Hindawi
Journal of Healthcare Engineering
Volume 2021, Article ID 6665573, 11 pages
https://doi.org/10.1155/2021/6665573

Research Article

A Machine Learning-Based Prediction Model for Preterm Birth in Rural India

Rakesh Raja, Indrajit Mukherjee, and Bikash Kanti Sarkar

Department of Computer Science & Engineering, Birla Institute of Technology, Mesra, Ranchi, India

Correspondence should be addressed to Rakesh Raja; rajarakeshchauhan@gmail.com

Received 20 November 2020; Revised 16 March 2021; Accepted 17 March 2021; Published 15 June 2021

Academic Editor: Giovanni Improta

Preterm birth (PTB) in a pregnant woman is the most serious issue in the field of Gynaecology and Obstetrics, especially in rural India. In recent years, various clinical prediction models for PTB have been developed to improve the accuracy of learning models. However, to the best of the authors' knowledge, most of them suffer from selecting the most accurate features from the medical dataset in linear time. The present paper attempts to design a machine learning model named as risk prediction conceptual model (RPCM) for the prediction of PTB. In this paper, a feature selection approach is proposed based on the notion of entropy. The novel approach is used to find the best maternal features (responsible for PTB) from the obstetrical dataset and aims to predict the classifier's accuracy at the highest level. The paper first deals with the review of PTB cases (which is neglected in many developing countries including India). Next, we collect obstetrical data from the Community Health Centre of rural areas (Kamdara, Jharkhand). The suggested approach is then applied on collected data to identify the excellent maternal features (text-based symptoms) present in pregnant women in order to classify all birth cases into term birth and PTB. The machine learning part of the model is implemented using three different classifiers, namely, decision tree (DT), logistic regression (LR), and support vector machine (SVM) for PTB prediction. The performance of the classifiers is measured in terms of accuracy, specificity, and sensitivity. Finally, the SVM classifier generates an accuracy of **90.9%**, which is higher than other learning classifiers used in this study.

1. Introduction

Preterm birth (PTB) is a serious public health problem that adversely affects both families and the society [1]. It is a leading cause of neonatal mortality and morbidity across the world and also the second major cause of child deaths under the age of five years [2]. Over the past two decades, PTB has been a significant research study in healthcare domain. Pregnancy and childbirth unlocked the door for medical experts and researchers to explore various effective strategies to reduce preterm birth in women having pregnancy-related complications. These strategies include healthcare services given to all pregnant women to control PTB and any medical interventions aimed to enhance the knowledge of women on early indications of pregnancy complications [3, 4]. The maternal history of a pregnant woman is a key part of the neonatal studies for providing certain clinical treatments to newborn babies regarding their health, disease, care, and outcomes. Newborn babies are very special. They do not have any previous medical background, and their early neonatal path is directly connected to the maternal history of their mothers [5–7]. The healthcare services also incorporate the arrangements of essential social and economic support for women before, during, and after pregnancy including educational, medical, and other training programs that facilitate healthy motherhood.

In general, treatments of diseases (including PTB) are made by the physicians based upon their knowledge (experience). However, on the one hand, manual diagnosis may not be often right as physician's experience varies from expert to expert. On the other hand, manual treatment is a time-consuming job. Further, shortage of medical experts is increasing everyday with population explosion and developing countries like in India, large number of women belong to lower or middle income families. They do not get proper healthcare facilities or awareness regarding health education to know about any complication that arises during pregnancy, especially in rural area. Further, people often are

afraid of doctors' prescription since doctors in most cases misguide the patients suggesting unnecessary tests (like double marker test, fetal echocardiography, urine test, and FT4 test which are used to determine any pregnancy complications) which are very expensive. Also, doctor's appointment fees are mostly on higher side. Besides, doctors could sometime diagnose the cases wrongly. After all, preterm delivery is the most critical issue in Gynaecology and Obstetrics and a major health concern for every pregnant woman. It may require several ultrasound sonography (USG) tests in addition to doctor's appointment fee for diagnosing high-risk patients, and these altogether may amount huge expense that may be beyond the income limit of many families. So, designing the computerized system (i.e., e-healthcare system) for birth prediction from past diagnosis data is the essential solution for quick and accurate decision to be taken for any adverse pregnancy outcome in order to save lives and cost.

Notably, a pioneering renovation is taking place in the Obstetrical community due to the advancement in technology and digitization of medical records. Data analytics is one of the most promising tool for research and development in the area of medicine [8–15]. Nowadays, machine learning techniques (e.g., neural networks, support vector machine, logistic regression, and Decision Trees) are playing important role in designing the disease predictive model to address the growing needs of human experts in the medical world [16–20]. However, medical datasets are highly imbalanced, conflicting in nature, and uncertain. So, designing the effective intelligent model for medical datasets is a challenging task. PTB dataset is one such clinical dataset. Numerous predictive models based on standard intelligent methods have been introduced by the researchers for prediction of PTB [21]. However, they usually suffer from several drawbacks like lack of understandability and inefficiency in making quick and correct decision. Further, early detection and diagnosis play important role in controlling such complications. Symptoms (text) based machine intelligent models may play vital role in early detection of such cases. The delay in receiving the clinical judgement for preterm delivery increases the risk of pregnancy complications which in turn increases the risk of prenatal mortality. Due to its direct association with prenatal mortality, neonatal health is also very important in the obstetrical community [7]. According to the UNICEF study released in 2015, 35% of neonatal death is due to PTB. The rate of PTB in rural areas of most developing countries is increasing due to lack of health facilities and insufficient number of healthcare workers.

In light of these considerations, the present study aims to design a novel conceptual model (by employing machine learning techniques) and its implementation for detection of PTB in pregnant women. In fact, the system can be used as a decision support system to assist the medical staff and healthcare workers for predicting premature delivery. More specifically, the present study focuses on novel feature selection (entropy-notion) approach to identify the most important maternal features (text-based symptoms) responsible for preterm delivery and aims to predict the classification accuracy.

The remaining sections of the paper are organized as follows. Section 2 describes the basic concept of PTB and feature selection. Section 3 elaborates the related work that has been carried out to predict PTB. Section 4 describes the methodology of this research. The experimental design and results are presented in Section 5. Finally, Section 6 deals with conclusion and future scopes.

