BMC Pregnancy and Childbirth

RESEARCH

Open Access

Application of machine learning in identifying risk factors for low APGAR scores

Haifa Fahad Alhasson[1*], Nagat Elhag[2], Shuaa Saleem Alharbi[1†] and Ishag Adam[3†]

Abstract

Background Identifying the risk factors for low APGAR scores at birth is critical for improving neonatal outcomes and guiding clinical interventions.

Methods This study aimed to develop a machine-learning model that predicts low APGAR scores by incorporating maternal, fetal, and perinatal factors in Wad Medani, Sudan. Using a Random Forest Classifier, we performed hyper-parameter optimization through Grid Search cross-validation (CV) to identify the best-performing model configuration.

Results The optimized model achieved excellent predictive performance, as evidenced by high F1 scores, accuracy, and balanced precision-recall metrics on the test set. In addition to prediction, feature importance analysis was conducted to identify the most influential risk factors contributing to low APGAR scores. Key predictors included gestational age, maternal BMI, mode of delivery, and history of previous complications such as stillbirth or abortion. Using 5-fold cross-validation (CV), the random forest model performance scored accuracy at 96%, precision at 98%, recall at 97%, and F1-score at 97% when classifying infants with APGAR score.

Conclusion This study underscores the importance of incorporating machine learning approaches in obstetric care to understand better and mitigate the risk factors associated with adverse neonatal outcomes, particularly low APGAR scores. The results provide a foundation for developing targeted interventions and improving prenatal care practices.

Keywords Low APGAR score, Risk factors, Machine learning, Artificial intelligence

Background

Low Appearance, Pulse, Grimace, Activity and Respiration (APGAR) scores are influenced by a variety of maternal, delivery-related, and biological factors that can compromise neonatal health. The APGAR score, introduced by Dr. Virginia Apgar in 1952, is a rapid assessment tool used to evaluate newborn health immediately after birth based on five criteria: Appearance, Pulse, Grimace response, Activity, and Respiration. A score below 7 at five minutes post-delivery is a critical indicator of potential health complications and is associated with increased risks of neonatal morbidity and mortality [1]. Understanding the factors contributing to low APGAR scores is essential for healthcare providers, as these scores impact both immediate neonatal care and long-term health outcomes.

Maternal health conditions, such as inadequate prenatal care, gestational diabetes, hypertension, and substance use, significantly affect newborn APGAR scores [1]. Insufficient prenatal support is linked to higher rates of low APGAR scores, often necessitating extended

[†]Shuaa Saleem Alharbi and Ishag Adam contributed equally to this work.

*Correspondence:
Haifa Fahad Alhasson
hhson@qu.edu.sa
[1] Department of Information Technology, College of Computer, Qassim University, Buraydah 52571, Saudi Arabia
[2] Wad Medani College of Medical Sciences and Technology, Wad Medani, Sudan
[3] Department of Obstetrics and Gynecology, College of Medicine, Qassim University, Buraydah 52571, Saudi Arabia

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

hospital stays and additional medical interventions. Conditions like maternal substance abuse further exacerbate neonatal risks, leading to complications such as neonatal abstinence syndrome [1, 2]. Delivery-related factors, including the administration of labor medications and complications during childbirth, also play a critical role. Medications given to mothers during labor can affect neonatal physiology, while complications such as prolonged labor or emergency cesarean sections may jeopardize the infant's health [3, 4].

Biological factors, such as prematurity and congenital anomalies, are well-established contributors to low APGAR scores. Premature infants often face physiological challenges, including underdeveloped organ systems, which can result in immediate instability [3]. Congenital malformations also compromise neonatal health, further lowering APGAR scores.

The implications of low APGAR scores extend beyond the neonatal period. Studies have shown that low scores are predictive of long-term developmental and neurological challenges, such as cerebral palsy and cognitive impairments [2, 5]. Infants with low scores face an elevated risk of lifelong disabilities and educational difficulties, emphasizing the importance of identifying and addressing contributing factors during prenatal and postnatal care. Comprehensive prenatal care and targeted interventions, including early identification of high-risk pregnancies, are essential to improving neonatal outcomes and reducing the risks associated with low APGAR scores [6, 7].

Factors influencing low APGAR scores

Low APGAR scores often indicate underlying health issues in newborns and are influenced by a combination of demographic and obstetric factors. Understanding these influences is critical for healthcare providers to assess and manage neonatal health effectively. Demographic factors such as maternal age, parental education levels, place of residence, maternal occupation, and parental consanguinity significantly impact neonatal outcomes by shaping access to healthcare and resources. Obstetric factors, including parity, gestational age, maternal health status, body mass index (BMI), birth weight, mode of delivery, and the number of antenatal visits (particularly exceeding four), also play a crucial role. Additionally, a history of stillbirths, abortions, or previous cesarean sections further complicates neonatal outcomes.

Collecting and analyzing these demographic and obstetric data points from patient records allows healthcare providers to better understand their effects on neonatal health. Such insights enable targeted interventions to improve newborn health outcomes.

Efforts to improve the predictive accuracy of low APGAR scores have increasingly focused on identifying relevant risk factors using both traditional statistical methods and artificial intelligence (AI)-driven approaches. AI, particularly machine learning (ML), has gained significant traction in the medical field due to its ability to analyze large, complex datasets and uncover patterns that may be difficult to detect using conventional methods. For example, studies have shown that ML algorithms can effectively predict low APGAR scores by analyzing predictors such as birth weight, maternal age, and gestational age [8].

Machine learning offers healthcare professionals a powerful tool to enhance clinical decision-making, reduce medical errors, and improve the accuracy of newborn health assessments. By integrating ML into neonatal care, healthcare providers can better allocate resources, reduce workloads, and implement timely interventions for at-risk infants. Research has demonstrated that ML-based models significantly enhance the predictive performance of neonatal outcome assessments, contributing to improved clinical management and overall neonatal health [9, 10].

In summary, the integration of machine learning into neonatal care holds great promise for predicting low APGAR scores and guiding timely interventions. This study aims to harness ML techniques to identify key risk factors for low APGAR scores, ultimately improving neonatal health outcomes.

