ISSN: 0976-3104
ISSUE: COMPUTER SCIENCE

PREDICTION OF PRETERM BIRTH USING DATA MINING-A SURVEY

Prema N. S.[1*], Pushpalatha M. P.[2]

[1 *]*Department of Information Science and Engineering, Vidyavardhaka College of Engineering, Mysuru, INDIA*

[2]*Department of Computer Science and Engineering, Sri Jayachamarajendra College of Engineering, Mysuru, INDIA*

ABSTRACT

Background: *Preterm birth (PTB), the birth of infant before 37 weeks of gestation and it is the leading reason for perinatal mortality and morbidity. The reason for PTB is unclear, although many data mining techniques are used in identifying the risk factors. The reasons may be because of large and high dimensional data set. Many data mining (DM) methods are used to explore the risk factors of PTB and to predict preterm birth. In this paper, we describe the work carried out on application of data mining techniques towards the analysis and prediction of preterm birth.* ***Methods:*** *This survey paper is prepared by considering two criteria one the data mining techniques used for analysis and prediction and the other is the risk factors considered for the preterm birth prediction. Since in the works considered for survey the authors have not used a common or standard data set for the study/analysis, the comparisons of the works were difficult, here we have made an attempt to explore the data mining techniques for preterm birth prediction and the major risk factors of preterm birth.* ***Conclusion:*** *The most commonly used data mining technique is classification in that the usual techniques are Support Vector Machine (SVM) and Logistic regression. The most commonly considered risk factors are socio-demographic, behavioral (life style) and Pregnancy History.*

KEY WORDS
Data mining, Preterm birth, Logistic regression, Naive bayes, SVM, Neural networks, Decision tree

INTRODUCTION

Preterm birth is one of the public health risks in the society causing death and complications during pregnancy. A delivery that occurs twenty weeks after the start of labor and before 37 weeks of pregnancy is called preterm birth [1]. WHO reports that, about 15 million babies are born prematurely every year around the world and that is more than one in 10 of all babies born globally. In India, 3,341,000 preterm cases in every year and 361,600 children below five years die because of preterm complications [2]. The cause of preterm birth is complex, multi-factorial and not fully understood [3].

Types of preterm births

i. Spontaneous: This may occur after spontaneous on set labor or by prelabour premature rupture of membrane (PPROM). The cause of this type of preterm birth is very difficult to identify in up to half of all cases.
ii. Provider initiated: It is type of elective or induction labor before completion of 37 weeks of gestations for different reasons like fetal or maternal condition or other medical reasons.

Most occurring are spontaneous preterm births but some are due to early induction for medical or non-medical reasons. About 45-50% of preterm births are because of unknown reasons/causes, 30% are due to premature rupture of membrane(PROM) and other 15-20% is elective or medically indicated preterm deliveries.

Received: 30 Oct 2018
Accepted: 15 Dec 2018
Published: 10 Jan 2019

The main Risk factors for preterm birth are

i. Age of the women at pregnancy.
ii. History of preterm birth.
iii. Multiple pregnancies.
iv. Chronic diseases such as diabetes, hypertension, anaemia, asthma, thyroid disease.
v. Infection.
vi. Genetic influences.
vii. Nutritional factors: Under nutrition, obesity, nutritional deficiencies.
viii. Life style-women: smoking, consumption of alcohol, amusing drugs, Stress, excessive physical work [4].

Medical data mining

Data mining is the method of extracting useful knowledge from large repository of data. Medical data mining is an application of data mining, where data mining methods used for the analysis of medical data. The Medical data mining approaches are applied for the following tasks: treatment, prognosis, diagnosis, management and monitoring. The aim of medical data mining is to help and support physicians, care patients and improve public health.

The two main approaches of data mining are prediction and description. Prediction includes classification and regression,description includes clustering and association analysis.

In this paper, we describe work carried out on application of data mining techniques towards the analysis and prediction of preterm birth. We mainly considered the two things one is with respect to the data

***Corresponding Author**
Email:
prema.gowda@gmail.com

IIOAB JOURNAL
ISSN: 0976-3104

mining techniques used for the prediction and other is the risk factors /attributes considered for prediction.

RESEARCH PROGRESS

Many data mining models were developed for the prediction of PTB and the performances of the models are measured in terms of accuracy, AUC or ROC etc. The following are the models developed by various researchers.

A data mining model was developed on the raw data to identify the risk of PTB, monitoring and to control the PTB associated problems. The DM methods are applied on five different scenarios by combining different variables and also oversampling is done to balance the data by replicating the PTB cases so that the dataset reaches equal distribution of positive and negative instances. The risk factors considered for preterm birth prediction are pregnancy characteristics, Gestation and physical conditions of pregnant women [5].