2. Background of the Present Research

2.1. Preterm Birth (PTB): A Comprehensive Overview. Preterm or premature birth is defined as birth, for any reason, occurring before 37 completed weeks (or less than 259 days) of pregnancy. Every year, about fifteen million babies are born prematurely (before 37 completed weeks of gestation), and this is nearly equal to one-tenth of all babies around the world [22]. According to the WHO reports studied in 2005, 12.9 million births or 9.6% of all births across the world occurred prematurely [23]. The rate of preterm birth, however, significantly varies across the world. Preterm birth reflects the most prominent reason for neonatal morbidity and mortality [24].

2.1.1. Categorization of PTB. PTB can be classified into different categories based on gestational age at birth. The gestational age is defined as the time from the first day of the last menstrual period (LMP) of a woman to birth [21]. The four categories of PTB are as follows:

(i) *Extreme PTB (under 28 Weeks).* It is the birth that takes place before 28 weeks of pregnancy

(ii) *Very PTB (28 to 32 Weeks).* It is the birth that takes place between 28 and 32 weeks of pregnancy

(iii) *Moderate PTB (32 to 34 Weeks).* It is the birth that takes place between 32 and 34 weeks of pregnancy

(iv) *Late PTB (34 to 37 Weeks).* It is the birth that takes place between 34 and 37 weeks of pregnancy

2.1.2. Medical Terminologies. For the purpose of clarity of the present study, the used terminologies are illustrated in Table 1.

2.1.3. Health Impact of PTB. PTB is the main risk factor for newborn mortality and morbidity. It is a leading cause of neonatal mortality and morbidity across the world and also the second major cause of child deaths under the age of five years [25]. It arises between 5 and 10% of all deliveries and involves 70% of neonatal mortality and up to 75% of neonatal morbidity [26]. Premature infants are more likely to suffer than normal birth and are at higher risk of brain paralysis, sensory impairment, respiratory failure, and so on. More than $13 billon of premature cost for maternity service is anticipated only in the USA [27, 28]. Most survivors of PTB face serious problems, often a lifetime of disability, including learning disabilities, visual, and hearing problems.

Table 1: Definitions used in the present study.

Terminology	Description
Antenatal care	Antenatal care (ANC) refers to the fundamental, clinical, and nursing care suggested for ladies during pregnancy
Neonate	A neonate or a newborn infant is a child under 28 days of age
Neonatal death	A death during the first 28 days of life (0–27 days) is termed as a neonatal death
Live birth	A birth at which a child is born alive is termed as live birth
Term birth	A birth at the end of a normal duration of pregnancy between 37 and 40 weeks of gestation is termed as term birth
Maternal death	A maternal death is the death of a woman while pregnant or within 42 days of termination of pregnancy
Stillbirth	Stillbirth is the delivery, after the 20th week of pregnancy, of a baby who has died
Abortion	Termination of a pregnancy either medically or induced
Miscarriage	Natural loss of pregnancy during first trimester
Gestational age	Gestational age (GA) refers to the time from the first day of a woman's last menstrual period to birth

In fact, babies born premature have more health problems compared with babies born at term birth. Term birth refers to babies that are born at 37 to 40 weeks of gestation. Furthermore, babies born at preterm are reported to be at an elevated risk of long-term health problems [29]. Unfortunately, after many years of research in obstetrics, yet the rate of PTB has not decreased [30]. Birth weight is generally associated with PTB and results in its own categorization. Usually, birth weight is simpler to measure precisely and is a first estimation of gestational age. Obviously, the most challenging issue in Gynaecology and Obstetrics is how to control the preterm delivery in pregnant women.

2.2. Feature Selection (FS). The term feature selection in the machine learning, also known as feature subset selection, refers to the process of selecting a subset of excellent features during construction of the predictive model. The presence of redundant and irrelevant features in any datasets (especially in medical datasets) can reduce the accuracy of the model's prediction and also have the negative impact on the performance of the model. The main goal of any feature selection method is to select the best subset of features by removing redundant and irrelevant features from the datasets in order to reduce the training time and enhance the classifier's predictive performance. In fact, feature selection is typically used as a preprocessing step in data mining. There are three standard approaches of the feature selection algorithm, namely, filter method, wrapper method, and embedded method. For more details about feature selection, one may refer to [31–33].

(i) *Filter Method*. The filter method measures the relevance of features based on the nature of data. The selection of features is independent of the classifiers used. The filter method is much faster compared with the wrapper method and provides an average accuracy for all the classifiers used. Some of the examples of filter methods are information gain, chi-square test, variance threshold, and so on.

(ii) *Wrapper Method*. The wrapper method finds the best subset of features based on a specific machine learning algorithm that we are trying to fit on a given dataset. The evaluation criteria are simply the predictive power of the particular classifier. The wrapper method has higher performance accuracy compared with the filter method but requires more computational time to find best features for a dataset with high-dimensional features. Some of the examples of wrapper methods are forward selection, backward elimination, genetic algorithms, and so on.

(iii) *Embedded Method*. The embedded method incorporates the advantages of both filter and wrapper methods. In this approach, feature selection is done during the process of model training and is usually unique to particular learning classifiers. This approach basically determines the importance of feature, i.e., which features to accept and which to reject, while making a prediction. The most typical embedded technique is the decision tree algorithm. This method typically falls somewhere between the filter method and wrapper method in terms of time complexity. Some of the examples of embedded methods are lasso regression, ridge regression, elastic net, and so on.

3. Related Works

This section focuses mainly on the existing methodologies related to prediction of PTB using machine learning, statistical analysis, and data mining techniques. Some of them are discussed in this section. The study of Mercer et al. [34] was designed to develop a risk-score-based model for predicting PTB. The model can be trained using a multivariate logistic regression technique to explore various risk factors using clinical data available between 23 and 24 weeks' gestation. Goodwin et al. employed the machine learning model to generate 520 predictive rules for PTB with the application of data mining techniques [35]. The study in [36] discussed the deep learning models for predicting preterm delivery using existing electronic medical records (EMRs) of mothers available in healthcare centres.

Weber et al. [37] performed a cohort study to predict spontaneous preterm. The prediction of PTB was performed using numerous classifiers, namely, K-nearest neighbours, lasso regression, and random forests. This study has taken

into the consideration of demographic, race-ethnicity, and maternal characteristics. Mailath-Pokorny et al. [38] explored the predictive features for preterm delivery that occurs within 2 days after admission and before 224 days of gestation using the multivariate logistic regression model. The predictive features considered are age of the mother, gestational age during admission, maternal history, vaginal bleeding, cervical length, preterm history, and preterm premature rupture of membranes (PPROM) in their study. Son and Miller presented a prediction model for PTB using cervical length measurement in women with a singleton gestation. To accomplish better predictive performance, they attempted to determine the best cut points of cervical length [39].