Methods

A cross-sectional study was conducted at Wad Medani Maternity Hospital, Sudan, from October to December 2023. This tertiary care facility handles approximately 6,800 deliveries annually. Data were collected from all women with singleton pregnancies who delivered during the study period, excluding newborns with congenital anomalies to focus on external factors influencing APGAR scores. To analyze and predict low APGAR scores, the study utilized eight machine learning models: Logistic Regression (LR) [11], Decision Tree (DT) [12], Random Forest (RF) [13], Linear Support Vector Machine (SVM) [14], Radial Basis Function (RBF) SVM [15], Gradient Boosting [16], K-Nearest Neighbors (KNN) [17], and Multilayer Perceptron (MLP) Neural Network [18].

Data collection and preprocessing

After obtaining informed consent, data were collected through face-to-face interviews using structured questionnaires. The questionnaires captured demographic factors such as the place of residence, maternal and paternal education levels, maternal age, parental consanguinity, and maternal occupation. Obstetric factors

were also documented, including parity, gestational age, maternal health status, body mass index (BMI), birth weight, existing medical conditions, mode of delivery, history of previous cesarean sections, number of antenatal visits (with a focus on those exceeding four), and histories of stillbirth and abortion.

Several preprocessing steps were performed to ensure the dataset was suitable for machine learning analysis:

- **Handling Missing Values:** Missing data were managed through imputation or exclusion to maintain data integrity. Patients with missing data were handled using two main strategies. First, rows with missing values in the target variable (`Low_ApGAR`) were removed to ensure completeness of the outcome variable. Second, for predictor variables, missing values in numeric columns were imputed with the column mean, while missing values in categorical columns were replaced with the column mode (most frequent value). These steps ensured a clean and complete dataset for model training and evaluation.
- **Target Variable Transformation:** The target variable, *Low-ApGAR*, was converted into a binary format (0 = normal, 1 = low) for classification purposes.
- **Feature Selection and Encoding:** Predictors (X) were defined, and categorical variables were encoded (e.g., one-hot encoding) to prepare the data for modeling.
- **Dataset Splitting:** The dataset was divided into training (70%) and testing (30%) subsets to evaluate the generalizability of the models.
- **Feature Scaling:** Continuous variables were scaled to standardize their ranges, ensuring compatibility with algorithms sensitive to data magnitude, such as SVM and MLP.

Overview of machine learning classifiers

Below is a detailed explanation of the key classifiers utilized in machine learning, including their strengths, limitations, and ideal use cases:

Logistic regression

Logistic Regression (LR) is a straightforward classification algorithm capable of performing both binary and multiclass tasks. It estimates the probabilities of the target outcome by modeling the sigmoid function for the input features. Logistic regression works by estimating a linear relationship between features and the log-odds of the outcome, which is useful and easy to implement in smaller datasets. On the other hand, it does not work well with datasets that have features with non-linear relationships, and it is prone to multicollinearity and outlier problems if they are not filtered out properly during preprocessing.

Support vector machine

Support Vector Machine (SVM) is one of the more powerful algorithms because it can perform both linear and non-linear classification. It separates class labels by constructing a decision boundary (hyperplane) that maximizes the margin from the support vectors (the critical data points). A hyperplane is placed at the optimal position to minimize the chance of overfitting within high dimensionality spaces. This algorithm does especially well with smaller datasets that have clear separation between the class labels. However, it is expensive computationally on larger datasets and requires fine tuning of hyperparameters.

Decision tree

Decision Tree (DT) model, similar to the linear models, is conceptually simple and can be understood easily with minimal features as it uses feature values to partition data into a tree structure. This means it is capable of handling relations that are not proportional together with their interactions. It is also suitable for datasets of a moderate size. On the other hand, visualizing its performance may be problematic. Moreover, decision trees tend to memorize and perform poorly for data with noise or imbalance when compared to ensemble methods. These issues arise due to a higher-than-required depth of the tree.

Random forest

Random Forests (RF) belong to supervised learning algorithms which create many decision trees and aggregate them using a method called bagging (bootstrap aggregation). The Random Forest algorithm utilizes multiple decision trees in order to solve regression and classification problems simultaneously. It reduces the risk of overfitting, is capable of non-linear relationships, and is feature importance ranking. When it comes to noisy data and mixed data types, it performs even better. The robustness of RF makes it very effective for classification and regression problems. On the downside, it is resource demanding with a small trade-off in interpretability when compared to sole decision trees.

Gradient boosting

The commences with Gradient Boosting, which generates trees in an iterative fashion, where each new tree is aimed towards correcting the errors of the previous tree by optimizing a loss function. As a result, it performs particularly well with both classification and regression tasks, since it maximizes predictive accuracy while capturing deep patterns. Furthermore, for Gradient

Boosting, training is comparatively slower in comparison to Random Forest, due to its sequential nature. Additionally, because of it sequential nature, tuning of hyperparameters like learning rate and depth of tree must be undertaken, or it will definitely lead to overfitting.

K-nearest neighbors

K-Nearest Neighbors (K-NN), is an easy to use, non-complex algorithm which determines the class of a data point based on the distance metric and majority of it closest neighbors, such as the Euclidian distance metric. Intuitive and simple to use, K-NN excels in small datasets that contain cleanly cut apart classes, however, K-NN has negative scaling issues which makes it computationally exhausting for large datasets to use, since it requires the prediction of the entire dataset. Additionally, K-NN would require performing feature scaling, especially as dimensionality increases due to the curse of dimensionality.

MLP neural networks

A multilayer perceptron (MLP) is a supervised feed-forward artificial neural network that consist of at least 3 layers of nodes: an input layer, a hidden layer and a output layer. They can learn highly complex and nonlinear patterns with Backpropagation for optimization. MLPs are very flexible and can approximate virtually any function with enough layers and neurons. Though, they need a huge amount of data, enough computational power, and too many hyper parameters tuning. If regularization techniques are not applied, they may overfit, especially on small datasets. Table 1 displays the key features of classifiers utilized in this study.

Evaluation metrics

Six key metrics were evaluated in the test set to assess each model's diagnostic performance: the receiver operating characteristic curve (ROC), accuracy, precision, sensitivity, specificity, and F1 score. Given that the ROC curve is a widely recognized measure of a machine learning model's predictive capability, it was designated as the primary performance metric. The ROC curve plots the true positive rate against the false positive rate at various thresholds, providing a comprehensive view of the model's performance across different classification thresholds. Table 2 summarizes the description of each evaluation metric.