Batoul ahadi et.al used statistical methods; support vector machine with different kernel functions and logistic regression models for preterm birth prediction. The data used is collected in 3 parts during nine months of pregnancy in every three months using questionnaires. The risk factors considered are demographic and pregnancy characteristics. They have applied Wrapper feature subset selection approach for the selection of best features [6].

The prediction model for PTB was developed by implementing risk of preterm delivery by Creasy [7], on Maternal-fetal Units Network (MFMU) data set. The data set used includes all most all risk factors of preterm birth in different time points Because of large size of the data set they have done preprocessing of data to remove noise, handled missing values by combining attributes into different groups. The authors conclude that identifying the risk factors of PTB is not elusive it requires an efficient model for correct prediction. They have mainly focused on first time mothers (nulliparous women) because of unavailability of pregnancy history [8]. In their further work they have demonstrated that model selection and nonlinear kernel methods for the prediction of PTB are better approaches for promising results [9].

Tuyen Tran et.al have presented methods for preterm birth prediction which includes quantifying, discovering risk factors, derivation of interpretable prediction rules and usage of stabilized sparse logistic regression for deriving linear prediction models. The authors also used Randomized Gradient Boosting a hybrid model to estimate the upper-bound accuracy for the data [10].

Cluster analysis techniques were used to get mutual exclusive clusters, as there were very low similarities in the set of attributes; they went for classification techniques, the accuracy of the classifiers are improved by applying cross validation. The classification techniques used are SVM and Naïve Bayes with the highest accuracy of 90% and 88% respectively [11].

Hsiang-Yang Chen et.al has explored the major risk factors of preterm birth using neural networks and decision tree. A neural network is used to find the top 15 risk factors then they have used decision tree C5.0 for classification. They have achieved an average accuracy of 80% and their results shows that hemorrhage in pregnancy and multiple birth are the major risk factors for preterm birth. Further they have also considered paternal risk factors like drinking alcohol, smoking and occupation in their study [12].

The prediction of PTB was done by using associative classifier by Yavar Naddaf et.al, they have also applied number of classification methods on a data set by considering both maternal and fetal records in predicting preterm birth. The Performance of the methods is very poor and there is no much improvement in the performance after applying feature selection techniques [13].

Christina Catley et.al developed an artificial neural network(ANN) model for the prediction of preterm birth on Perinatal Partnership Program of Eastern and South eastern Ontario (PPPESO) database based on physician input, eight obstetrical variables namely Age, Current pregnancy number of babies, Parity, Baby's gender, intention of breastfeed, Smoking, Previous term and preterm babies. The authors also shown the effect of change in the prior distribution of data on the performance of back propagation feed-forward ANN and they have assessed the effectiveness of weight elimination cost function in improving the accuracy [14]. In their further work they have applied ANN for prediction of high risk preterm birth. The back propagation feed forwarded ANN was developed by considering variables describing pregnancy history for high risk preterm birth prediction, the performance is measured in terms of sensitivity (36%&37%) and specificity (88% &92%) [15].

M. M. Van Dyne et.al used Learning from Examples using Rough sets (LERS) where rules are generated directly from the data for preterm birth prediction. The LERS model was able to predict 78% correctly of full term cases and 90% of preterm cases when full term and preterm rules were run separately and 73% of accuracy for combined cases [16].

Linda Goodwin and Sean Meher used the following techniques for preterm birth prediction namely neural networks, logistic regression, CART and software (PVRuleMiner and FactMiner) there was very small differences in the performances of the techniques. The data set used for the study is from Duke University

IIOAB JOURNAL
ISSN: 0976-3104

medical center data repository, they have considered about 32 demographic parameters for preterm birth prediction [17].

The usage of data mining techniques for preterm birth prediction was shown by considering about seven demographic parameters. The obtained results were interesting but the concern is whether the particular demographic data selected would represent the general population [18].

A secondary analysis was done by Courtney et al. describing that the preterm prediction model generated in [18] using demographic parameters is generalizing to a larger population with a modest result. In the study the data set used is birth certificate data, the demographic parameters considered for prediction are only five namely Age, marital status, ethnicity/race, education and country. The acceptable average AUC result for classifiers is 0.58 for different mining techniques [19].