Elaveyini et al. [40] explored the major risk factors of preterm birth using artificial neural networks. PTB prediction was based on the feed-forward backpropagation algorithm. Over the past decades, majority of research studies have been done to enhance the accuracy of prediction of PTB [41]. Researchers are continually making their best efforts to analyse and explore the principal risk factors for preterm delivery [42–44]. The present article focuses on the machine learning approaches for prediction of birth cases in rural community.

3.1. Shortcomings in the Existing Clinical Models. In recent years, using feature selection approach, a significant number of clinical prediction model have been developed to improve the accuracy of learning models. However, to the best of authors' knowledge, most of them suffer from selecting the most accurate features from the medical dataset in linear time. Hence, there is a scope for improving the performance of machine learning classifiers and reducing learning time.

3.2. Novel Contribution. A novel feature selection approach based on the notion of entropy is introduced in this study to address the identified issues of the existing models. The key role of the novel approach is to find the subset of optimal features from the medical dataset in order to improve the prediction's accuracy and ultimately reduce the machine learning time.

4. Research Methodology

4.1. Objective. The finding of this research study can be utilized to fulfill the three following main objectives:

(i) A machine learning-based risk prediction conceptual model (RPCM) for PTB can be introduced with the help of novel feature selection approach using entropy-notion to predict the birth cases (TB and PTB) from the obstetrical records.

(ii) The suggested approach is used to identify the excellent (text-based symptoms) features responsible for PTB. Furthermore, medical experts' (physicians and obstetricians) opinions are also considered through review of medical records of patients and survey analysis. The model can be extended to select the regions for pregnancy consultation.

(iii) The predictive model can be beneficial for rural India to identify the important maternal features in order to predict the possibility of PTB in the gestation of women. This information can support rural medical staff for taking effective decisions for adverse pregnancy outcome—that aim to reduce the diagnosis cost.

4.2. The Proposed Feature Selection Approach Based on the Notion of Entropy. According to the study in [45], attributes having strong correlation cannot be the part of feature subset. Besides, more the attributes are independent among themselves and more information gain they will have which would eventually give better outcomes over unseen data. The present research focuses on medical (obstetrical) datasets which are more sensitive in nature, so feature selection approach is more effective for such datasets. In light of this point, a feature selection (entropy-notion) approach is presented here to extract the most relevant features from obstetrical (term-preterm) dataset. These features are utilized to classify all birth cases into term birth and PTB. A conceptual model of the proposed approach is shown in Figure 1.

The proposed approach is stated as follows:

(i) Suppose that D is a medical dataset having n attributes, say A_i for $i = 1, 2, 3, \ldots, n$.

Let F_0 denote a set of features in the original dataset D.

Initially, $F_0 = \{A_1, A_2, A_3, \ldots, A_n\}$.

Since D is divided into three distinct subsets as D_1, D_2, and D_3, so after applying the proposed approach, we get three feature subsets, namely, F_1, F_2, and F_3 from these data subsets.

F is considered as a resultant feature set derived from F_1, F_2, and F_3. Initially, $F_k = F_0$ for $k = 1, 2, 3$.

Let P be a classification problem described by a set of n attributes, say A_i for $i = 1, 2, 3, \ldots, n$ and also consider that F represents the set of features derived from the original dataset.

Initialize, $F = F_0 = \{A_1, A_2, A_3, \ldots, A_n\}$.

for each data subset $D_i \in D$; where $i = 1, 2, 3$

do

for each attribute $A_i \in F_0$

do

Calculate Gain (S, A_i)//information gain for A_i

Using formula stated below,

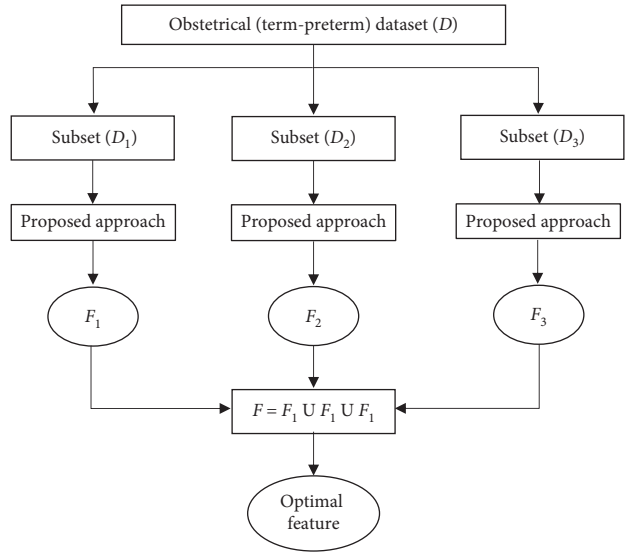

Figure 1: Conceptual model for feature selection approach.

$\text{Gain}(S, A_i) = \text{Entropy}(S) - \Sigma_{v_j \in A_i} (|S_{v_j}|/|s|)$ $\text{Entropy}(S_{v_j})$, where v_j denotes values of attribute A_i and $Entropy(S) = \Sigma p_m \log_2 p_m$, where S represents the number of instances in P and p_m is the nonzero probability of s_m instances (out of S) belonging to class m, out of c classes.

end for

Compute $r = (Max\_Gain(S, A_i) - Min\_Gain(S, A_i)/n)$, where $i = 1, 2, 3, \ldots, n$.

//Here, r is considered as a threshold value for selecting features

for each attribute $Ai \in F_0$

do

if $Gain(S, A_i) < r$

then

update $F_k = F_k - \{A_i\}$//removing A_i from F_k

end if

end for

end for

$F = F_1 \cup F_2 \cup F_3$//including all attributes of F_1, F_2, and F_3.

Note. The proposed feature selection approach in this study is a form of the filter method and is implemented in Java-1.4.

Time Complexity. The algorithm is simple and easy to understand. The running time of an algorithm is $O(n)$, where n is the number of attributes in the dataset.

4.3. The Proposed Framework: Risk Prediction Conceptual Model (RPCM). Based on novel feature selection (entropy-notion) approach and several studies in [46–49], RPCM is carefully designed to predict the risk of PTB in pregnant women. The workflow of the framework consisting of three stages (Stage-I, Stage-II, and Stage-III) is depicted in Figure 2, and then its each component is detailed.

