How to Cite:
Mohamed, M. V., & Mayilvahanan, M. (2022). A descriptive analysis on bagging based hybrid ensemble classification technique. *International Journal of Health Sciences*, *6*(S2), 12556–12566. https://doi.org/10.53730/ijhs.v6nS2.8328

A descriptive analysis on bagging based hybrid ensemble classification technique

M. Varusai Mohamed
Assistant Professor, Department of Computer Science, Einstein College of Arts and Science, Seethaparpanallur

Dr. Mayilvahanan
Professor, School of Computing, VISTAS, Vels University, Chennai

***Abstract*---**Machine learning involves data mining, which solves many problems in data science. An application in machine learning predicts the outcome based on available data. There are many predictive strategies available. The most important method is dividing the most powerful predictions. Some of them predict the results satisfactorily and some are accurately measured. This investigation has carried out a process called the bagging based hybrid ensemble process, which collects the accuracy of weak algorithms by combining multiple separators for development. This study helps us to see the integration process that improves the accuracy of predictor birth. This is not only the study of weak separation algorithms, but also the use of algorithms using medical data, which predicts prematurely. This study proves to be an effective method of bagging based classification to improve the prediction accuracy of 97.4%.

***Keywords*---**machine learning, bagging, hybrid classification, prediction, ensemble classification.

Introduction

The full gestation period is considered at 40 weeks of gestation. At 20 to 37 weeks of gestation, the onset of menopause should be delayed due to uterine contractions that result in the opening and closing of the cervix [1]. Identifying pregnant women at high risk of premature birth requires high accuracy to determine who will benefit from certain clinical interventions. Further improvement in outcomes to extend the gestation period from 28 to 36 weeks determines survival and reduces the cost of intensive care for newborns. 70% of premature births are more likely to die than premature babies [2]. Most premature babies are at risk for life-threatening complications such as cerebral palsy and other neurological disorders, hydrocephalus, blindness, deafness,

respiratory infections, and behavioural problems [3]. Despite advances in birth control, preterm birth rates remain almost 8-12% in the USA over the past two decades. Although the total number of premature births has not changed, there has been significant improvement in the survival of premature infants, but at a cost that is very costly in our national health budget [4].

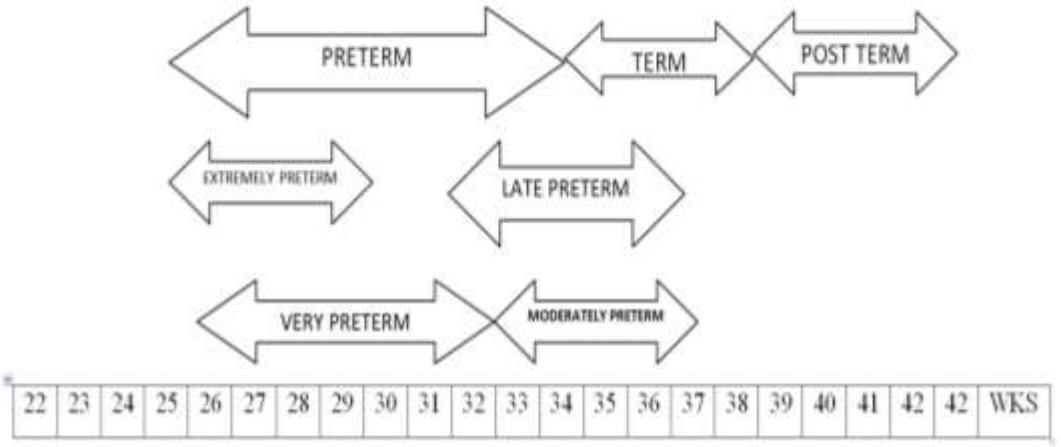

Figure-1 Information of preterm period

The addition of maternal biomarkers (e.g., cervical length, fetal fibronectin, and serum analytees) improves prognosis, but gradually [7, 8] a few methods can be used to predict fetal weight in clinical activities, including miscarriage, i - parturient symphysio. -Fundal height and abdominal girth measurements, as well as obstetric ultrasound. Among them, the ultrasound-based measurement method is the most reliable and effective, and is widely used by obstetricians in China. Its goal is to use a class of well-designed retreat models with multiple levels of children's boundaries [5]. However, there are many limitations to such a method. First, these types of regressions were proposed by different doctors, and they do not usually apply to all people in the world.

