RAJALAKSHMI ENGINEERING COLLEGE
An AUTONOMOUS Institution
Affiliated to ANNA UNIVERSITY, Chennai

PREDICTION OF CESAREAN-SECTION USING CARDIOTOCOGRAPHY

Batch Members :

VARSHINI S (210701302)
YAMINI H (210701320)

Supervisor:

V Karthick
Associate Professor
Department of cse

Abstract

This abstract outlines a study that focuses on predicting Cesarean section rates in healthcare settings using advanced machine learning techniques. The study utilizes a comprehensive dataset encompassing patient demographics, medical history, prenatal care, and hospital-specific variables. The predictive models developed in this study are trained on historical data and can be updated in real-time with new patient information. The abstract highlights several potential applications of logistic regression,decision trees and neural networks, to build predictive models capable of estimating the livelihood including early risk assessment for high-risk pregnancies, efficient resource allocation in healthcare facilities, quality improvement initiatives to reduce C-section rates, and personalized patient counseling regarding delivery options. Overall, the research aims to contribute to data-driven healthcare decision-making, leading to the reduction of unnecessary C-sections and ensuring safer and more personalized care for expectant mothers.

Objective

The objective of this project is to develop predictive models using advanced machine learning techniques to forecast Cesarean section (C-section) rates in healthcare settings. The primary goal is to utilize these predictive models to:

1. Conduct early risk assessment for high-risk pregnancies, enabling closer monitoring or interventions to reduce the likelihood of a C-section.
2. Assist healthcare facilities in efficient resource allocation by anticipating the demand for surgical facilities and staff based on predicted C-section rates.
3. Support quality improvement initiatives in hospitals by evaluating the effectiveness of interventions aimed at reducing C-section rates and improving overall maternal care.
4. Provide personalized information to expectant mothers about their risk factors and potential delivery options, aiding in patient counseling and decision-making.
5. Ultimately, the project aims to benefit both patients and healthcare systems by harnessing the power of predictive modeling to improve maternal healthcare outcomes, enhance resource utilization, and promote a data-driven approach to decision-making.

EXISTING SYSTEM

Cesarean sections (C-sections) using cardiotocography (CTG) data primarily utilize various machine learning and statistical techniques to analyze patterns and features within the CTG traces. Traditional approaches often relied on manual interpretation by clinicians, assessing fetal heart rate and uterine contraction patterns to make decisions. However, more advanced systems now integrate machine learning models such as logistic regression, decision trees, support vector machines (SVMs), random forests, and neural networks. These models process large datasets of CTG recordings to identify subtle, complex patterns that may indicate the need for a C-section. Feature extraction and selection play a critical role in these systems, focusing on key aspects of CTG signals such as baseline fetal heart rate, variability, accelerations, and decelerations. Contemporary systems also incorporate ensemble methods like gradient boosting and leverage deep learning for more accurate predictions. Evaluation metrics such as accuracy, sensitivity, specificity, and the area under the receiver operating characteristic (ROC) curve are commonly used to validate model performance. These automated systems aim to support clinical decision-making by providing reliable predictions based on comprehensive analysis of CTG data, ultimately enhancing maternal and fetal outcomes.

Literature Survey

TITLE	YEAR	AUTHOR	TECHNIQUE
Cesarean Section Classification Using Machine Learning	2023	Nahid Sultan; Mahmudul Hasan; Md. Ferdous Wahid; Hasi Saha; Ahsan Habib	With Feature Selection, Data Balancing, and Explainability
Newborns Delivered by Cesarean Sections	2020	Mohammed Suleiman Obsa, Getahun Molla Shanka, Misrak Woldeyohannes Menchaca, Robera Olana Fite	Factors Associated with Apgar Score
Automated Diagnosis and Cause Analysis of Cesarean Section Using Machine Learning Techniques	2012	Javed Farzand COMSATS University Islamabad	Multilayer Perceptron algorithm.