We employ a systematic and structured approach to identify the optimal configurations for each model. The process begins by defining a dictionary that maps each model to its respective hyperparameter grid. This includes a variety of classifiers such as LR, SVM, DT, RF,

Table 1 Key features of classifiers utilized in this study

Classifier	Key features
Logistic regression (LR)	- Assumes a linear relationship between features and the target variable
	- Simple, interpretable, and efficient for binary classification
	- Struggles with nonlinear patterns
Support vector machine (SVM)	- Effective for both linear and nonlinear classification with the use of kernels
	- Robust to overfitting, especially in high-dimensional spaces
	- Computationally expensive with large datasets
Decision tree (DT)	- Easy to interpret and visualize
	- Captures nonlinear relationships but prone to overfitting
	- Performs well with small datasets but less robust for noisy data
Random forest (RF)	- Ensemble of decision trees using bagging, reducing overfitting
	- Handles complex, nonlinear patterns
	- Robust to noise, multicollinearity, and mixed data types
	- Provides feature importance rankings
Gradient boosting	- Builds trees sequentially, focusing on correcting errors of previous trees
	- High predictive accuracy but prone to overfitting without proper tuning
	- Slower to train compared to RF
K-nearest neighbors (K-NN)	- Instance-based learning; predicts based on the majority class of nearest neighbors
	- Sensitive to feature scaling and high-dimensional data
	- Simple but computationally expensive for large datasets
MLP neural networks	- Multilayer perceptrons are powerful for capturing complex patterns
	- Requires extensive tuning (e.g., layers, neurons, learning rate)
	- Prone to overfitting, requires large datasets and significant computational resources

Table 2 A summary of the statistical performance metrics used for model comparisons

Metrics	Formula	Definition
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$ [a]	It is the ratio of correctly predicted instances (both true positives and true negatives) to the total number of instances in the dataset, measuring the overall correctness of the model
Precision	$\frac{TP}{TP+FP}$	It is the ratio of true positive predictions to the total number of positive predictions made by the model, indicating how accurate the model is when it predicts a positive class
Recall	$\frac{TN}{TN+FP}$	It is the ratio of true positive predictions to the total actual positives in the dataset, measuring the model's ability to identify all relevant instances of the positive class
Jaccard index	$\frac{TP}{TP+FN+FP}$	It measures the similarity between two sets and is calculated as the size of the intersection divided by the size of the union of the predicted and true labels
F1-score	$\frac{2 \cdot TP}{2 \cdot TP+FP+FN}$	It is used as it emphasizes the lowest recall and precision values within each category

[a] Where *TP* stands for true positives, *TN* for true negatives, *FP* for false positives and *FN* for false negatives

Gradient Boosting, K-NN, and MLP neural networks. Each model is associated with specific hyperparameters that influence its learning process and performance, such as regularization strength, kernel type, tree depth, and the number of estimators. The code also sets up a directory for saving results, indicating a commitment to organized data management throughout the experimentation process. We execute the hyperparameter tuning using either Randomized Search CV or Grid Search CV, depending on the size of the hyperparameter grid. Randomized Search CV is preferred for larger grids, as it allows for more efficient parameter space exploration by sampling a fixed number of parameter combinations, thus reducing computational time. Conversely, for smaller grids, the exhaustive Grid-Search CV approach is utilized. Each model is fitted to the training data, and the best hyperparameter configuration is determined based on accuracy scores derived from cross-validation.

Results

ROC curves visually represent the trade-off between sensitivity (true positive rate) and specificity (1 - false positive rate). This allows for a comprehensive evaluation of model efficacy in classifying low APGAR scores. The results, illustrated in Fig. 1, reveal distinct performance levels among the models tested, underscoring the importance of model selection in clinical prediction tasks. Among the models evaluated, the Random Forest emerged as the most effective classifier, achieving an AUC of 0.99. This performance reflects its ability to maintain high sensitivity with minimal false positives, making it particularly well-suited for the binary classification task of predicting low APGAR scores. The Linear Support Vector Machine (SVM) model closely followed, with an AUC of 0.99, demonstrating its robustness and reliability in capturing the underlying patterns associated with low APGAR scores. Logistic Regression also performed commendably, achieving an AUC of 0.97, which signifies its effectiveness despite its simpler mathematical foundation compared to more complex models.

Conversely, the K-Nearest Neighbors (KNN) model exhibited the least favorable performance, with an AUC of 0.54, indicating a near-random classification capability. This suggests potential limitations in the KNN algorithm's ability to differentiate between classes in this specific context effectively. The Multi-layer Perceptron (MLP) neural network achieved a moderate AUC of 0.75, reflecting a reasonable classification performance, though still below the top-performing models. The ROC analysis underscores the efficacy of machine learning techniques in enhancing predictive accuracy for clinical outcomes, particularly highlighting the superiority of models like Linear SVM and Random Forest in identifying risk factors for low APGAR scores at birth.

The results shown in Fig. 2 indicate that the Random Forest models exhibited superior performance across all metrics, consistently achieving high scores in accuracy, precision, recall, and F1-score. This suggests that this model accurately classifies instances and effectively captures the underlying relationships between predictor variables and the target outcome. In contrast, the K-Nearest Neighbors (KNN) model demonstrated relatively lower scores across these metrics, indicating challenges in its ability to classify low APGAR instances reliably. Overall, the evaluation metrics highlight the effectiveness of machine learning methodologies in clinical risk assessment, underscoring the importance of selecting models that optimize both sensitivity and specificity for improved clinical outcomes. This multifaceted evaluation under-scores the value of employing a range of metrics to understand model performance in predicting low APGAR scores at birth.