As a result, the direct application of such a category of models to the Chinese people may result in inaccuracies, especially in the case of overweight or low birth weight. Second, there are strict requirements for sonographers and certain standards for ultrasound imaging equipment. Factors such as a deformed fetal head, the presence of oligohydramnios and abdominal fat, and low image quality may affect the final measure. One limitation is that access to obstetric ultrasound is still poor in some rural areas with limited resources and this has had a significant impact on child weight measurement. In addition to the standard methods introduced, machine learning methods can be used in this field [6]. Historical data on prenatal testing can be analyzed and relationships between fiction organizations can be assessed through their training, integration, planning, and learning ability. .The use of machine learning techniques may reveal test algorithms that are more sensitive than conventional[7].

Diverse machine learning algorithms are compared using the opposite validation, a method used to test how well a guessing model works without sample matching the model. The novel sample was separated into a training set to fit the model and a test set to assess the quality of the fit. In this way, the general potential of the results is assessed while minimizing the risk of over-modelling of available data [8].

The organisation of this paper is as follows: section 1 elaborates about basics of preterm delivery, role of machine learning methods in preterm prediction, In section 2 existing classification algorithms for preterm labour prediction are shown. In section 3 proposed methodologies is elaborated. Section 4 shows the experimental analysis with graphs. Finally the paper ends with section 5 conclusion and future work.

Literature Survey

Saranya et al., (2019) The proposed system adopts a machine learning approach that includes addition, subtraction, etc., to identify the factors involved in causing a premature baby for the mother to predict. Key factors include diabetes, birth control number, drugs, and cervical dysfunction. Other accurate features include body position, Pityriasis Rosea, Systolic & Diastolic Blood Pressure, Stroke Volume Index, etc., by analyzing the above features, using the K- Nearest Neighbor algorithm and the C 4.5 system can predict whether the mother will have a baby prematurely or not. If there is a high probability of having a premature birth, the system automatically raises the mother with a risk proposal based on an important factor [9]

Alisha Kamat, et al., (2015) used two classification algorithms such as Naive Bayes and ID3 to determine the method of delivery based on a few parameters available in obstetric ultra sonography, and blood and urine tests for pregnant women. The result demonstrates high accuracy and memory thus ensuring the accuracy of these partition algorithms for effective prediction. This program can be helpful in recommending women to take a second view in situations where doctor's predictions are very different [10].

Ilia Vovsha et al. (2016) compared three methods for obtaining predictable models: a vector machine (SVM) method with straight and non-linear characters, a retrospective retrieval of different model options and a rule-based model determined by medical experts to predict premature birth. . This method of analysis highlights previous processing techniques used to manage natural forces, sounds and spaces in the data and defines the techniques used to manage a distributed class distribution. Strong tests show significant improvements in predictability of premature birth compared with previous activity [11].

Srimani et al., (2013) proposed mining data integration (EDMM) methods, learning algorithms with a combination of many basic models. Tests are performed on five sets of medical data and the results prove that there is a significant improvement in the function of the primary detectors, and this will certainly facilitate effective medical diagnosis, which will have an impact on patient health index.

Further, it has been concluded that only selected class dividers should be used in each data set, and in some cases mentioned, merged class dividers do not need to be suggested. Proper classification is recommended in order to achieve complete accuracy in relation to a specific medical data set [12]. Peng ren etal., (2015) a new method of early risk assessment analysis using EMG recordings that first used the EMD and IMF functions which are Empirical Mode Decomposition and Intrinsic Mode respectively. All these functions are compared with previous methods to achieve maximum level of AUC as 0.986 [13].