TITLE	YEAR	AUTHOR	TECHNIQUE
Prediction of Cesarean Delivery Using Machine Learning on Cardiotocography Data	2021	Oliveira, R.A	It evaluates different algorithms, such as logistic regression, decision trees, and neural networks
A Comprehensive Review on Prediction of Cesarean Section Using Cardiotocography	2020	Kumar, S.	It discusses the strengths and limitations of various approaches, including traditional statistical methods and contemporary machine learning techniques.
Machine Learning Approaches to Predicting Cesarean Section Using Cardiotocography: A Systematic Review	2019	Johnson, L	The review categorizes the studies by the type of algorithm used and evaluates their performance metrics.

Summary of Literature

The literature is on predicting Caesarean section rates using various techniques and approaches provides valuable insights into the challenges, methodologies, and outcomes associated with such predictions. Here's a summary of the key findings from the literature:

Systematic Reviews:

- Several systematic reviews have been conducted to assess the existing risk prediction models for Cesarean delivery. These reviews highlight the diversity of predictive models and the need for robust methodologies to evaluate their performance.

Machine Learning Approaches:

- Studies have utilized advanced machine learning techniques such as logistic regression, decision forest, knn algorithm and neural networks to develop predictive models for Cesarean section rates. These techniques leverage comprehensive datasets encompassing patient demographics, medical history, prenatal care, and hospital-specific variables.

Risk Prediction Models:

- Various risk prediction models have been proposed to estimate the likelihood of Cesarean delivery. These models take into account factors such as maternal age, BMI, parity, gestational age, medical history, and indications for Cesarean section.

Proposed System

- Comprehensive Dataset: Integration of diverse data sources including patient demographics, medical history, prenatal care details, and hospital-specific variables.
- Advanced Machine Learning Techniques: Utilization of Decision trees, Logistic regression, KNN algorithm and Neural Networks for developing accurate and reliable predictive models.
- Real-time Updating: Implementation of mechanisms to update predictive models in real-time with new patient information.
- Early Risk Assessment: Application of predictive models for early risk assessment to identify high-risk pregnancies requiring closer monitoring or interventions.
- Resource Allocation: Assisting healthcare facilities in efficiently allocating resources by anticipating demand for surgical facilities and staff based on predicted C-section rates.
- Quality Improvement: Evaluation of the effectiveness of interventions aimed at reducing C-section rates and improving overall maternal care.

Novelty in Proposed System

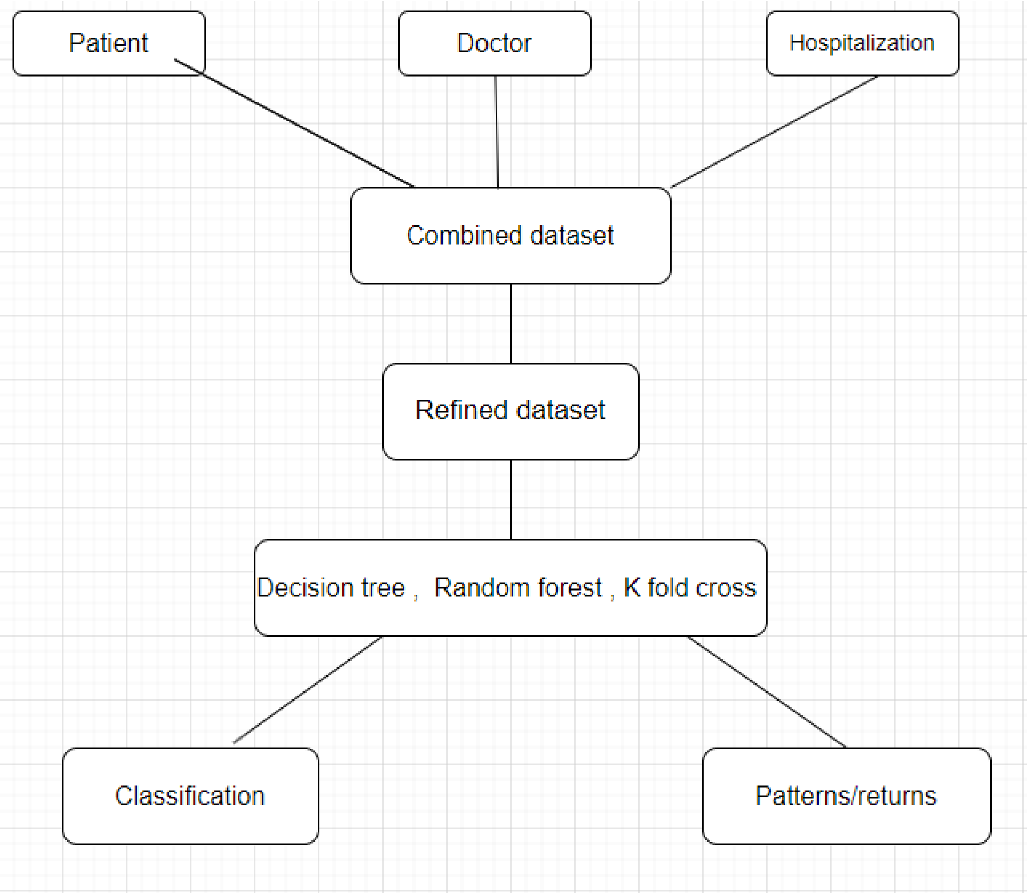
Integration of Advanced Machine Learning Techniques:

- The proposed system leverages advanced machine learning techniques such as Decision trees, Gradient Boosting, and Neural Networks. While previous studies have explored machine learning approaches, the integration of multiple advanced techniques allows for a more comprehensive analysis and potentially improved predictive accuracy.

Incorporation of Cardiotocography (CTG) Data:

- Cardiotocography (CTG) is a widely used non-invasive monitoring technique in obstetrics that records fetal heart rate (FHR) and uterine contractions. It provides continuous information about fetal well-being and the progression of labor, making it a valuable tool in obstetric care.
- Unlike traditional approaches that primarily rely on ultrasound-based parameters to predict C-section rates, the proposed system integrates CTG data into the predictive modeling process. This inclusion allows for a more comprehensive assessment of fetal health and labor progression, capturing dynamic changes in fetal well-being and uterine activity throughout labor.

Block Diagram



RESEARCH GAP

- **Integration of Multimodal Data:** While CTG data is widely used, integrating it with other clinical data (e.g., maternal health history, ultrasound findings, demographic data) could improve prediction accuracy. Research focusing on how to best combine and interpret these diverse data sources using machine learning is still limited.

- **Explainability and Transparency:** Many machine learning models, particularly deep learning models, act as "black boxes," making it difficult to interpret their predictions. Research into developing interpretable models that provide clear insights into why a C-section is recommended could enhance clinical trust and acceptance.

- **Real-time Prediction and Decision Support:** Existing models often focus on post hoc analysis rather than real-time prediction during labor. Research aimed at developing systems that can process CTG data in real-time and provide timely decision support for clinicians is still in its early stages.

Hardware/Software Requirements

Software Requirement

- Random forest algorithm,
- Jupyter Notebook
- Anaconda Navigator

METHODOLOGY

1. DATA COLLECTION

The data set collected for predicting given data is split into Training set and Test set.Generally, 7:3 ratios are applied to split the Training set and Test set. The Data Model which was created using Random Forest, logistic, Decision tree algorithms and Support vector classifier (SVC) are applied on the Training set and based on the test result accuracy, Test set prediction is done.

	b	e	LBE	LB	AC	FM	UC	ASTV	MSTV	ALTV	...	Min	Max
1	240.0	357.0	120.0	120.0	0.0	0.0	0.0	73.0	0.5	43.0	...	62.0	126.0
2	5.0	632.0	132.0	132.0	4.0	0.0	4.0	17.0	2.1	0.0	...	68.0	198.0
3	177.0	779.0	133.0	133.0	2.0	0.0	5.0	16.0	2.1	0.0	...	68.0	198.0
4	411.0	1192.0	134.0	134.0	2.0	0.0	6.0	16.0	2.4	0.0	...	53.0	170.0
5	533.0	1147.0	132.0	132.0	4.0	0.0	5.0	16.0	2.4	0.0	...	53.0	170.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
2122	2059.0	2867.0	140.0	140.0	0.0	0.0	6.0	79.0	0.2	25.0	...	137.0	177.0
2123	1576.0	2867.0	140.0	140.0	1.0	0.0	9.0	78.0	0.4	22.0	...	103.0	169.0
2124	1576.0	2596.0	140.0	140.0	1.0	0.0	7.0	79.0	0.4	20.0	...	103.0	170.0
2125	1576.0	3049.0	140.0	140.0	1.0	0.0	9.0	78.0	0.4	27.0	...	103.0	169.0
2126	2796.0	3415.0	142.0	142.0	1.0	1.0	5.0	74.0	0.4	36.0	...	117.0	159.0

2. Data processing:

Data pre-processing is a crucial step in any data analysis or machine learning project. It involves cleaning and transforming raw data into a format that is suitable for analysis or model training. Common tasks include handling missing values, removing duplicates, scaling features, and encoding categorical variables.