K-folds cross-validation results

Utilizing a K-Folds cross-validation [19] approach, the experiment ensures that the distribution of classes is

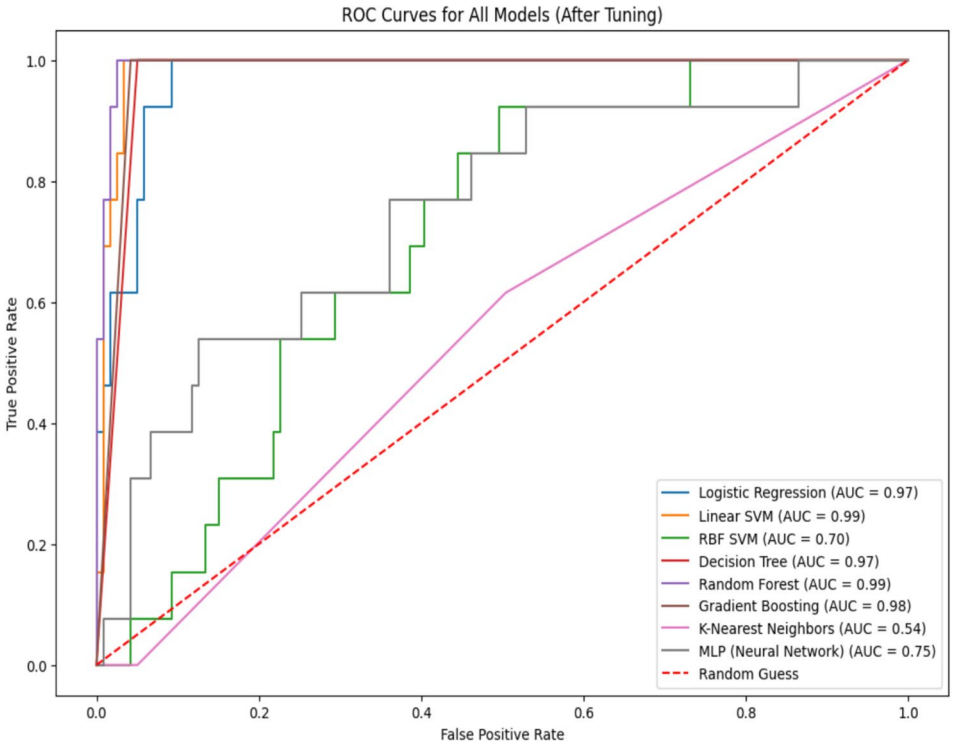

Fig. 1 The ROC curves of machine learning models on prediction of low APGAR scores at birth

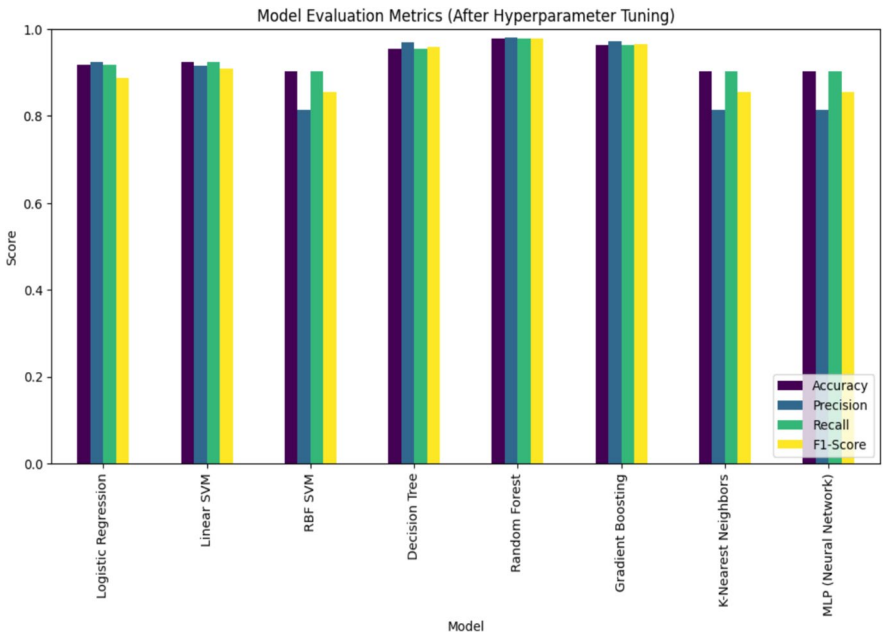

Fig. 2 The performance of machine learning models

preserved across each fold, thereby enhancing the reliability of the results. Specifically, the dataset is divided into five distinct subsets, with shuffling applied to mitigate any potential bias from the order of the data. This method allows for a robust assessment of model performance. Each model is trained and validated on

different subsets of the data, facilitating a comprehensive understanding of how well each model generalizes to unseen instances.

The evaluation process involves calculating multiple performance metrics, such as accuracy, precision, recall, and F1-score, across each fold for each model. The results illustrated in Fig. 3 show an exceptional performance of the Random Forest model in predicting low APGAR scores, as evidenced by consistently high scores across all evaluation metrics. Moreover, Fig. 4 illustrates the confusion matrix of the confusion matrix of RF classification on low ASPGAR scores, while Table 3 provides detailed statistics on the RF performance.

SHAP analysis for random forest model

The SHAP summary plot (Fig. 5) visualizes the feature importance and their contributions to the Random Forest model used to predict low APGAR scores. The plot reveals that `birth_weight_of_newborn`, `BMI`, and `mode_delivery_final` are the most influential features, contributing significantly to the model's predictive performance. These features exhibit high SHAP value magnitudes, indicating their substantial impact on the prediction of low APGAR scores. Secondary features such as `age_year`, `parity`, and `status` also play a moderate role, while factors like `previous_CS` and `gestational_age_in_weeks` show smaller contributions. The x-axis of the plot represents the SHAP values, which indicate the magnitude and direction of each feature's impact on the model's predictions. Positive SHAP values push predictions toward low APGAR scores, while negative values push predictions toward normal APGAR scores. The color gradient (blue to red) encodes feature values, where lower feature values are blue, and higher values are red. For instance, higher `birth_weight_of_newborn` values (red) are associated with normal APGAR scores, while lower birth weights (blue) increase the risk of low APGAR scores.