Nynke R. van denet al., (2014) [14] used multi-item retrieval to obtain independent traits associated with preterm birth, premature and premature birth. History of female pregnancy and early diagnosis of maternal weight loss, malaria and anemia as risk factors for premature birth; HIV status does not put them at risk of premature birth. Traditional, simple, repetitive, logical retrospective logos and Log-Gaussian models (with continuous variation) do not work better than standard log models (with constant variability) because they are better suited to the data.

Guillermo Marshall et al. (2015) analyzed the associations between hospital mortality and preterm infants and hospitalization features using a multidisciplinary retrospective model and a multidisciplinary retrospective model. Sample test and reverse confirmation techniques are used to validate the statistical model on hospital mortality. The new level of risk is compared to the two existing points using the area below the curve of the receiver's performance factor [15].

Proposed Approach

Classification is one of the most important functions in data mining. There are many types of data separation algorithm. Separation algorithms also play an important role in analyzing and predicting clinical data. The proposed system uses flexible separation methods to predict heart disease more accurately and effectively.

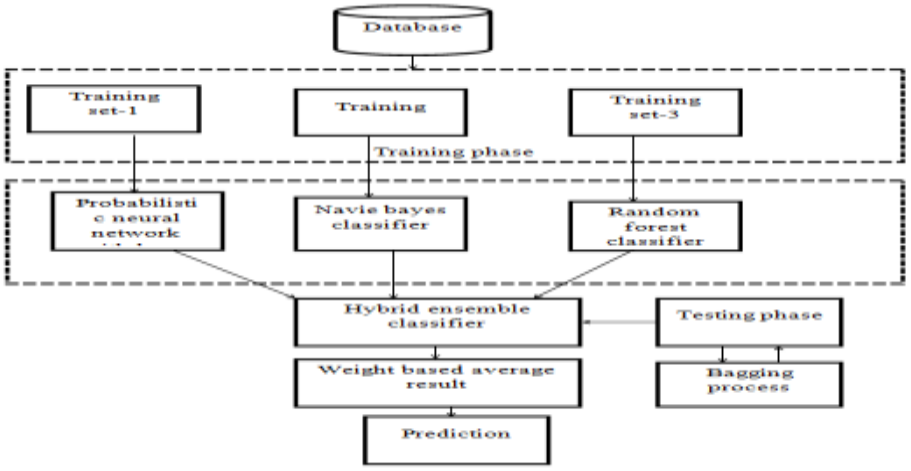

Figure- 2: Proposed architecture for classification

Probabilistic neural network with lasso regression

PNN provides a common solution to pattern separation problems by following a mathematical method, called Bayesian separators.
Euclidean (Edj) process for data training and testing. The typical Euclidean activity can be defined as:

ED = $\sqrt{\sum W(ij) - f(i)\text{^}2}$ (1)

The effects of the first hidden layer neuron are,

X1(n) = $\frac{1}{(1+\exp(w1(n)*f(n)+b1(n)}$ (2)

The neuron effects of the second hidden layer are,

X2(n) = $\frac{1}{(1+\exp(w2(n)*f(n)+b2(n)}$ (3)

Outputs of the network are,

Y(n) = $\frac{1}{(1+\exp(wh(n)*f(n)+bh(n)}$ (4)

The back row represents the negative chance associated with a given input combination. The scatter structure is defined by the line number,

y=β0 + β1 x1+ β2x2+…..+ βixn for i = 1…n (5)

The deviation between the drop line and a single data point is a variance that our model cannot explain. This inexplicable diversity is also called residual. The small square method is used to reduce the rest.

The lasso regression's variance α^2 is estimated as $S^2 = \frac{\sum e\ (i)\text{^}2}{n-p-1}$ (6)

Where,
p is the number of independent variables and
n the sample size.

After obtaining the lasso regression equation, the suitability and usefulness of the mathematical is assessed. The main criterion for linear fitness is R^2.
R^2 = total difference / defined difference.