3. Data Analysis and Visualization:

Data analysis involves exploring and summarising data to extract meaningful insights. Visualization plays a key role in understanding patterns and relationships within the data. Techniques such as charts, graphs, and plots are used to present data visually, making it easier to comprehend and interpret.

4. Random forest:

Random Forest is an ensemble learning technique used in machine learning for both classification and regression tasks. It operates by constructing a multitude of decision trees during training and outputs the mode (for classification) or mean prediction (for regression) of the individual trees.

5. Randomisation:

Subset of Features: For each decision tree, a random subset of features is considered at each split. This introduces diversity among the trees. -Bootstrap Aggregating (Bagging): The training data for each tree is created by randomly sampling, with replacement, from the original training dataset. This creates different datasets for each tree.

6. Training Process:

Create Subsets: Random subsets of the training data are created using both random sampling of instances (with replacement) and random subsets of features. Build Trees: Decision trees are built using the created subsets. Aggregate Predictions: For classification tasks, the final prediction is determined by a majority vote (mode) among the predictions of individual trees. For regression tasks, it's the average of individual tree predictions. In summary, Random Forest is a powerful and versatile machine learning algorithm that leverages the strength of multiple decision trees to make accurate and robust predictions in various types of tasks.

ALGORITHM USED

Random forest:

Algorithm 1: FHR Classsification using Random Forest
Input: $S_i \subseteq \{(d_1, y_1), \ldots, (d_n, y_n)\}$
Step1: Perform row and column sampling
Step2: Decision tree DT_i *for each* S_i
Step3: Prediction $P_i \leftarrow$ *output of each* S_i
Output: $P \leftarrow$ *majority vote of* $P_1, \ldots, P_i$

Support vector machine

Algorithm 3: FHR Classsification using Support Vector Machine
Input: $\Delta = \{(d_1, y_1), \ldots, (d_n, y_n)\}$, *class labels*
$C = [c_1\|, c_2, c_3]$, *initial weight vector* $w = [w_i]^T$, *hyper-parameters, training set* T
Step1: Define the hyperplane $w^T * d + b \leftarrow 0$
Step3: Maximize the margin
$w * d + b \vee \frac{w}{\|w\| \vee \leftarrow \frac{1}{\|w\| \vee}}$
between the hyperplane and the plane median
Step4: Minimize $\|w\| \vee$ *to maximize the margin*
Output: w_{T+1}

Bagging:

Algorithm 4: FHR Classsification using Bagging

Input: $\Delta = \{(d_1, y_1), \ldots, (d_n, y_n)\}$, *class labels, base learning algorithm* L, *number of learning rounds j, training set* T
Step1: $\Delta_t \leftarrow$ *Bootstrap* Δ
Step3: $h_t \leftarrow L\Delta_t$
Step4: $j \leftarrow j + 1$
Step5: Repeat steps 1–3 till j covers the entire training set T
Output:

$$\hat{y}(d) = argmax \sum_{j=1}^{T} (y = h_t(d))$$

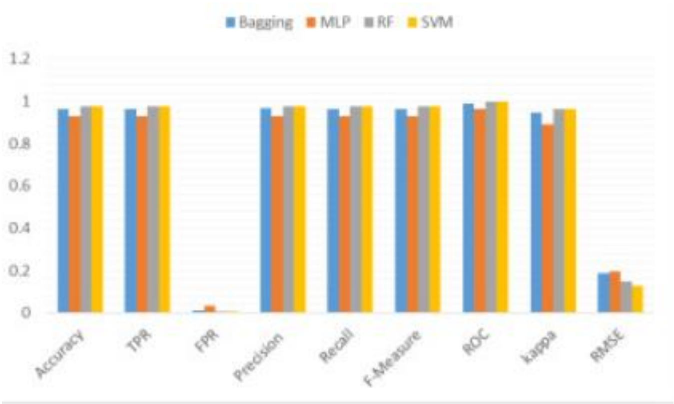
K fold cross validation:

Algorithm 5: Algorithm for the k-fold cross validation

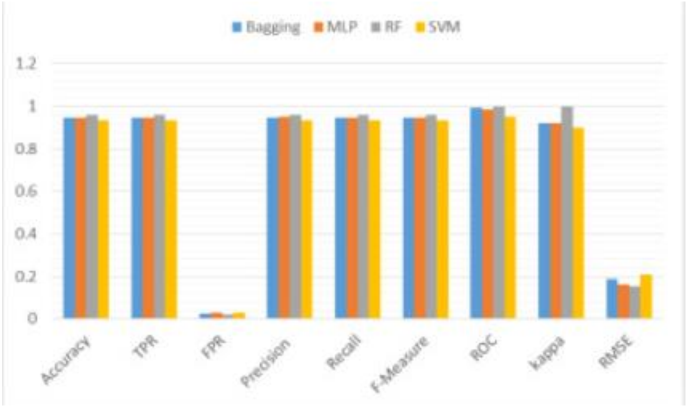
Input: $D = \{d_1, d_2, \ldots, d_n\}$, *class labels*
Step1: Initialize $i \leftarrow 1, k \leftarrow 5$
Step2: Split D *into k-folds*
Step3: In iterationi, the i^{th} *fold is used for testing and the rest are used for training*
Step4: $i \leftarrow i + 1$
Step5: Repeat steps 3–4 while $i \leq k$
Step6: Evaluate model on test score

RESULTS AND DISCUSSION

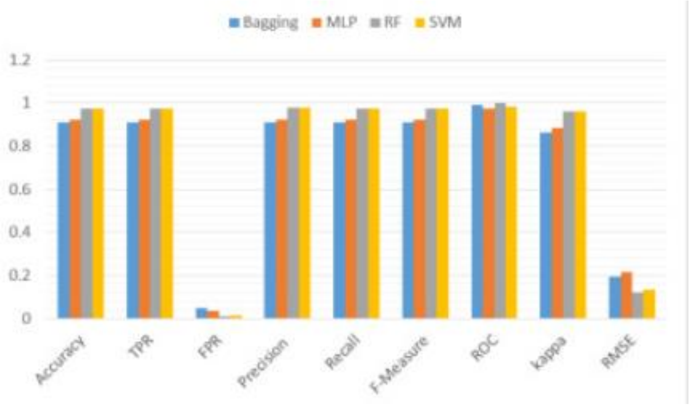
Accuracy of random forest



Accuracy of SVM



Accuracy of k fold



CONCLUSION:

In conclusion, the application of the Random Forest algorithm in predicting caesarean sections demonstrates its efficacy in addressing complex classification tasks within the realm of machine learning. Through the utilization of ensemble learning and decision tree- based models, Random Forest exhibits robustness in handling high-dimensional data and mitigating over fitting. The accuracy and efficiency of this approach in predicting caesarean sections pave the way for its potential utilization in various other fields, ranging from medical diagnostics to industrial quality control. Furthermore, by leveraging the insights derived from this study, future research endeavours can explore further refinements and adaptations of the Random Forest algorithm, ultimately contributing to advancements in both theoretical understanding and practical applications of machine learning techniques.

FUTURE SCOPE

In future work, further enhancement and exploration of the Random Forest algorithm for predicting caesarean sections could be undertaken.

