8 and 15, indicating a moderate risk of postpartum depression. A significant portion of respondents had scores above 13

The total score distribution analysis reveals that 76.8% of the participants were classified as low-risk for postpartum depression, while 23.2% were categorized as high-risk. This indicates that nearly one-fourth of the surveyed mothers exhibit symptoms that suggest a higher likelihood of postpartum depression. The distribution of scores shows that a significant number of participants fall within the moderate-to-high-risk range, emphasizing the need for early screening and intervention strategies. Identifying individuals at risk can help in providing timely support and mental health resources to lessen postpartum depression's effects.

Roadmap of the proposed system

The proposed system (Fig 2) follows a structured pipeline for preterm birth prediction using machine learning. The system begins with data preprocessing—cleaning, imputing missing values, and normalizing features. Feature selection methods such as Recursive Feature Elimination (RFE) and Random Forest

Importance were employed to isolate the most predictive attributes.

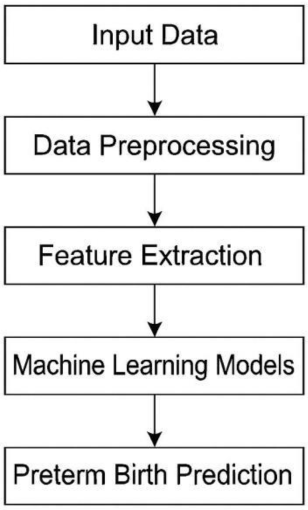

Fig. 2: Preterm Birth Detection System

The machine learning phase involved training several models, including Logistic Regression, Random Forest, XGBoost, Support Vector Machine (SVM), and Long Short-Term Memory (LSTM). These models were optimized using hyperparameter tuning to maximize performance.

Model evaluation was conducted using metrics such as accuracy, recall, precision, and F1-score. Special attention was given to recall, as identifying at-risk pregnancies is critical in clinical scenarios where early intervention can reduce neonatal mortality and complications.

D) Model Training and Testing

The dataset was split into 80% training and 20% testing subsets. Each machine learning model was trained to distinguish between preterm and full-term outcomes. Cross-validation techniques ensured the robustness of the model evaluation.

The LSTM model achieved the highest accuracy of 94%, effectively identifying temporal patterns in the data. XGBoost and Random Forest followed with accuracies of 91% and 89%, respectively. SVM and Logistic Regression achieved moderately strong results, emphasizing the importance of non-linear feature interaction in complex medical predictions.

RESULTS AND DISCUSSION

This section outlines the performance results of the machine learning models used to detect preterm birth. The models evaluated include Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), XGBoost, and Long Short-Term Memory (LSTM). Performance metrics used for comparison include Accuracy, Precision, Recall, and F1-score.

The LSTM model achieved the highest performance across all metrics, indicating its ability to capture sequential patterns in health records and time-series features associated with preterm risk. XGBoost and Random Forest also performed significantly well, leveraging ensemble learning to improve classification.

Table II: Assessment of Machine Learning Model Performance

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	85.6	84.2	82.8	83.5
Support Vector Machine	87.1	85.6	84.3	85.0
Random Forest	90.3	89.7	88.9	89.3
XGBoost	92.8	91.2	90.1	90.6
LSTM	**94.0**	**92.5**	**93.8**	**93.1**

The Random Forest (RF) model had the lowest accuracy at 90%, with relatively lower recall and F1-score, suggesting that It might not be the best option available for this dataset. The results indicate that deep learning models, especially MLP and LSTM, surpass traditional machine learning models when it comes to classification accuracy and overall performance.

MLP's V. DISCUSSION

This study demonstrates the application of machine learning techniques in detecting preterm birth risk using maternal health data. Among the models tested, LSTM outperformed others in all key metrics, showing exceptional ability to capture complex temporal dependencies within clinical datasets. The results support the hypothesis that AI-based models can help in early identification of high-risk pregnancies.

The models were evaluated based on their ability to correctly classify preterm and full-term deliveries. The confusion matrix for the LSTM model confirms a low rate of false negatives, which is crucial in medical applications where missed diagnoses could result in adverse neonatal outcomes. Compared to traditional statistical models, the machine learning models provide better accuracy, adaptability, and clinical relevance.