4.3.1. Key Components of the Proposed Model. The proposed model consists of some key components, namely, healthcare centre, patient survey, maternal and neonatal records, data preprocessing, machine learning, and birth outcome. Each of these is discussed as follows:

(i) *Healthcare Centre*. A healthcare centre is a part of a network of hospitals employed by a group of general physicians, nurses, and healthcare professionals that provide healthcare facilities to people in a certain area. In addition to standard medical treatments, one of the main goals of the primary healthcare centre is maternal care during pregnancy especially in rural India. This is because people from rural India avoid contacting healthcare professionals and practitioners for pregnancy care which increases the cases of maternal and neonatal deaths.

(ii) *Patient Survey*. A comprehensive care to mother and child is primarily concerned to all healthcare systems in India. The term survey describes any study that consists of requesting people to respond queries. This entails researcher-developed questionnaires and personal interviews with pregnant women during their antenatal care visits.

(iii) *Maternal and Neonatal Records*. Maternal and neonatal records play a vital role in deciding the way healthcare services are provided, accessed, and affected by health outcomes. It stores the statistical reports describing the use of prenatal services, maternal risk factors, and birth outcomes for all patients residing in rural area. PTB is one of the most frequent complication of pregnancy. It occurs due to several medical reasons and is affected by

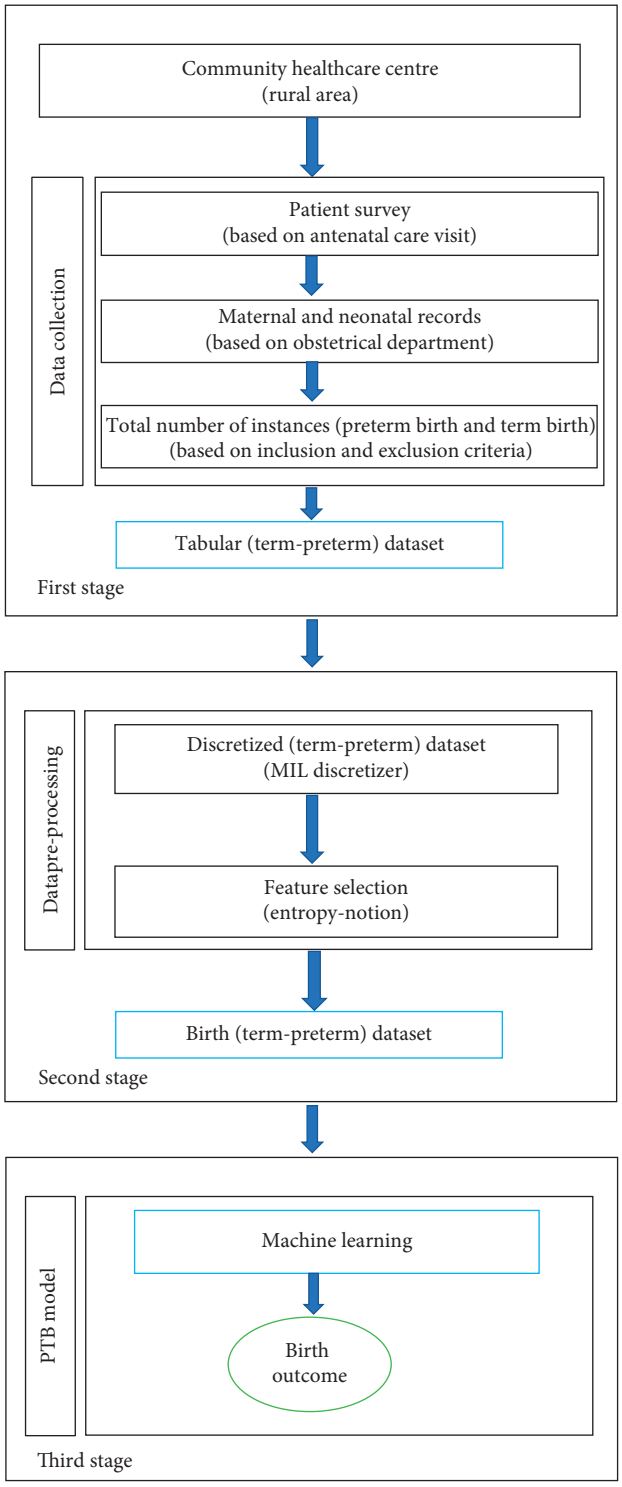

Figure 2: Framework of the proposed model.

some of the important maternal features based on human experts (experience) and several research studies [50–53]. These maternal features are critical in nature to predict cases of PTB. The total number of birth instances is taken from the obstetrical data.

(iv) *Data Discretization*. A technique of converting continuous values of attribute into a finite set of intervals and associating a new discrete value with each interval is known as data discretization. Since any classifiers prefer to handle discrete values rather than continuous values for the learning process, data discretization plays a crucial role in the process of machine learning. The study in [54] suggests that data discretization improves the quality of discovered knowledge, and it is based on the concept of information theory.

(v) *Feature Selection*. One of the core concepts in machine learning is the feature selection. Feature selection is the process of selecting those features from the input datasets which highly impact the performance of the predictive model. The present study focuses on feature selection approach based on entropy notion as already discussed in Section 4.2.

(vi) *Data Preprocessing*. The tabular dataset collected from obstetrical data is preprocessed and converted into a normalized form with the help of MIL discretizer [55, 56].

(vii) *Machine Learning (ML)*. The present study focuses on applying machine learning algorithms [46, 49] for PTB prediction. ML is a method of data analysis that automates analytical model building. Classification is one of the most popular approaches for applying ML methods (e.g., DT, LR, and SVM). These techniques are used to in medical domain for classification, prediction, and diagnosis purposes.

(viii) *Birth Outcome*. This component is very crucial in preventing preterm delivery in pregnant women during antenatal care clinics. The predicted birth outcome can also be used to properly analyse the key maternal features responsible for PTB.

4.4. Details of Stage-I. The main role of the first stage of framework is to collect obstetrical data from the Community Healthcare Centre, and it is detailed in this section.

4.4.1. Study Design. The study was conducted in the Community Health Centre, Kamdara (Gumla), situated in rural area of Jharkhand, during a period from July 2018 to September 2020. The hospital provides obstetric and gynaecological services to all categories of women, whether registered for antenatal care or referred. The approval for the study was taken from the Institutional Ethics Committee.