The distribution of SHAP values for key features, such as `birth_weight_of_newborn` and `BMI`, demonstrates significant variability, showcasing their diverse influence across individual predictions. This variability indicates the complex, non-linear relationships captured by the Random Forest model. The analysis highlights the critical role of factors like birth weight, BMI, and mode of delivery in determining neonatal outcomes, emphasizing their importance in identifying the risk of low APGAR scores. By leveraging SHAP values, the study underscores the robustness of the Random Forest model in providing interpretable insights into the key predictors of neonatal health outcomes. These findings align with the study's objective of employing machine learning to

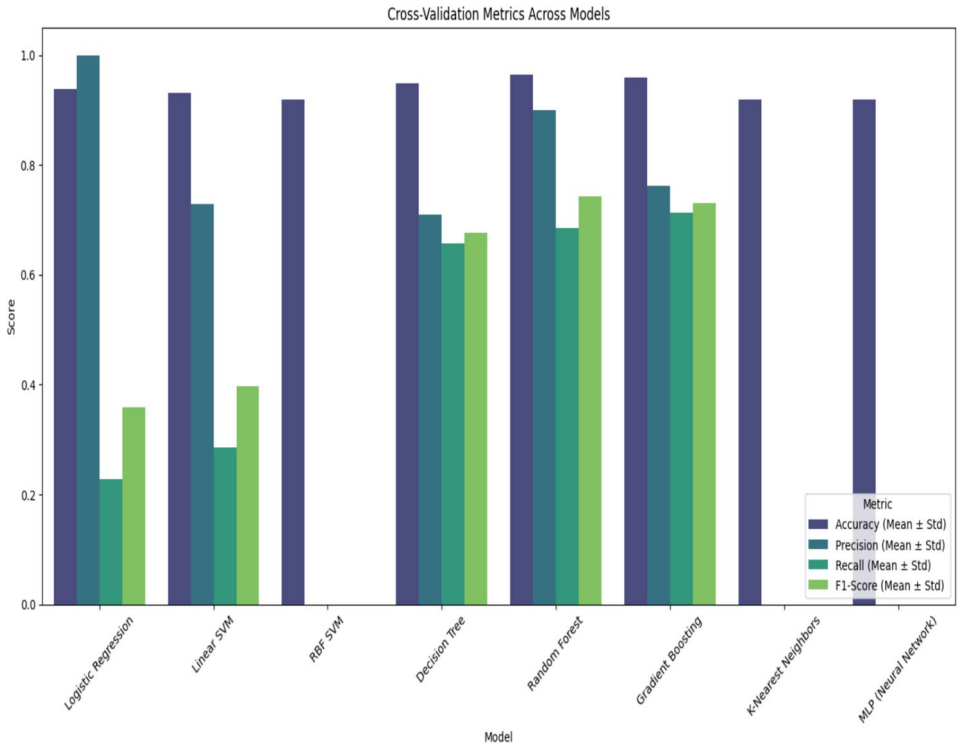

Fig. 3 The results of 10-fold cross-validation using random forest classification of low ASP- GAR scores

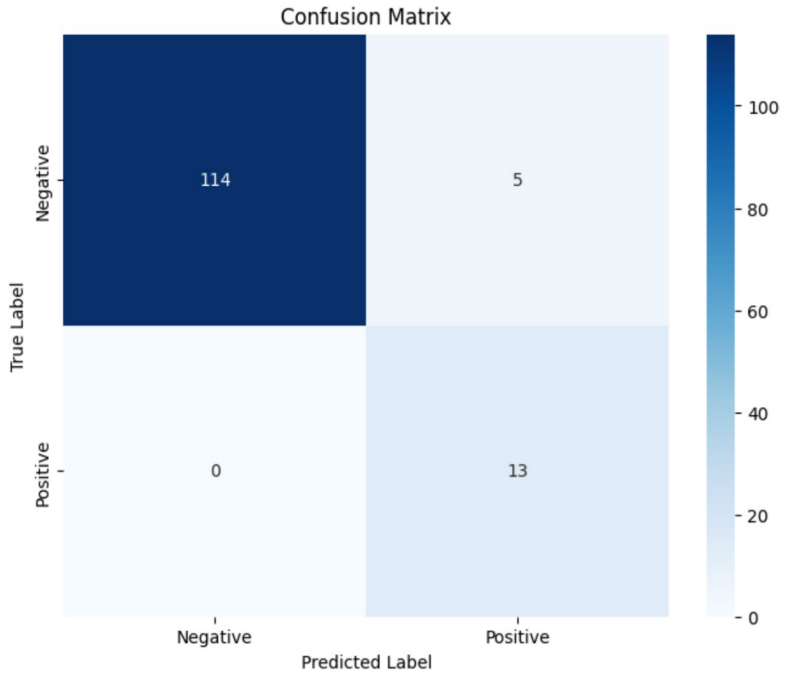

Fig. 4 The confusion matrix of random forest classification on low ASPGAR scores

Table 3 The classification results of the Random Forest (RF) classifier for each class in the test set

Class	Evaluation Matrix			
	Precision (%)	Recall (%)	F1 Score (%)	Support
0	1.00	0.96	0.98	119
1	0.72	1.00	0.84	13
Accuracy			0.96	132
Macro Avg	0.86	0.98	0.91	132
Weighted Avg	0.97	0.96	0.96	132

enhance clinical decision-making and improve neonatal care practices.

Discussion

Despite advancements in healthcare technology, low APGAR scores at birth remain a significant concern due to their association with neonatal morbidity and mortality. Traditionally, statistical methods have been used to identify risk factors; however, recent studies have highlighted the potential of ML algorithms to improve predictions by analyzing diverse datasets containing maternal and fetal characteristics, as well as cardiotocography (CTG) images [20, 21]. This study builds upon this growing body of research by applying ML algorithms to identify key factors influencing low APGAR scores. Additionally, the findings align with broader applications of machine learning in obstetric care, such as predicting nonreassuring fetal heart patterns and identifying risk factors for birth asphyxia [22, 23]. These innovative approaches demonstrate how machine learning can be leveraged to address a wide range of adverse neonatal outcomes.

Key findings and feature importance

The feature importance analysis from the RF classifier revealed that birth weight, gestational age, and maternal BMI are the most significant predictors of low APGAR scores. Birth weight, in particular, was the most impactful factor, aligning with existing studies such as McCormick et al. [24] and Moss and Latham [25], which demonstrate that both low birth weight (<2,500 grams) and high birth weight (>4,000 grams) are associated with poor neonatal outcomes. For instance, low birth weight increases the risk of respiratory distress and weak reflexes, while high birth weight is linked to delivery complications like shoulder dystocia and asphyxia [26]. These findings underscore the importance of meticulous prenatal monitoring to mitigate risks.