Hybrid ensemble classifier

The data stream contains a continuous sequence of events: {a1, a2, ..., a (n)}, where a1 is the first event in the stream, and a (n) is the latest, most recent example. they arrived. One event is observed at a time, not at intermittent intervals. In each example ai is a n-dimensional element vector containing a number of attributes, Bi = {B1, B2, ..., Bn}, with class label, Ai - 2 {A1, A2, ..., An }. Each attribute has an attribute value, Ci = {Ci1, Ci2, - ..., CIP}. An example of training ai is labeled the value of class Bi, so pear (ai, Bi) is called a training pattern with label. We refer to cases {a1, a2, ..., a (n) as historical data and a (n) +1 as an experimental (or targeted) example. Initially, the equivalent weight, 1 / N was given for each example, y (i) in the actual training data set, D.

Error rate calculation

DTj error rate is calculated by the sum of the weight of the undivided events. There, error (xi) is an error in the case of xi. If, for example, xi is not well separated, it means that error (xi) is the same. If not, error (xi) is zero, if for example, xi is correctly separated.
err(DTj) = wi × err(xi)

The weights of the properly set conditions were reviewed after issuing the rules. If, for example, xi in jth iteration is organized correctly, its weight is repeated by err (DTj) 1 − err (DTj)). Then the weights of all conditions (including the poorly classified conditions) were made normal. Therefore, the weight of the undivided conditions increases and the weights of the properly positioned positions have been reduced. Finally, a subset of Dm data is separated and created from the Dj in poorly separated conditions.

Dataset

Personal data

Status :married or unmarried
Level of education: primary or post primary
Maternal occupation: unemployed, self employed
Age : <20, 20-34, >35

Medical Risk factors

Underweight : yes/no
Diabetics : yes/no
Urinary infection : yes/no
Placenta prevail : yes/no
Abnormality in fetus : yes/no
High blood pressure : yes/no
Blood clotting problems : yes/no
Shortime period between pregnancies: yes/no
Obesity : yes/no
Bleeding : yes/no
Polyhydramnios : yes/no

Other risk factors

Smoking during pregnancy : yes/no
Alcohol during pregnancy : yes/no
Limiting health care : yes/no
Illegal drugs usage : yes/no
Domestic violence : yes/no
Lack of social support : yes no
Stress : yes/no
Long working hours with long period of standing : yes/no

Exposure to certain environmental pollutants : yes/no

Bagging process

Random loading selects specific patterns in the installed test set. The newly tested test set will have the same number of patterns as the actual test set with fewer omissions and duplicates. The new test set is known as Bootstrap replicate. In bag packaging, bootstrap samples are retrieved from the data and a separator is trained for each sample. Voting from each category divider is combined, and the result of the division is chosen based on the majority of the vote or rating. Studies show that wrapping can be used to maximize the effectiveness of a weak separator.

Bagging based Hybrid ensemble classification technique

Start weights and bias in N, where N is Network
while the situation is real {
in each tuple of training X in D {
per unit of input layer j {
Allow D = {d1, d2, d3,... dn} to be a given database
ai = {a1, a2, a3... ..an}
bi = {b1, b2, b3... .bn}
ci = {c1, c2, c3 ..... cn}
H = {}, a set of class dividers (here, 3 dividers}
C = {c1, c2, c3}, set of dividers
Wi = weight of features
C (ij) = f (j)
f (j) = (f1, f2, f3 ..... f15) = W (i)
fori = 1,2,3,4... ..15 and j = 1,2,3... .30
Edj = Normal Euclidean
ED = √ (∑▒ ⟦W (ij) -f (i) ^ 2⟧)
X (n) = (X1 (n), X2 (n), X3 (n)... .. Xn (n))
Y (n) = (Y1 (n), Y2 (n), X3 (n)... .. Yn (n))
y = β0 + β1 x1 + β2x2 + ... + βixn
of α2 = S ^ 2 = (∑▒ ⟦e (i) ^ 2⟧) / (n-p-1)
L = α2
because i = 1 to make L
F (i) = {Sample Z-bag with backup data}
N (i) = Modeled trained using C (i) in S (i)
H = H * F (i)
Next i
because = 1 to L
K (i) = Y (n) and X (n) are divided by H (i)
Next i
Result = limit (K (i): i = 1,2,..., n)