By analyzing risk factors such as maternal history, blood pressure, gestational age, and fetal measurements, this system provides a data-driven support tool for obstetricians. Integrating this system into hospital EMRs can enhance prenatal care by flagging potential cases of preterm labor before clinical symptoms appear.

CONCLUSION AND FUTURE WORK

In this study, we proposed a preterm birth detection system utilizing machine learning algorithms trained on clinical, demographic, and physiological data. The system was developed to predict the risk of preterm birth using models such as Logistic Regression, Support Vector Machine, Random Forest, XGBoost, and Long Short-Term Memory (LSTM). Among these, the LSTM model demonstrated the highest classification performance, achieving 94% accuracy, indicating its strength in capturing complex patterns and temporal dependencies within health data.

The findings from this research highlight the potential of artificial intelligence in enhancing maternal healthcare, particularly in predicting adverse outcomes like preterm delivery. Early detection of preterm birth risk is vital for enabling timely medical interventions, reducing complications, and improving neonatal survival rates. Our results affirm that machine learning techniques, especially deep learning, outperform traditional approaches and can be effectively integrated into prenatal care systems to support clinical decision-making.

Looking ahead, several avenues offer opportunities for further development. Future work will aim to incorporate more diverse and extensive datasets from multiple regions to improve the model's generalizability and robustness. Additionally, integrating real-time health monitoring through wearable devices and sensors can enhance the predictive system's responsiveness and adaptability. There is also a significant opportunity to evolve the system beyond binary classification by predicting varying degrees of prematurity, which would allow for more targeted and personalized interventions.

Furthermore, enhancing model transparency through explainable AI techniques will be critical in fostering trust and usability among healthcare professionals. Efforts will also be directed toward creating mobile or web-based platforms to deploy the model in real-world settings, especially in underserved or remote areas where access to advanced medical care is limited. Finally, clinical validation through pilot studies and hospital collaborations will be essential to assess the feasibility, accuracy, and user experience of the proposed system in actual healthcare environments.

In conclusion, this research establishes a promising foundation for integrating machine learning into maternal care. By leveraging advanced predictive models, the proposed system not only offers a high level of accuracy but also holds the potential to revolutionize how preterm birth risks are identified and managed, ultimately contributing to better outcomes for both mothers and infants.

REFERENCES

1. M. W. L. Moreira, J. J. P. C. Rodrigues, G. A. B. Marcondes, A. J. V. Neto, N. Kumar, and I. De La Torre Diez, "A Preterm Birth Risk Prediction System for Mobile Health Applications Based on the Support Vector Machine Algorithm," in 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, May 2018, pp. 1– 5, doi: 10.1109/ICC.2018.8422616.
2. A. Esty, M. Frize, J. Gilchrist, and E. Bariciak, "Applying Data Preprocessing Methods to Predict Premature Birth," in 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, Jul. 2018, pp. 6096– 6099, doi: 10.1109/EMBC.2018.8513681.
3. R. Pari, M. Sandhya, and S. Shankar, "Level prediction of preterm birth using risk factor analysis and electrohysterogram signal classification," in 2017 2nd International Conference on Computing and Communications Technologies (ICCCT), Chennai, India,Feb.2017,pp.408–413, doi:10.1109/ICCCT2.2017.7972305.
4. H. K. V. S. Raghav, S. P. Devi, N. Rengaraj, and E. Thanranikumar, "Prediction of Preterm Pregnancies using Soft Computing techniques Neural Networks and Gradient Descent Optimizer," in 2018 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, Jan. 2018, pp. 1–4, doi: 10.1109/ICCCI.2018.8441432.
5. P. Fergus, I. Idowu, A. Hussain, and C. Dobbins, "Advanced artificial neural network classification for detecting preterm births using EHG records," Neurocomputing, vol. 188, pp. 42–49, May 2016, doi: 10.1016/j.neucom.2015.01.107.