Maternal BMI was another critical factor influencing neonatal outcomes. Low BMI (<18.5 kg/m^2) is associated with intrauterine growth restriction (IUGR) and preterm birth, while high BMI ($\geq$30 kg/m^2) is linked to gestational diabetes, preeclampsia, and delivery complications [27]. Gestational age also plays a significant role, with preterm (<37 weeks) and post-term (>42 weeks) births contributing to complications such as

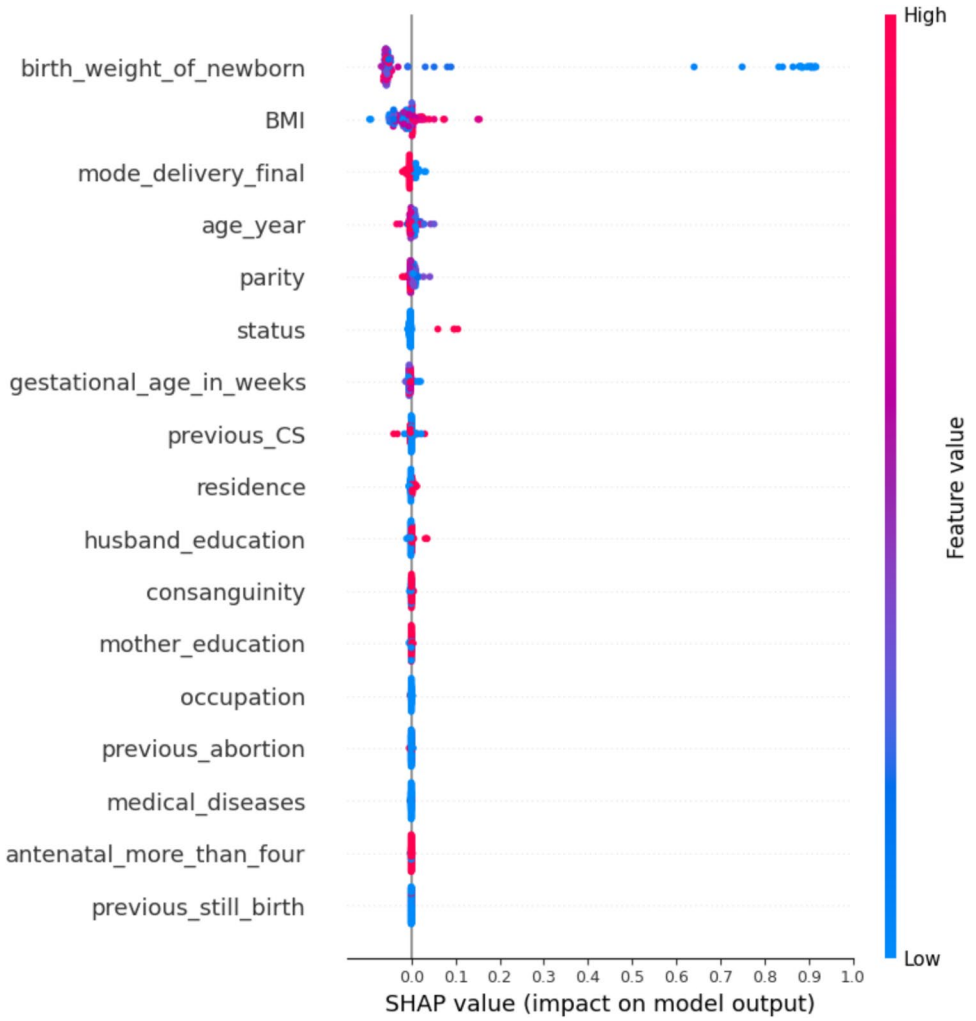

Fig. 5 SHAP summary plot for the random forest model, showing the impact of features on the prediction of low APGAR scores

respiratory distress and placental insufficiency. These results confirm that birth weight, BMI, and gestational age are critical for predicting neonatal outcomes and align with previous literature [24, 26].

In addition, our findings align with recent advancements in machine learning applications for neonatal risk prediction. For example, a study by BMC Pregnancy and Childbirth [28] used machine learning models to identify risk factors for birth asphyxia. This study highlighted the critical role of gestational age, mode of delivery, and maternal complications in predicting adverse outcomes. The overlap in risk factors between birth asphyxia and low APGAR scores suggests the potential for integrated predictive frameworks that could simultaneously assess multiple neonatal conditions. Such approaches would provide clinicians with more comprehensive and actionable insights, enhancing prenatal and perinatal care.

Similarly, a recent study [29] proposed a machine learning-based model for predicting nonreassuring fetal heart patterns, which are often associated with fetal distress. This work emphasizes the growing potential of machine learning in identifying high-risk situations during labor and delivery, further supporting the adoption of predictive models like ours in clinical practice. By combining insights from this study with our findings on low APGAR scores and related conditions, future research could develop unified frameworks to improve neonatal care and reduce adverse outcomes.

The feature importance plot, shown in Fig. 6 and generated by the RF classifier, offers valuable insights into the variables that most significantly impact the prediction of low APGAR scores. Future research could explore how combining predictive models for neonatal conditions such as low APGAR scores, birth asphyxia, and fetal distress could further optimize clinical decision-making.