Selection Criteria. The selection of patients (women) depends on the following inclusion-exclusion criteria:

Inclusion criteria include the following:

(i) Women registered for ANC and having birth at the Community Health Centre
(ii) Women having birth occurring at the gestational age of 28 weeks or more
(iii) Women who delivered a live birth

Exclusion criteria include the following:

(i) Women having still birth
(ii) Women having birth of twins
(iii) Women referred to other hospitals

4.4.2. Data Collection. The basic step of Stage-I is to collect data based on patient survey and maternal records available in the obstetrics department. Initially, 1800 records were collected during a research period. Then, 1300 records were selected for further study based upon inclusion-exclusion criteria. The collected records include all instances of term birth and PTB. A manual analysis is performed to select all maternal features which are involved during pregnancy (based on medical experts' opinion and several research studies) [51, 52, 57, 58]. The description of the obstetrical dataset (original) after data collection is summarized in Table 2.

Initially, all instances are in a raw-form which are compiled into a tabular-form using MS Excel program. As a result, a tabular (term-preterm) dataset is prepared for the research purpose. The tabular (term-preterm) dataset used in this work is a binary class dataset.

The feature values in this dataset are of the form-string, integer, and continuous. The tabular (term-preterm) dataset consists of 1300 instances, composed of thirty-six different features which are taken into consideration before, during, and after pregnancy. These features are listed in Table 3. The questionnaire used for data entry during patient survey was mainly focused on their background details, medical history, previous pregnancy details, current pregnancy details, baby details, and medical disorders in current pregnancy.

4.5. Description of Stage-II. The collected data from tabular (term-preterm) dataset are preprocessed at the second stage of the framework. This stage deals with two main operations, namely, data discretization and feature selection.

4.5.1. Data Discretization. During data preprocessing, tabular (term-preterm) dataset is converted into a normalized form with the help of data discretization process. This gives a discretized (term-preterm) dataset. This dataset is utilized to select most accurate features by applying suggested feature selection approach based on the notion of entropy. The initial statistics of discretized (term-preterm) dataset is shown in Table 4.

In reality, attributes of any medical dataset may contain mixture of string, continuous, outliers, and missing data. Many classifiers cannot handle continuous attributes but each of them can operate on discretized attributes [55]. Besides, performance of classifiers can be significantly improved by replacing continuous attributes with its discretized values. Depending upon the amount of missing data and the criticality of the feature in which the data is missing, it may impact the accuracy of prediction. In this study, the missing value in any feature is replaced with the mean value of that feature, and minimum information loss (MIL) data discretizer [12, 54, 59] is employed here for data processing, which make data compatible with the machine learning algorithm.

4.5.2. Feature Selection. After that, the proposed feature selection approach is taken into consideration to select the most probable features (responsible for PTB) from the discretized (term-preterm) dataset. As a result, seventeen different features are selected from this dataset. These maternal features (listed in Table 5) are also considered as major risk factors for PTB as suggested by medical experts and several research studies. Then, a final birth (term-preterm) dataset, consisting of these selected features, is prepared for the last stage of framework. The birth dataset also contains 1300 instances of term birth and PTB.

4.6. Description of Stage-III. Finally, a machine learning-based prediction model for PTB is built at this stage. This section describes the actual construction of the suggested system.

4.6.1. Machine Learning PTB Model. The aim of this research is to find a suitable classifier which can predict the PTB with more accuracy. The three classifiers, namely, decision tree (DT), logistic regression (LR), and support vector machine (SVM) are used in this analysis. The method of selecting classifier in this study is illustrated in Figure 3. Model fitting was carried out by dividing the input dataset into training dataset and test dataset at a ratio of 70% and 30%, respectively. The training set is used in learning phase and test set is used in prediction phase, to determine the best model. Researchers may find ample information about several machine learning classifiers from the articles [60–63].

4.6.2. Evaluation of Machine Learning Classifiers. The empirical measures can be extracted from the confusion matrix in order to evaluate the performance of the learning classifier [64]. A confusion matrix shows the accuracy of the solution to a classification problem. Table 6 depicts the confusion matrix, which summarizes the number of instances predicted correctly or incorrectly by a classification model.

Furthermore, the other parameters used to measure the classifier's performance are correct classification rate (CCR) or accuracy, true positive rate (TPR) or sensitivity, true negative rate (TNR) or specificity, false positive rate (FPR), false negative rate (FNR), precision, recall, and F1 score. A formal definition of these performance metrics is shown in Table 7.

5. Experimental Design and Results

5.1. Experimental Design. A birth (term-preterm) dataset with 1300 patients' observations is obtained in order to perform the experiment. The experiment is carried out with the help of Python and Scikit-Learn library or under WEKA toolbox (http://www.cs.waikato.ac.nz/ml/weka). The observations in the birth dataset are carefully reviewed for prediction of birth cases. This is in fact a binary class dataset in

Table 2: Summary of the obstetrical (term-preterm) dataset.

Problem name	Number of features	Number of classes	Number of instances
Birth case	36	2	1300

Table 3: Maternal features associated with PTB.

S. no.	Feature ID	Feature name
1	PID	Patient identification
2	WA	Woman age
3	LMP	Last menstrual period
4	EDD	Estimated delivery date
5	G	Gravida
6	P	Parity
7	A	Abortion
8	L	Living
9	EL	Educational level
10	H	Height
11	W	Weight
12	BMI	Body mass index
13	BP	Blood pressure
14	HB	Hemoglobin
15	ANC	Antenatal care visit
16	ADD	Actual delivery date
17	OH	Obstetric history
18	PCS	Previous caesarean section
19	GA	Gestational age
20	BW	Birth weight
21	GDM	Gestational diabetes mellitus
22	FHR	Fetal heart rate
23	MG	Multiple gestation
24	ND	Normal delivery
25	MH	Previous medical history
26	LBW	Low birth weight
27	ASPX	Asphyxia
28	HT	Hypertension
29	PE	Preeclampsia
30	LV	Live birth
31	SB	Still birth
32	OB	Obesity
33	AN	Anemia
34	TH	Thyroid
35	NS	Neonatal status
36	PTB	Preterm birth

Table 4: Summary of discretized (term-preterm) dataset.

Outcome	*N*
Number of features	36
Number of classes	2
Total instances	1300
Term birth	991
Preterm birth	309

which all births occurring between 28^{th} to 37^{th} weeks are termed as PTB class with label "1" whereas all births after 37^{th} weeks are termed as term birth (TB) class with label "0." According to the study, around **24%** of the findings in the dataset are of PTB with label "1" and remaining **76%** are of TB with label "0." Hence, PTB class is dominated by TB class, and we can say that PTB is the minority class and TB is the majority class. Therefore, there is a need of a good sampling technique for medical datasets [24, 52]. In this context, synthetic minority oversampling technique (SMOTE) is used to balance the target dataset [65]. This can be achieved by replicating the PTB cases until it reaches approximately 50% of the dataset. This gives a new balanced (term-preterm) dataset.