Performance Analysis

In order to compare the results of our proposed wallet method is preferred, which is why it is used separately in all three categories in the database. Comparative

analysis of differential classification algorithms for the above-mentioned database was performed. Some algorithms show good accuracy and some algorithms do poorly. To improve the performance of the weak sections, ensemble algorithms are used.

Sensitivity = $\frac{TP}{TP+FN} * 100$
Specificity = $\frac{TN}{TN+FP} * 100$
CA = (t/n).100

Where,
t is the total of sample cases correctly classified
n is the total of sample cases

Method	Sensitivity		Specificity		Classification Accuracy	
	Hybrid Method	Basic Method	Hybrid Method	Basic Method	Hybrid Method	Basic Method
Probabilistic neural network with lasso regression	57%	42.5%	59%	13.5%	45%	41%
Navie bayes classifier	69%	51.4%	51%	42%	49%	33%
Random forest classifier	71%	66.4%	38%	31%	81%	78%

Figure -3 Comparison of hybrid and basic method in terms of various parameters

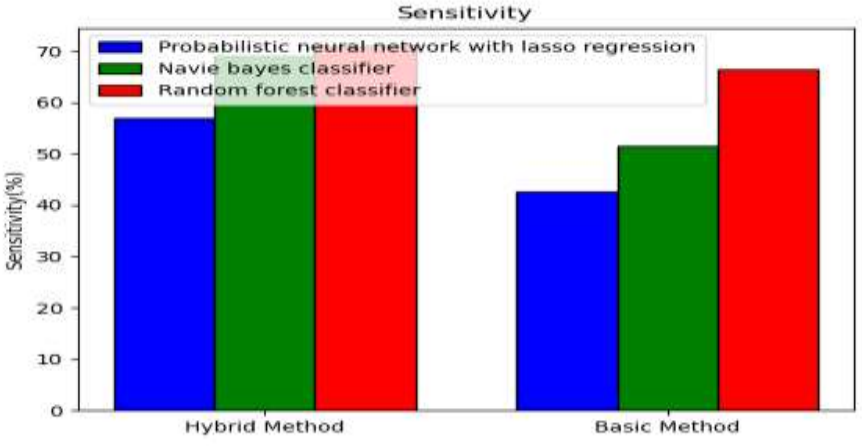

Figure-4 analysis of sensitivity

Figure 4 shows a percentage sensitivity comparison between three dividers with or without input method. It noted that through the fundraising process individual

designers received a sensitivity of PNN 57%, navie bayes 69% and informal jungle 71%.