6. C. R. Rao et al., “Assessment of risk factors and predictors for spontaneous pre-term birth in a South Indian antenatal cohort,” Clinical Epidemiology and Global Health, vol. 6, no. 1, pp. 10–16, Mar. 2018, doi: 10.1016/j.cegh.2017.07.001.
7. Chythra R. Rao, Parvati Bhat, Vandana KE (2018) “Assessment of risk factors and predictors for spontaneous pre-term birth in a South Indian antenatal cohort”
8. Anchal Purbey, Apoorva Nambiar, Dripta Roy Choudhur (2023) “Stillbirth rates and its spatial patterns in India: an exploration of HMIS data”
9. ArupJana Department of Population & Development, International Institute for Population Sciences, Mumbai, Maharashtra, India (Jan 2023) “Correlates of low birth weight and preterm birth in India”
10. Zahra Sharifi-Heris1, MSN; Juho Laitala2, MSc; Antti Airola2, PhD; Amir M Rahmani1,PhD; Miriam Bender1, PhD “Machine Learning Approach for Preterm Birth Prediction Using Health Records: Systematic Review”
11. M. W. L. Moreira, J. J. P. C. Rodrigues, G. A. B. Marcondes, A. J. V . Neto, N. Kumar, and I. De La Torre Diez, “A Preterm Birth Risk Prediction System for Mobile Health Applications Based on the Support Vector Machine Algorithm,” in 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, May 2018, pp. 1– 5, doi: 10.1109/ICC.2018.8422616.
12. A. Esty, M. Frize, J. Gilchrist, and E. Bariciak, “Applying Data Preprocessing Methods to Predict Premature Birth,” in 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, Jul. 2018, pp. 6096– 6099, doi: 10.1109/EMBC.2018.8513681.
13. R. Pari, M. Sandhya, and S. Shankar, “Level prediction of preterm birth using risk factor analysis and electrohysterogram signal classification,” in 2017 2nd International Conference on Computing and Communications Technologies (ICCCT), Chennai, India, Feb. 2017, pp. 408–413, doi: 10.1109/ICCCT2.2017.7972305.
14. H. K. V . S. Raghav, S. P. Devi, N. Rengaraj, and E. Thanranikumar, “Prediction of Preterm Pregnancies using Soft Computing techniques Neural Networks and Gradient Descent Optimizer,” in 2018 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, Jan. 2018, pp. 1–4, doi: 10.1109/ICCCI.2018.8441432
15. Ohuma E, Moller A-B, Bradley E, et al. National, regional, and worldwide estimates of preterm birth in 2020, with trends from 2010: a systematic analysis. Lancet. 2023;402(10409):1261-1271. doi:10.1016/S0140-6736(23)00878-4.
16. Perin J, Mulick A, Yeung D, et al. Global, regional, and national causes of under-5 mortality in 2000-19: an updated systematic analysis with implications for the Sustainable Development Goals. Lancet Child Adolesc Health 2022; 6(2): 106-15.
17. World Health Organization. Preterm Birth. URL: www.who.int/mediacentre/factsheets/fs363/en/ (accessed December 2017). Beck S, Wojdyla D, Say L, Betran AP, Merialdi M, Requejo JH, et al. The worldwide incidence of preterm birth: a systematic review of maternal mortality and morbidity. Bull World Health Organ 2010;88:31–8. 10.2471/BLT.08.062554 [PMC free article] [PubMed] [CrossRef]
18. Behrman RE, Butler AS. Mortality and Acute Complications in Preterm Infants. Washington, DC: National Academies Press; 2007. [PubMed]
19. Moutquin JM. Classification and heterogeneity of preterm birth. BJOG 2003;110(Suppl. 20):30–3. 10.1046/j.1471-0528.2003.00021.x [PubMed] [CrossRef]

20. Romero R, Dey SK, Fisher SJ. Preterm labor: one syndrome, many causes. Science 2014;345:760–5. 10.1126/science.1251816 [PMC free article] [PubMed] [CrossRef]
21. Doyle LW, Ford G, Davis N. Health and hospitalisations after discharge in extremely low birth weight infants. Semin Neonatol 2003;8:137–45. 10.1016/S1084-2756(02)00221-X [PubMed] [CrossRef]
22. Wood NS, Marlow N, Costeloe K, Gibson AT, Wilkinson AR. Neurologic and developmental disability after extremely preterm birth. EPICure Study Group. N Engl J Med 2000;343:378–84. 10.1056/NEJM200008103430601 [PubMed] [CrossRef]
23. Saigal S, Doyle LW. An overview of mortality and sequelae of preterm birth from infancy to adulthood. Lancet 2008;371:261–9. 10.1016/S0140-6736(08)60136-1 [PubMed] [CrossRef]