Table 5: List of excellent features in discretized (term-preterm) dataset.

Feature code	Feature name	Feature type
WA	Woman age	Numeric
PT	Parity	Numeric
GD	Gravida	Numeric
BMI	Body mass index	Ordinal
ANC	Antenatal care visit	Numeric
GA	Gestational age	Numeric
FHR	Fetal heart rate	Numeric
BP	Blood pressure	Ordinal
HB	Hemoglobin	Numeric
GDM	Gestational diabetes mellitus	Binary
PE	Preeclampsia	Binary
HT	Hypertension	Binary
OH	Obstetric history	Binary
EL	Education level	Ordinal
CS	Previous caesarean section	Binary
MH	Previous medical history	Binary
PTB	Preterm birth (target variable)	Binary

5.2. Results and Discussion. A total of 1300 patients (women) were selected in this study based on inclusion-exclusion criteria. Out of 1300 pregnant women, 309 women were having preterm birth and rest 991 women were having term birth. Thus, the incidence of PTB is 23.78% of total pregnant women. In this work, the performance of DT, LR, and SVM classifiers is evaluated in terms of accuracy, specificity, and sensitivity [66]. With these indicators, it is possible to compare the proposed model performance with three classifiers. Tables 8 and 9 present the performance metrics of classifiers for the original dataset and balanced dataset, respectively.

Based on the results shown in Tables 8 and 9, we can observe that the accuracy of three different classifiers is roughly around 85%. With respect to the original dataset, the accuracy of SVM is 86.1% which is highest, followed by LR and DT. The results were additionally improved (after applying SMOTE) with the balanced dataset. The accuracy of SVM classifier in the balance dataset increases from 86.1% to

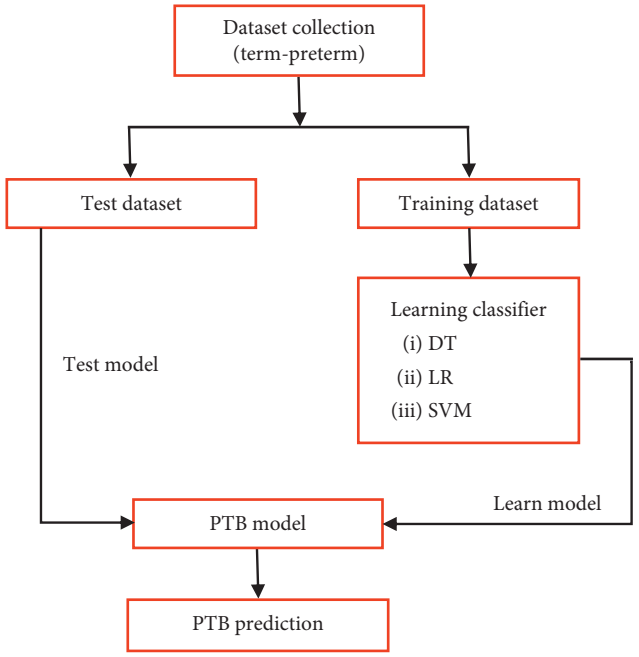

Figure 3: A conceptual PTB prediction model.

Table 6: Confusion matrix.

	Predictive positive	Predictive negative
Actual positive	True Positive (TP)	False Negative (FN)
Actual negative	False Positive (FP)	True Negative (TN)

Table 7: Performance metrics for machine learning classifiers.

Metrics	Formula
CCR	$((TP + TN)/(TP + FP + FN + TN))\%$
TPR	$TP/(TP + FN)$
TNR	$TN/(TN + FP)$
FPR	$FP/(TN + FP)$
FNR	$FN/(TP + FN)$
Precision	$TP/(TP + FP)$
Recall	$TP/(TN + FN)$
F_1 score	$2 * TP/(2 * TP + Fp + FN)$

Table 8: Performance metrics of the classifiers—original dataset.

Classifiers	Accuracy	Sensitivity	Specificity
DT	0.777	0.702	0.930
LR	0.841	0.863	0.971
SVM	**0.861**	0.801	0.702

Table 9: Performance metrics of the classifiers—balanced dataset.

Classifiers	Accuracy	Sensitivity	Specificity
DT	0.796	0.713	0.972
LR	0.872	0.832	0.954
SVM	**0.909**	0.891	0.783

90.9% compared with original dataset. In summary, the SVM model is the best classifier in the experiment.

6. Conclusion and Future Scope

In this study, the proposed model (RPCM) can be used for prediction of PTB based on excellent features (text-based symptoms) available in obstetrical data. The work focuses on feature selection (entropy-notion) approach by applying machine learning classifiers (DT, LR, and SVM) in order to classify all birth cases into term birth and PTB. Comparing the performances of the classifiers, it is evident that SVM classifier is the most suitable classifier as it achieves an accuracy of 90.9%. According to the findings of this study, the identified risk factors (excellent features) will be helpful in the prediction of PTB, especially in rural community. The developed model supports the decision-making process in maternity care by identifying and alerting the pregnant women at risk of preterm delivery thereby preventing possible complications, reducing the diagnosis cost, and ultimately minimizing the risk of PTB. The present system can be regarded as a successful innovation in Obstetrics to give clinical support to patients during pregnancy consultations. In particular, RPCM claims to assist healthcare professionals to make effective and timely decisions without consulting specialists directly.

The limitation of the present research is that the risk factors for PTB are limited in size and dataset is small, which could be increased to improve the performance of the PTB prediction in the future studies. However, expert knowledge and clinical judgement may still be needed to interpret this risk and take appropriate action in individual cases.

Data Availability

The data used to support the finding of this study are available from the corresponding author upon reasonable request. The data are not publicly available due to privacy and ethical restrictions of Institutional Ethics Committee.

Disclosure

The content of this paper represents the views of the authors and do not necessarily reflect the views of the Community Health Centre.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors acknowledge the staff of Department of Gynaecology and Obstetrics at Community Health Centre (Kamdara, Jharkhand) for their efforts in the completion of this research work. The authors gratefully thank Dr. Ruchi Bhushan, MBBS, MS (OBG), for her valuable comments and helpful discussions. The authors would also like to convey their sincere gratitude to Mrs. Priyanka and Ms. Prerna for their neverending support and motivation.