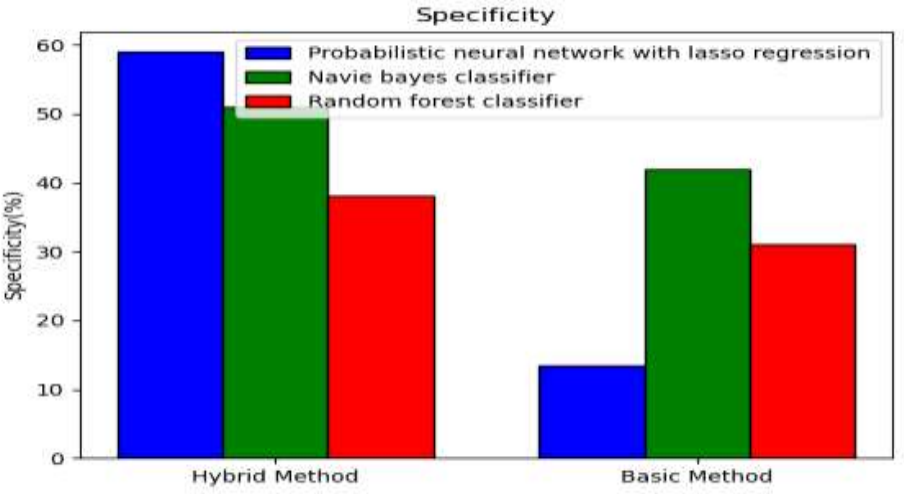

Figure-5 analysis of sensitivity

Figure 5 shows the percentage specificity comparison between the three dividers having a method of assembling the bags on and off. Note that through the bagging method individual dividers receive a certain PNN rate of 59%, navie bayes 61% and the random forest 38%.

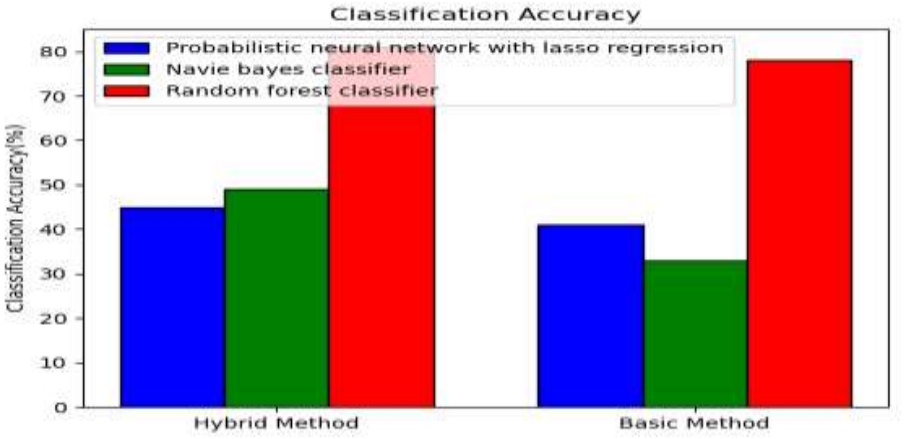

Figure- 6 analysis of classification accuracy

Figure 6 shows a comparative accuracy of the percentage categories between the three dividers with the method of assembling the bags on and off. It noted that through the bagging method individual dividers obtained a PNN classification rate of 45%, navie bayes of 49% and a random forest of 81%.

Method	Overall Accuracy
Probabilistic neural network with lasso regression	83%
Navie bayes classifier	59.4%
Random forest classifier	78%
Bagging based hybrid ensemble classification technique	97.4%

Figure-7 Comparison of overall accuracy

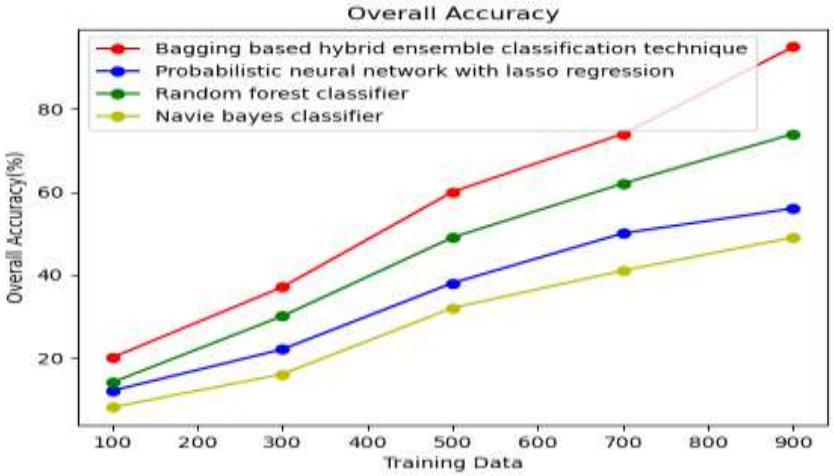

Figure-8: comparison of proposed ensemble classification method

Although the experiments achieved acceptable results by constructing a mixed integration model, another important challenge was comparing current research with previous methods. Figure 6 shows that absolute accuracy is measured by three dividers. It is noteworthy that the process of splitting the proposed financial combination reaches an accuracy of 97.4% and therefore has the potential to predict premature births successfully.