This could involve several avenues of research and development:

1. Feature Engineering
2. Algorithm Optimisation
3. Ensemble Methods
4. Model Interpretability
5. Cross-Validation and Validation Strategies
6. Deployment and Integration
7. Domain-Specific Extensions

Advantages

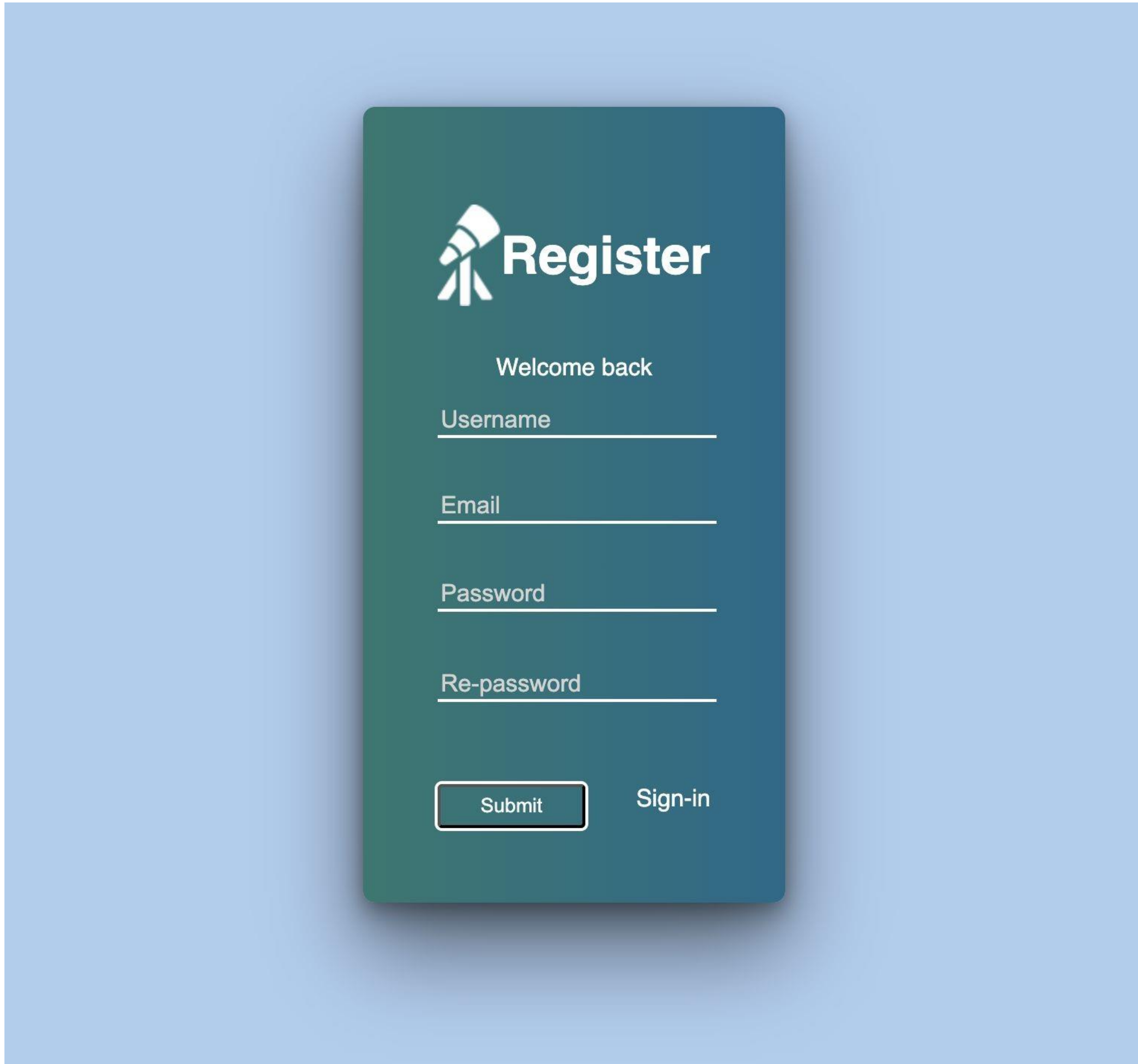
Random Forest is a powerful algorithm for predicting Cesarean section because it:

- Handles large datasets efficiently.
- Can handle both numerical and categorical data.
- Reduces overfitting by aggregating the predictions of multiple decision trees.
- Provides feature importance, helping identify the most influential factors in predicting Cesarean section.
- Is robust to outliers and missing values.
- Allows for easy interpretation and visualization of results.
- Typically requires minimal parameter tuning compared to other algorithm.