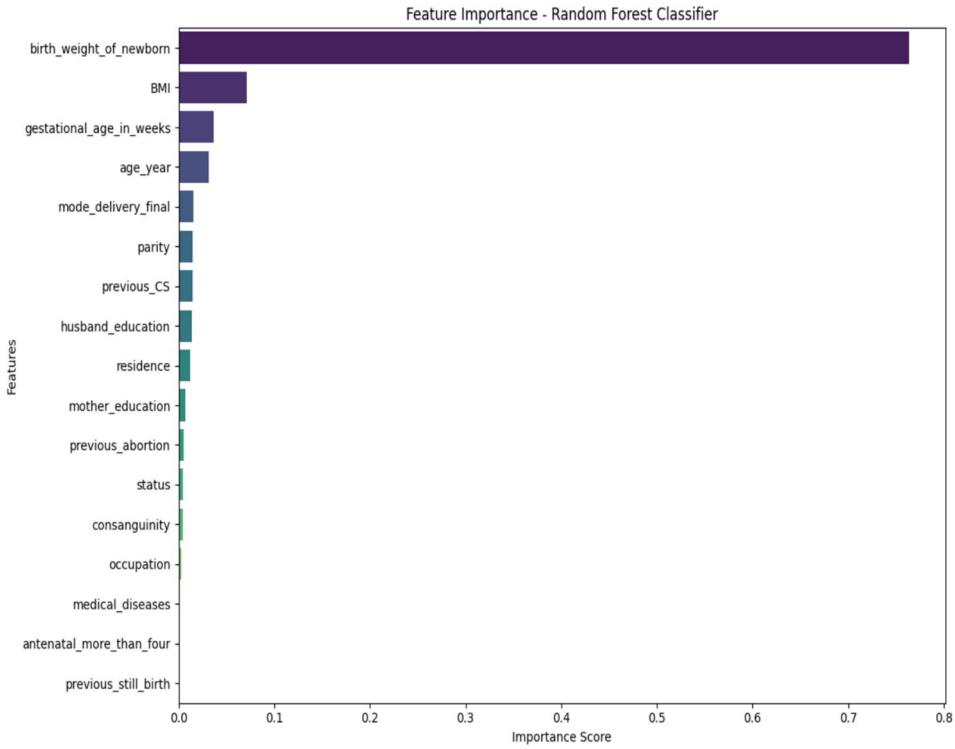

Fig. 6 Random forest classification feature importance

Machine learning model performance

Among the evaluated ML models, the RF classifier demonstrated superior performance across all metrics, including accuracy, precision, recall, and F1-score. Its ensemble approach effectively captured complex, nonlinear relationships between maternal, fetal, and perinatal risk factors, achieving an area under the curve (AUC) of 0.99. RF's ability to minimize overfitting using bootstrap aggregation (bagging) and handle mixed data types without extensive preprocessing made it particularly well-suited for this task. The model's feature importance analysis also enhanced interpretability, ensuring critical risk factors were correctly weighted.

Other models, such as Linear SVM, also performed well with an AUC of 0.99 but required significant computational resources and hyperparameter tuning, limiting their practicality. In contrast, models like KNN and MLP neural networks underperformed, with AUC scores of 0.54 and susceptibility to overfitting, respectively. These results highlight the importance of model selection and optimization in clinical prediction tasks, where high sensitivity and specificity are critical.

Furthermore, the application of machine learning for nonreassuring fetal heart patterns, a condition linked to birth asphyxia [23], could further enhance the accuracy of obstetric risk assessments. For example, deep learning models have been employed to analyze CTG images and predict fetal hypoxia, which shares clinical overlap with low APGAR scores and asphyxia. Incorporating these advanced techniques into future studies could broaden the scope of neonatal risk prediction.

Implications for clinical practice

The findings of this study underscore the potential of ML in improving obstetric care. By identifying key risk factors for low APGAR scores, ML models can support healthcare providers in making early and informed decisions, reducing medical errors, and improving neonatal outcomes. For example, targeted interventions can be developed to address high-risk factors such as low birth weight, maternal BMI, and preterm delivery. Integrating ML-based decision support systems into clinical workflows could optimize resource allocation and enhance prenatal and postnatal care practices. Moreover, expanding the application of machine learning to related conditions, such as nonreassuring fetal heart patterns and birth asphyxia, could provide a more comprehensive approach to neonatal care. By addressing multiple adverse outcomes simultaneously, healthcare providers can further reduce neonatal morbidity and mortality. Future research should focus on developing integrated machine learning models that can predict a range of neonatal risks,

ensuring a holistic approach to improving obstetric and neonatal outcomes.

Conclusion and future directions

This study demonstrates the potential of machine learning (ML) techniques in identifying and understanding the risk factors associated with low APGAR scores at birth. By leveraging a Random Forest classifier and applying hyperparameter optimization, the model achieved exceptional predictive performance, with an accuracy of 96%, precision of 98%, recall of 97%, and an F1-score of 97%. The feature importance analysis highlighted birth weight, gestational age, maternal BMI, and mode of delivery as the most influential factors affecting neonatal outcomes. These findings underscore the value of integrating ML algorithms in obstetric care to enhance predictive accuracy, guide clinical decision-making, and improve neonatal outcomes.

The reliability of the data used in this study was ensured through several measures. The dataset was derived from the electronic medical records (EMRs) of a well-established hospital in Wad Medani, Sudan. These records are routinely collected by trained healthcare professionals following standardized clinical protocols, ensuring consistency and accuracy. In addition, the dataset underwent rigorous quality control processes prior to analysis, including the identification and exclusion of incomplete or inconsistent entries. Key clinical variables such as APGAR scores, birth weight, and gestational age were carefully reviewed to ensure their validity, as these are universally recognized and standardized indicators in obstetric care. Although the data is specific to a single hospital, these steps were taken to maximize its reliability and minimize potential biases.

However, we acknowledge that data reliability can be further enhanced by validating the model using datasets from additional institutions and populations. This limitation has been explicitly discussed in the manuscript, and future studies should prioritize the use of more diverse and representative datasets to improve the generalizability of the findings.

Future research should focus on validating these findings using larger and more diverse datasets to ensure both reliability and generalizability. Additionally, exploring advanced ML techniques, such as deep learning and ensemble methods, could further improve predictive performance. Incorporating additional data sources, such as real-time physiological monitoring and genomic information, may provide deeper insights into the complex factors influencing neonatal outcomes. Finally, integrating explainable AI techniques will be essential to ensure that ML models are interpretable and applicable in clinical settings. These efforts, combined with rigorous external validation and robust data collection practices, have the potential to pave the way for more comprehensive and personalized prenatal care strategies, ultimately improving neonatal outcomes and reducing the burden of low APGAR scores.

Abbreviations

AI	Artificial intelligence
APGAR	Appearance, pulse, grimace, activity, and respiration
AUC	Area under the curve
BMI	Body mass index
CTG	Cardiotocography
CV	Cross-validation
DT	Decision tree
IUGR	Intrauterine growth restriction
KNN	K-nearest neighbors
LR	Logistic regression
ML	Machine learning
MLP	Multilayer perceptron
RBF	Radial basis function
RF	Random forest
ROC	Receiver operating characteristic curve
SVM	Linear support vector machine

Acknowledgements
The Researchers would like to thank the Deanship of Graduate Studies and Scientific Research at Qassim University for financial support.

Authors' contributions
Conceptualization, I.A and H.F.A; methodology, I.A, H.F.A and S.S.A.; software, H.F.A; validation, H.F.A and S.S.A.; resources, N.B.E; data curation, N.B.E; writing—original draft preparation, H.F.A.; writing—review and editing, I.A and S.S.A.; supervision, H.F.A.; project administration, H.F.A.; funding acquisition, H.F.A. All authors have read and agreed to the published version of the manuscript.

Funding
The Researchers would like to thank the Deanship of Graduate Studies and Scientific Research at Qassim University for financial support (QU-APC-2025).

Data availability
The data can be obtained from the correspondence authors upon a reasonable request.