Conclusion

This paper analyzes the accuracy of premature birth predictions using a set of class dividers. Database from the UCI machine learning archive was used for training and testing purposes. The ensemble algorithm bagging is selected with classifications used for testing. It was found that the bags were used and the accuracy was found to be improved by a high value of 97.4% .The test results show that the estimation of standard systems with the appropriate classification system is much higher (58% - 97%) than conventional manual methods. (15.3% - 29%). The future task is to focus on additional attributes in the form of feature selection.

References

1. Dangare CS, Apte SS. Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques. *Int J Comput Appl*. 2012;47(10):44–48.
2. M. Lucovnik, W.L. Maner, L.R. Chambliss, R. Blumrick, J. Balducci, et al., "Noninvasive uterine electromyography for prediction of preterm delivery," American journal of obstetrics and gynecology, vol. 204, no. 3, pp.228.e1–10, 2011
3. G. Fele-Žorž, G. Kavšek, Ž. Novak-Antolič, and F. Jager, "A comparison of various linear and non-linear signal processing techniques to separate uterine EMG records of term and pre-term delivery groups," Medical & Biological Engineering & Computing, vol. 46, no. 9, pp. 911-922, 2008
4. N.E Huang, Z. Shen, S.R. Long, et al. "The empirical mode decomposition and Hilbert spectrum for nonlinear and nonstationary time series analysis," IEEE Signal Processing Letters, vol.11, no.2, pp. 112-114,2004
5. Savitz DA, Terry JW Jr., Dole N, Thorp JM Jr., Siega-Riz AM, Herring AH. Comparison of pregnancy dating by last menstrual period, ultrasound scanning, and their combination. American journal of obstetrics and gynecology. 2012; 187(6):1660–6
6. Miller, S. L., and Huppi, P. S. 2016. The consequences of fetal growth restriction on brain structure and neurodevelopmental outcome. Journal of Physiology 594(4):807–823
7. Rao CR, Bhat P, KE V, Kamath V, Kamath A, Nayak D, Shenoy RP, Bhat SK. Assessment of risk factors and predictors for spontaneous pre-term birth in a South Indian antenatal cohort. Clin Epidemiol Glob Heal. 2018;6:10–6.
8. Wang S, Yao X. Relationships between diversity of classification ensembles and single-class performance measures. IEEE Trans Knowl Data Eng. 2013;25:206–19.
9. Saranya N , Pavithra R," Prediction of Premature Baby using Machine Learning Algorithm" IOSR Journal of Nursing and Health Science, Volume 8, Issue 2 Ser. VIII. (Mar. - Apr .2019), PP 85-90
10. Kamat, Alisha, Veenal Oswal, and Manalee Datar. "Implementation of classification algorithms to predict mode of delivery." *International Journal of Computer Science and Information Technologies* 6.5 (2015): 4531-4.
11. Vovsha, Ilia, et al. "Using kernel methods and model selection for prediction of preterm birth." *arXiv preprint arXiv:1607.07959* (2016).
12. Srimani, P. K., and Manjula Sanjay Koti. "Medical diagnosis using ensemble classifiers-a novel machine-learning approach." *Journal of Advanced Computing* 1 (2013): 9-27.
13. Ren, Peng, et al. "Improved prediction of preterm delivery using empirical mode decomposition analysis of uterine electromyography signals." *PloS one* 10.7 (2015).
14. Nynke R. van den Broek, Rachel Jean-Bapsite, James P.Nelison, "Factors Associated with preterm, early preterm and late preterm birth in Malawi," PLOS ONE, Volume 9, Issue 3, March 2014, pp.1-8
15. Guillermo Marshall, et al., "A New Score for Predicting Neonatal Very Low Birth Weight Mortality Risk in the NEOCOSUR South American Network," Journal of Perinatology, 25, 2015, pp.577-582.