In our work we have also used the statistical models for the prediction of PTB; the main risk factors considered are Age, Number of Times Pregnant, Diabetes, Obesity and Hypertension. The data set considered the details of pregnant women having either diabetes mellitus or gestational diabetes mellitus. The data set was balanced by using data oversampling techniques Synthetic Minority Oversampling Techniques (SMOTE). Comparing with work carried out by other researcher's highest accuracy achieved but the risk factors considered are limited and data sample size is also small [20].

Table 1: Data mining methods and accuracy for preterm birth prediction

Sr. No.	Data Mining Methods	Performance measures	Reference	Year
1	Decision Tree Generalized Linear Model Support Vector Machine Naïve Bayes	93% 86% 93% 74% (Accuracy)	[5]	2015
2	Support Vector Machine Logistic Regression	56% 67% (Accuracy)	[6]	2016
3	Support Vector Machine (Linear, Polynomial, RBF) Logistic Regression(Lasso, Elastic Net)	60% (Accuracy average)	[8]	2016
4	Logistic Regression Randomized Gradient Boosting(Ensemble Method)	62%/81.5% (Sensitivity/Specificity)	[10]	2016
5	Hierarchical Clustering Naïve Bayes Support Vector Machine	- 88% 90% (accuracy)	[11]	2010
6	Neural Networks & Decision Tree C5.0	80% (accuracy)	[12]	2010
7	Logistic Regression Naïve Bayes SVM Neural Networks Decision Tree C4.5 Associative Classifier	0.57 (Average AUC)	[13]	2006
8	ANN	36%/90% (Sensitivity/Specificity	[15]	2006
9	Neural Networks CART	0.64 0.65 (ROC)	[17]	2000
10	Logistic Regression Neural Networks SVM Bayesian Classifier CART	0.605 0.57 0.57 0.59 0.56 (AUC)	[19]	2008
11	Support Vector Machine Logistic Regression	87% 86% (Accuracy)	[20]	2018

IIOAB JOURNAL
ISSN: 0976-3104

Table 2: The most common Risk factors considered for preterm birth prediction

Factors Contributing to Preterm Birth	Variables	References
Demographic and Socioeconomic	Maternal Age Race/Ethnicity Educational Status Maternal Status Socioeconomic Status	[13] [8] [9] [11] [5] [12] [17] [18]
Behavioral Characteristics/ Life style	Alcohol Tobacco Recreational Drugs Psychological And Social Stress	[12] [11] [5] [8] [9] [10]
Maternal Health/ Chronic conditions	Body Mass Index (BMI) Diabetes Hypertension Anemia Asthma Thyroid Disease	 [5] [13] [20]
Current Fetal Conditions/Pregnancy Characteristics	Multiple Fetuses Infertility Treatments Infant weight Drugs used during pregnancy	 [5] [10] [6]
Pregnancy History/ Genetic Characteristics	Previous Preterm Births Diabetes Hypertension Obesity	[8] [9] [10] [13]
Biological Characteristics	Infections	[8] [9]
Others	Ultrasonography Insurance Details Cervical Measurements	 [5] [8] [9]

CONCLUSION

In this paper we have made an attempt study the work related to preterm term birth prediction using data mining techniques. The work considered for survey are studied based on the risk factors considered for prediction and the data mining techniques used, the comparative study was difficult as none of the authors have considered either the dataset or the risk factors in common.

The following conclusions can be drawn after the study of various works done in the application of data mining in preterm birth prediction.

1. Data set: All the works carried out in literature have not considered any data set in common for the comparison of performance of their proposed data mining techniques. All most all have used different data sets with varying size and dimensions.
2. Data sharing and privacy: As the work related to personal medical details of individual which will be circulated as commercial product which would help in research work in healthcare industry but at the same time it might threat privacy protection.
3. DM methods: The most commonly used data mining method is classification for predicting preterm birth and also for exploring risk factors of preterm birth. The very regularly used classification techniques are Support vector machine with different kernel functions and Logistic regression. Very few authors have used clustering and Association analysis techniques for preterm birth prediction but their performance is very poor compared to that of classification methods.
4. The main risk factor identified for the prediction is previous preterm birth but it is difficult in first time mothers because of unavailability of pregnancy history.
5. Performance measures: The classification techniques performances are measured in terms of accuracy, sensitivity/specificity ROC and AUC the details are shown in [Table 1].

IIOAB JOURNAL
ISSN: 0976-3104

6. Major risk factors: The different risk factors considered for PTB are shown in [Table 2]. The most commonly considered risk factors are socio-demographic, behavioral (life style) and Pregnancy History.