References

[1] H. Blencowe, S. Cousens, D. Chou et al., "Born Too Soon: the global epidemiology of 15 million preterm births," *Reproductive Health*, vol. 10, no. 1, p. S2, 2013.

[2] L. Liu, S. Oza, D. Hogan et al., "Global, regional, and national causes of child mortality in 2000-13, with projections to inform post-2015 priorities: an updated systematic analysis," *The Lancet*, vol. 385, no. 9966, pp. 430–440, 2015.

[3] B. Govindaswami, P. Jegatheesan, M. Nudelman, and S. R. Narasimhan, "Prevention of prematurity," *Clinics in Perinatology*, vol. 45, no. 3, pp. 579–595, 2018.

[4] L. J. E. Meertens, P. van Montfort, H. C. J. Scheepers et al., "Prediction models for the risk of spontaneous preterm birth based on maternal characteristics: a systematic review and independent external validation," *Acta obstetricia et gynecologica Scandinavica*, vol. 97, no. 8, pp. 907–920, 2018.

[5] M. J. Perez, J. J. Chang, L. A. Temming et al., "Driving factors of preterm birth risk in adolescents," *American Journal of Perinatology Reports*, vol. 10, no. 3, pp. e247–e252, 2020.

[6] S. Shrestha, S. S. Dangol, M. Shrestha, and R. P. B. Shrestha, "Outcome of preterm babies and associated risk factors in a hospital," *Journal of the Nepal Medical Association*, vol. 50, no. 180, 2010.

[7] C. Catley, M. Frize, R. C. Walker, and D. C. Petriu, "Predicting preterm birth using artificial neural networks," in *In Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05)*, pp. 103–108, Dublin, Ireland, June 2005.

[8] A. Belle, R. Thiagarajan, S. M. R. Soroushmehr, F. Navidi, D. A. Beard, and K. Najarian, "Big data analytics in healthcare," *BioMed Research International*, vol. 2015, Article ID 370194, 16 pages, 2015.

[9] R. Raja, I. Mukherjee, and B. K. Sarkar, "A systematic review of healthcare big data," *Scientific Programming*, vol. 2020, Article ID 5471849, 15 pages, 2020.

[10] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *IEEE Access*, vol. 5, pp. 8869–8879, 2017.

[11] A. Mansoul, B. Atmani, and S. Benbelkacem, "A hybrid decision support system: application on healthcare," 2013, http://arxiv.org/abs/1311.4086.

[12] B. K. Sarkar, "A two-step knowledge extraction framework for improving disease diagnosis," *The Computer Journal*, vol. 63, no. 3, pp. 364–382, 2020.

[13] M. Seera and C. P. Lim, "A hybrid intelligent system for medical data classification," *Expert Systems with Applications*, vol. 41, no. 5, pp. 2239–2249, 2014.

[14] A. Bhardwaj and A. Tiwari, "Breast cancer diagnosis using genetically optimized neural network model," *Expert Systems with Applications*, vol. 42, no. 10, pp. 4611–4620, 2015.

[15] P. J. Lisboa and A. F. G. Taktak, "The use of artificial neural networks in decision support in cancer: a systematic review," *Neural Networks*, vol. 19, no. 4, pp. 408–415, 2006.

[16] T. M. Mitchell, "Machine learning and data mining," *Communications of the ACM*, vol. 42, no. 11, pp. 30–36, 1999.

[17] A. Callahan and N. H. Shah, "Machine learning in healthcare," in *Key Advances in Clinical Informatics*, pp. 279–291, Academic Press, Cambridge, MA, USA, 2017.

[18] L. Fu, "Knowledge discovery based on neural networks," *Communications of the ACM*, vol. 42, no. 11, pp. 47–50, 1999.

[19] G. P. Zhang, "Neural networks for classification: a survey," *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)*, vol. 30, no. 4, pp. 451–462, 2000.

[20] W. Klosgen and J. M. Zytkow, *Handbook of Data Mining and Knowledge Discovery*, Oxford University Press, Oxford, UK, 2002.

[21] R. Pari, M. Sandhya, and S. Sankar, "Risk factors based classification for accurate prediction of the Preterm Birth," in *Proceedings of the 2017 International Conference on Inventive Computing and Informatics (ICICI)*, pp. 394–399, IEEE, Coimbatore, India, November 2017.

[22] "The global burden of preterm birth," *The Lancet*, vol. 374, 2009.

[23] S. Beck, D. Wojdyla, L. Say et al., "The worldwide incidence of preterm birth: a systematic review of maternal mortality and morbidity," *Bulletin of the World Health Organization*, vol. 88, no. 1, pp. 31–38, 2010.

[24] S. Saigal and L. W. Doyle, "An overview of mortality and sequelae of preterm birth from infancy to adulthood," *The Lancet*, vol. 371, no. 9608, pp. 261–269, 2008.

[25] I. Rudan, K. Y. Chan, J. S. Zhang et al., "Causes of deaths in children younger than 5 years in China in 2008," *The Lancet*, vol. 375, no. 9720, pp. 1083–1089, 2010.

[26] M. C. Hogan, K. J. Foreman, M. Naghavi et al., "Maternal mortality for 181 countries, 19802008: a systematic analysis of progress towards Millennium Development Goal 5," *The Lancet*, vol. 375, no. 9726, pp. 1609–1623, 2010.

[27] World Health Organization (WHO), *Commission on Information and Accountability for Women's and Children's Health. Keeping Promises, Measuring Results*, WHO, Geneva, Switzerland, 2015.

[28] S. R. Walani and J. Biermann, "March of Dimes Foundation: leading the way to birth defects prevention," *Public Health Reviews*, vol. 38, no. 1, pp. 1–7, 2017.

[29] Y. Dong and J.-L. Yu, "An overview of morbidity, mortality and long-term outcome of late preterm birth," *World Journal of Pediatrics*, vol. 7, no. 3, pp. 199–204, 2011.

[30] J. R. G. Challis, S. J. Lye, W. Gibb, W. Whittle, F. Patel, and N. Alfaidy, "Understanding preterm labor," *Annals of the New York Academy of Sciences*, vol. 943, no. 1, pp. 225–234, 2001.

[31] H. Witten Ian and F. Eibe, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Publishers, San Francisco, CA, USA, 2nd edition, 2005.

[32] Y. Li, T. Li, and H. Liu, "Recent advances in feature selection and its applications," *Knowledge and Information Systems*, vol. 53, no. 3, pp. 551–577, 2017.

[33] H. Huan Liu and L. Lei Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 4, pp. 491–502, 2005.