Progress of the project

- Implementation of the decision algorithm.
- Evaluation of the trained model's performance using metrics like accuracy, precision, recall.
- User can successfully view the output after the completion of the program.
- Utilization of techniques like cross-validation to assess model generalization and robustness.

Output Images


Register
Welcome back
Username
Email
Password
Re-password
Submit
Sign-in

PREDICTION OF SCISSORION SECTIONS

Home

baseline value (medical expert)

LBE

baseline value

LB

accelerations

AC

foetal movement

FM

uterine contractions

UC

abnormal short term variability

ASTV

References

1. T. Desyani, A. Saifudin, and Y. Yulianti, “Feature selection based on naive Bayes for caesarean section prediction,” IOP Conf. Ser., Mater. Sci. Eng., vol. 879, no. 1, Jul. 2020, Art. no. 012091.

2. S. A. Abbas, A. U. Rehman, F. Majeed, A. Majid, M. S. A. Malik, Z. H. Kazmi, and S. Zafar, “Performance analysis of classification algorithms on birth dataset,” IEEE Access, vol. 8, pp. 102146–102154, 2020.

3. N. I. Khan, T. Mahmud, M. N. Islam, and S. N. Mustafina, ‘‘Prediction of cesarean childbirth using ensemble machine learning methods,” in Proc. 22nd Int. Conf. Inf. Integr. Web-Based Appl. Services, Nov. 2020, pp. 331–339.

4. D. M. W. Powers, ‘‘Evaluation: From precision, recall and F Measure to ROC, informedness, markedness and correlation,’’ 2020, arXiv:2010.16061.

1. F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, ‘‘Data imbalance in classification Experimental evaluation,’’ Inf. Sci., vol. 513, pp. 429–441, Mar. 2020.

6. L. N. Mahdy, K. A. Ezzat, H. H. Elmousalami, H. A. Ella, and A. E. Hassanien, ‘‘Automatic X-ray COVID-19 lung image classification system based on multi-level thresholding and support vector machine,’’ medRxiv, 2020.

7. D. Chicco and G. Jurman, ‘‘The advantages of the Matthews correlation coefficient (MCC) over f1 score and accuracy in binary classification evaluation,’’ BMC Genomics, vol. 21, no. 1, pp. 1–13, Dec. 2020.

8. D. Dave, H. Naik, S. Singhal, and P. Patel, ‘‘Explainable AI meets healthcare: A study on heart disease dataset,’’ 2020, arXiv:2011.03195.

9. M. S. Bin Alam, M. J. A. Patwary, and M. Hassan, ‘‘Birth mode prediction using bagging ensemble classifier: A case study of Bangladesh,’’ in Proc. Int. Conf. Inf. Commun. Technol. Sustain. Develop. (ICICT4SD), Feb. 2021, pp. 95–99.

10. T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, ‘‘A survey on missing data in machine learning,’’ J. Big Data, vol. 8, no. 1, pp. 1–37, Oct. 2021.

11. A. Tharwat, “Classification assessment methods,” Appl. Comput. Informat., vol. 17, no. 1, pp. 168–192, Jan. 2021.

12. D. Chicco, M. J. Warrens, and G. Jurman, “The Matthews correlation coefficient (MCC) is more informative than Cohen’s Kappa and brier score in binary classification assessment,” IEEE Access, vol. 9, pp. 78368–78381, 2021.

13. V. Belle and I. Papantonis, “Principles and practice of explainable machine learning,” Frontiers Big Data, vol. 4, p. 39, Jul. 2021.

14. M. S. Islam, M. A. Awal, J. N. Laboni, F. T. Pinki, S. Karmokar, K. M. Mumenin, S. Al-Ahmadi, M. A. Rahman, M. S. Hossain, and S. Mirjalili, “HGSORF: Henry gas solubility optimization-based random forest for C-section prediction and XAI-based cause analysis,” Comput. Biol. Med., vol. 147, Aug. 2022, Art. no. 105671.

15. A. Javeed, L. Ali, A. M. Seid, A. Ali, D. Khan, and Y. Imrana, “A clinical decision support system (CDSS) for unbiased prediction of caesarean section based on features extraction and optimized classification,” Comput. Intell. Neurosci., vol. 2022, pp. 1–13, Jun. 2022.