FUTURE WORK

In literature few works can be found where data mining techniques are applied for the prediction of preterm birth using Electro-hysterography(EHG) signal, EHG is used to measure electrical activity in the uterus. In this survey we have not included preterm birth prediction using EHG, which can be included in future.

CONFLICT OF INTEREST

None

ACKNOWLEDGEMENTS

The authors express gratitude towards the assistance provided by Accendere Knowledge Management Services Pvt. Ltd. in preparing the manuscripts. We also thank our mentors and faculty members who guided us throughout the research and helped us in achieving desired results.

FINANCIAL DISCLOSURE

None

REFERENCES

[1] Cunningham F, Leveno K, Bloom S, Hauth J, Rouse D. Spong CY.[2010] Williams ObstetricsUSA : The McGraw-Hill Companies, Inc. Medical Publishing Division.
[2] WHO. [Online] //www.who.int/topics/preterm_birth.
[3] National Health portal.[Online] www.nhp.gov.in/disease/reproductive-system/female-gynaecological-diseases-/preterm-birth
[4] Varney H, Kriebs JM, Gegor CL.[2004] Varney's midwifery: Jones & Bartlett Learning. Canada.
[5] Pereira S, Portela F, Santos MF, Machado J, Abelha A.[2015] Predicting Preterm Birth in Maternity Care by means of Data Mining. Progress in Artificial Intelligence 116-121.
[6] Batoul A, Hamid A, Soheila K, et al.[2016] Using support vector machines in predicting and classifying factors affecting preterm delivery. Journal of Paramedical Sciences. 7(3): ISSN 2008-4978.
[7] RK Creasy, BA Gummer, GC. Liggins.[1980] System for predicting spontaneous preterm. Obstetrics and Gynecology, 55(6):692-695.
[8] Ilia Vovsha, AnsafSalleb-Aouissi, Anita Raja, et al.[2016] Predicting Preterm Birth Is Not Elusive: Machine Learning Paves the Way to Individual Wellness. Proceedings of the 1st Machine Learning for Healthcare Conference, PMLR 56:55-72.
[9] lia Vovsha, Ashwath Rajan, Ansaf Salleb-Aouissi, et al. [2014] Predicting Preterm Birth Is Not Elusive: Machine Learning Paves the Way to Individual Wellness.Stanford University : Big Data Becomes Personal: Knowledge into Meaning :AAAI Spring Symposium Series.
[10] Tran T, Luo W, Phun D, Morris J, Rickard K. [2016] Preterm Birth Prediction: Deriving Stable and Interpretable Rules from High Dimensional Data. Proceedings of Machine Learning for Healthcare 2016 JMLR W&C Track (56).
[11] Adriana-Georgiana MALEA, Ştefan HOLBAN, Nicolae MELIȚĂ.[2010] Analysis and Determination of Risk Factors using R. 10th International Conference on DEVELOPMENT AND APPLICATION SYSTEMS, May 27-29
[12] Hsiang-Yang Chen, Chao-Hua Chuang ,Yao-Jung Yang , Tung-Pi Wuc[2011] Exploring the risk factors of preterm birth using data mining. Expert Systems with Applications.38 (5): 5384-5387.
[13] Yavar Naddaf, MojdehJalali Heravi and AmitSatsangi. [2008]. Predicting Preterm Birth Based on Maternal and Fetal Data. semanticscholar.
[14] C Catley, M Frize, RC Walker, DC Petriu. [2005] Predicting preterm birth using artificial neural networks. 18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05).1064-7125.
[15] Catley C, Frize M, Walker RC, petriu DC.[2006] Predicting High-Risk Preterm Birth Using Artificial Neural Networks. IEEE Transactions on Information Technology in Biomedicine, July 2006 , 10:540-549.
[16] Dyne MM Van, et al., et al.[1994]. Using Machine learning and expert systems to predict preterm delivery in pregnant women. USA : Tenth Conference on Artificial Intelligence for Applications.
[17] Goodwin L, Meher S.[2000] Data mining for preterm birth prediction. Proceedings of the 2000 ACM symposium on Applied computing, 46-51.
[18] Linda K Goodwin, Mary Ann Iannacchione, et al.[2001] Data mining methods find demographic predictors of preterm birth. Nursing research, 340-345.
[19] Courtney KL, Stewart S, Popescu M, Goodwin LK. [2008] Predictors of preterm birth in birth certificate data .Studies in health technology and informatics, February.136:555-560
[20] Prema NS, Pushpalatha MP. [2018] Machine learning approach for Preterm Birth Prediction Based on Maternal Chronic Conditions. International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT-2018).