[34] B. M. Mercer, R. L. Goldenberg, A. Das et al., "The preterm prediction study: a clinical risk assessment system," *American Journal of Obstetrics and Gynecology*, vol. 174, no. 6, pp. 1885–1895, 1996.

[35] L. K. Goodwin, M. A. Iannacchione, W. E. Hammond, P. Crockett, S. Maher, and K. Schlitz, "Data mining methods find demographic predictors of preterm birth," *Nursing Research*, vol. 50, no. 6, pp. 340–345, 2001.

[36] C. Gao, S. Osmundson, D. R. Velez Edwards, G. P. Jackson, B. A. Malin, and Y. Chen, "Deep learning predicts extreme preterm birth from electronic health records," *Journal of Biomedical Informatics*, vol. 100, p. 103334, 2019.

[37] A. Weber, G. L. Darmstadt, S. Gruber et al., "Application of machine-learning to predict early spontaneous preterm birth among nulliparous non-Hispanic black and white women," *Annals of Epidemiology*, vol. 28, no. 11, pp. 783–789, 2018.

[38] M. Mailath-Pokorny, S. Polterauer, M. Kohl et al., "Individualized assessment of preterm birth risk using two modified prediction models," *European Journal of Obstetrics & Gynecology and Reproductive Biology*, vol. 186, pp. 42–48, 2015.

[39] M. Son and E. S. Miller, "Predicting preterm birth: cervical length and fetal fibronectin," *Seminars in Perinatology*, vol. 41, no. 8, pp. 445–451, 2017.

[40] U. Elaveyini, S. P. Devi, and K. S. Rao, "Neural networks prediction of preterm delivery with first trimester bleeding," *Archives of Gynecology and Obstetrics*, vol. 283, no. 5, pp. 971–979, 2011.

[41] T. Khatibi, N. Kheyrikoochaksarayee, and M. M. Sepehri, "Analysis of big data for prediction of provider-initiated preterm birth and spontaneous premature deliveries and ranking the predictive features," *Archives of Gynecology and Obstetrics*, vol. 300, no. 6, pp. 1565–1582, 2019.

[42] M. Colstrup, E. R. Mathiesen, P. Damm, D. M. Jensen, and L. Ringholm, "Pregnancy in women with type 1 diabetes: have the goals of St. Vincent declaration been met concerning foetal and neonatal complications?" *The Journal of Maternal-Fetal & Neonatal Medicine*, vol. 26, no. 17, pp. 1682–1686, 2013.

[43] H. Blencowe, S. Cousens, M. Z. Oestergaard et al., "National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications," *The Lancet*, vol. 379, no. 9832, pp. 2162–2172, 2012.

[44] L. Liu, H. L. Johnson, S. Cousens et al., "Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000," *The Lancet*, vol. 379, no. 9832, pp. 2151–2161, 2012.

[45] M. A. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in *Proceedings of Seventeenth International Conference on Machine Learning*, pp. 359–366, Stanford, CA, USA, July 2000.

[46] K. Y. Ngiam and I. W. Khor, "Big data and machine learning algorithms for health-care delivery," *The Lancet Oncology*, vol. 20, no. 5, pp. e262–e273, 2019.

[47] J. G. Nam, S. Park, E. J. Hwang et al., "Development and validation of deep learning-based automatic detection algorithm for malignant pulmonary nodules on chest radiographs," *Radiology*, vol. 290, no. 1, pp. 218–228, 2019.

[48] Z. Ahmed, K. Mohamed, S. Zeeshan, and X. Dong, "Artificial Intelligence with Multi-Functional Machine Learning Platform Development for Better Healthcare and Precision Medicine," *Database*, vol. 2020, 2020.

[49] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019.

[50] S. Pereira, F. Portela, M. F. Santos, J. Machado, and A. Abelha, "Predicting preterm birth in maternity care by means of data mining," in *Portuguese Conference on Artificial Intelligence*, pp. 116–121, Springer, Berlin, Germany, 2015.

[51] K. L. Courtney, S. Stewart, M. Popescu, and L. K. Goodwin, "Predictors of preterm birth in birth certificate data," *Studies in Health Technology and Informatics*, vol. 136, pp. 555–60, 2008.

[52] N. S. Prema and M. P. Pushpalatha, "Machine learning approach for preterm birth prediction based on maternal chronic conditions," in *Emerging Research in Electronics, Computer Science and Technology*, pp. 581–588, Springer, Berlin, Germany, 2019.

[53] F. Cunningham, K. Leveno, S. Bloom, C. Y. Spong, and J. Dashe, *Williams Obstetrics, 24e*, McGraw-Hill, New York, NY, USA, 2014.

[54] R. Jin, Y. Breitbart, and C. Muoh, "Data discretization unification," *Knowledge and Information Systems*, vol. 19, no. 1, pp. 1–29, 2009.

[55] B. K. Sarkar, S. S. Sana, and K. Chaudhuri, "MIL: a data discretisation approach," *International Journal of Data Mining, Modelling and Management*, vol. 3, no. 3, pp. 303–318, 2011.

[56] G. Mitra, S. Sundareisan, and B. K. Sarkar, "A simple data discretizer," 2017, http://arxiv.org/abs/1710.05091.

[57] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8–17, 2015.

[58] N. Jothi, N. A. A. Rashid, and W. Husain, "Data mining in healthcare-a review," *Procedia Computer Science*, vol. 72, pp. 306–313, 2015.

[59] M. Boulle, "Khiops: a statistical discretization method of continuous attributes," *Machine Learning*, vol. 55, no. 1, pp. 53–69, 2004.

[60] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, Pearson Education India, New Delhi, India, 2016.

[61] I. H. Witten and E. Frank, "Data mining," *ACM Sigmod Record*, vol. 31, no. 1, pp. 76-77, 2002.

[62] A. Dhillon and A. Singh, "Machine learning in healthcare data analysis: a survey," *Journal of Biology and Today's World*, vol. 8, no. 6, pp. 1–10, 2019.

[63] M. A. Ahmad, C. Eckert, and A. Teredesai, "Interpretable machine learning in healthcare," in *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 559-560, Washington, DC, USA, August 2018.

[64] Q. Gu, L. Zhu, and Z. Cai, "Evaluation measures of the classification performance of imbalanced data sets," in *International Symposium on Intelligence Computation and Applications*, pp. 461–471, Springer, Berlin, Germany, 2009.

[65] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[66] W. Zhu, N. Zeng, and N. Wang, "Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations," *NESUG Proceedings: Health Care and Life Sciences*, vol. 19, 2010.