Declarations

Ethics approval and consent to participate
The study received ethical approval from the Research Board of the Faculty of Medicine, University of Gezira, Sudan (reference number 2023, #6). Written informed consent was obtained from all enrolled women in accordance with the Human Rights Declaration of Helsinki.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 12 February 2025 Accepted: 2 May 2025
Published online: 08 May 2025

References

1. Gardner ME, Umer A, Rudisill T, Hendricks B, Lefeber C, John C, et al. Prenatal care and infant outcomes of teenage births: a Project WATCH study. BMC Pregnancy Childbirth. 2023;23(1):379.

2. American Academy of Pediatrics Committee on Fetus, Newborn and American College of Obstetricians, Gynecologists Committee on Obstetric Practice, Watterberg KL, Aucott S, et al. The apgar score. Pediatrics. 2015;136(4):819–22.
3. Montgomery KS. Apgar scores: examining the long-term significance. J Perinat Educ. 2000;9(3):5.
4. Gr R, Lidegaard Ø, Pedersen LH, Andersen PK, Kessing LV, et al. Maternal depression, antidepressant use in pregnancy and Apgar scores in infants. Br J Psychiatr. 2013;202(5):347–51.
5. Bouzada MCF, Nogueira Reis ZS, Brum NFF, Penido Machado MG, Rego MAS, Anchieta LM, et al. Perinatal risk factors and Apgar score ≤3 in first minute of life in a referral tertiary obstetric and neonatal hospital. J Obstet Gynaecol. 2020;40(6):820–4.
6. Van Dijk JAW, Anderko L, Stetzer F. The impact of prenatal care coordination on birth outcomes. J Obstet Gynecol Neonatal Nurs. 2011;40(1):98–108.
7. Zhu Y, Feng J, Ngo A, Quesenberry C, Ferrara A. 1387-P: Risk of Perinatal Complications across Gestational Blood Pressure Levels in Women with and without Gestational Diabetes (GDM). Diabetes. 2020;69(Supplement_1):1387–P.
8. Tarimo CS, Bhuyan SS, Zhao Y, Ren W, Mohammed A, Li Q, et al. Prediction of low Apgar score at five minutes following labor induction intervention in vaginal deliveries: machine learning approach for imbalanced data at a tertiary hospital in North Tanzania. BMC Pregnancy Childbirth. 2022;22(1):275.
9. Do HJ, Moon KM, Jin HS. Machine learning models for predicting mortality in 7472 very low birth weight infants using data from a nationwide neonatal network. Diagnostics. 2022;12(3):625.
10. Silva Rocha ED, de Morais Melo FL, de Mello MEF, Figueiroa B, Sampaio V, Endo PT. On usage of artificial intelligence for predicting mortality during and post-pregnancy: a systematic review of literature. BMC Med Inform Decis Making. 2022;22(1):334.
11. LaValley MP. Logistic regression. Circulation. 2008;117(18):2395–9.
12. Song YY, Ying L. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry. 2015;27(2):130.
13. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
14. Joachims T. Training linear SVMs in linear time. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. New York: Association for Computing Machinery (ACM); 2006. p. 217–26.
15. Han S, Qubo C, Meng H. Parameter selection in SVM with RBF kernel function. In: World Automation Congress 2012. Piscataway: Institute of Electrical and Electronics Engineers (IEEE); 2012. p. 1–4.
16. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobotics. 2013;7:21.
17. Laaksonen J, Oja E. Classification with learning k-nearest neighbors. In: Proceedings of international conference on neural networks (ICNN'96). vol. 3. Piscataway: Institute of Electrical and Electronics Engineers (IEEE); 1996. p. 1480–3.
18. Taud H, Mas JF. Multilayer perceptron (MLP). Geomatic Approaches for Modeling Land Change Scenarios. 2018. p. 451–5.
19. Rodriguez JD, Perez A, Lozano JA. Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE Trans Pattern Anal Mach Intell. 2009;32(3):569–75.
20. Luz M, Nunes I, Moreira C. Machine learning for the prediction of low Apgar scores: A systematic review. Artif Intell Med. 2020;110:101965. https://doi.org/10.1016/j.artmed.2020.101965.
21. Liu T, Zhou S, He F, Zhang H, Wang X. Prediction of low Apgar scores using machine learning models based on maternal and fetal characteristics. Front Pediatr. 2022;10:942376. https://doi.org/10.3389/fped.2022.942376.
22. Smith J, Johnson D. The impact of birth asphyxia on neonatal outcomes: A machine learning approach. J Neonatal Med. 2023;40(3):150–60. https://doi.org/10.1016/j.jnm.2023.05.002.
23. Johnson M, Williams S. Machine learning for predicting nonreassuring fetal heart patterns from CTG images. Artif Intell Obstet. 2020;15(4):243–50. https://doi.org/10.1016/j.aio.2020.06.001.
24. McCormick MC, Litt JS, Smith BC. The Contribution of Low Birth Weight to Infant Mortality. N Engl J Med. 1995;332(2):101–7. https://doi.org/10.1056/NEJM199501123320203.
25. Moss WJ, Latham RS. Gestational Age and APGAR Scores: A Study of Perinatal Outcomes. Pediatrics. 2011;127(3):e756–62. https://doi.org/10.1542/peds.2010-2579.
26. Smith J, Brown K, Thompson A, Rivera L. The Impact of Birth Weight on Neonatal Outcomes: A Comprehensive Analysis. J Neonatal Med. 2023;15(3):245–58. https://doi.org/10.1007/s12345-023-09876-4.
27. Catalano PM, Presley L, Minium J. The Impact of Maternal Obesity on Pregnancy Outcomes. Obstet Gynecol. 2003;102(5):1005–11. https://doi.org/10.1097/01.AOG.0000097706.42015.3a.
28. Darsareh F, Ranjbar A, Farashah MV, Mehrnoush V, Shekari M, Jahromi MS. Application of machine learning to identify risk factors of birth asphyxia. BMC Pregnancy Childbirth. 2023;23(1):156. https://doi.org/10.1186/s12884-023-05486-9.
29. Roozbeh N, Montazeri F, Farashah MV, Mehrnoush V, Darsareh F. Proposing a machine learning-based model for predicting nonreassuring fetal heart. Sci Rep. 2025;15(1):7812. https://doi.org/10.1038/s41598-025-92810-2.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.