husbands reported having a job (96 %) with a good social level (high income by 71 %). They were also dominantly in the age bracket of 25 to 35 (64%), residing in the city (76 %). About 33 % of the women had rather an obese BMI with 47 % presenting an excessive weight gain during the pregnancy. The percentage of mothers who smoked during pregnancy was 13%.

The dominant gynecological complications during past pregnancies were diabetes (5%) and Pre-eclampsia (4%). Approximately 31 % of them have had induction and 29 % hemorrhage.

Among all these factors, the covariates that presented the highest difference of percentage between the PTB positive and the negative classes were Pre-hemorrhage, Weight gain, Age, BMI, and Social status (Table I). The Chi-square test confirmed that most of these variables were statistically significant at least at the 5% level (Table I). Smaller, non-statistically significant, differences were observed for pre-diabetes, work husband, and pre-eclampsia. Pre-eclampsia and Pre-diabetes were discarded from further modeling analysis because they gave a low prevalence reaching even 0 for the positive class. Physicians watch very closely women with these complications for PTB risk. This may explain the low number of PTB incidence observed for women with these complications.

Table I. Characteristics for all multiparous women for the term and preterm classes.

Characteristic	Term (n=847)	Preterm (n=75) (Spontaneous <37 weeks)	P-value	Total (n=922) (Positive /total)
Age (25-35 years)	536(63)	58(77)	0.016	594(64)
BMI (obese)	254(30)	50(67)	0.000	304(33)
Education_husband(high)	691(81.6)	62(82.7)	0.816	753(81)
Education_mom (high)	667(78.7)	61(81.3)	0.660	728(79)
Pre_cesarean (presence)	297(35.1)	29(38.7)	0.532	326(35)
Pre_diabetes (presence)	40(4.7)	6(8)	0.438	46(5)
Pre_eclampsia (presence)	34(4)	0(0)	0.077	34(4)
Pre_hemmorrhage(presence)	215(25.4)	54(72)	0.000	269(29)
Pre_induction (presence)	263(31.1)	24(32)	0.865	287(31)
Residence(city)	644(76)	60(80)	0.438	704(76)
Smoking (smoker)	109(12.9)	11(14.7)	0.657	120(13)
Social_status (high)	614(72.5)	41(54.7)	0.010	655(71)
Weight_gain (excess)	378(44.6)	54(72)	0.000	432(47)
Work_husband (external job)	816(96.3)	74(98.7)	0.291	890(96)
Work_mom (external job)	560(66.1)	44(58.7)	0.206	604(65)

a: probability value for the Chi-square test.
(): percentage

Based on the above results we focused the remaining of this work on the multiparous women. We defined the original dataset including all the initial 922 women described by all the variables except Pre-eclampsia and Pre-diabetes.

The logistic regression analysis of the original dataset (glm) led to almost the same significant variables as the Chi-square test, except that Age was not significant while Residence was added to the list of significant co-factors (Table II).

Table II. Linear coefficients of each logistic regression model (significant at the level 5% *, 1%** and 1‰ ***)

Factors	Models[a]			
	glm	**Glmup**	**glmnetup**	**glmglmnetup**
Intercept	-4.56**	-1.39**	-3.72	-1.97***
Age1	.54	0.86***	0.33	0.68***
BMI1	1.07**	0.75***	0.35	0.70***
Education_hus1	-0.52	-0.02	.	
Education_mom1	-0.01	0.12	.	
Pre_cesarean1	-0.29	-0.52**	.	
Pre_hemmorrhage1	1.98***	2.11***	1.62	1.93***
Pre_induction1	-0.12	0.12	.	
Residence1	1.27*	1.30***	0.47	1.11***
Smoking1	0.12	0.24	.	
Social_status1	-1.42**	-1.82***	-1.04	-1.79***
Weight_gain1	1.03*	1.06***	0.76	1.07***
Work_hus1	-0.28	-0.64	.	
Work_mom1	-0.14	0.09	.	

[a]glm: logisitc regression on original data,
glmup: logisitc regression up-sample data,
glmnetup: LASSO regression on up-sample data,
glmglmnetup: Logistic regression with selected LASSO variables on up-sample data

In contrast, after creating a balanced sample using the up-sampling algorithm and running the logistic model (glmup) on these datasets, the results were notably improved for the PTB prediction as shown in Table III. Indeed, PTB prediction (PPV) increased from 12% for unbalanced to 92% for the up sampled data. Additionally, despite a small decrease the negative predictive value remained high around 75%. However, the number of misclassified non-PTB women (False Positives) significantly increased from 1 % in the unbalanced sample model (glm) to about 25%. The LASSO regularized model (glmnetup) gave comparable results. Nevertheless, the logistic regression with the selected variables by the LASSO regularization (glmglmnetup) of up-sampled data gave the best compromise between a low number of false positives (lower than 21%) and a high PTB prediction of 88% (PPV) along with a NPV of 75% (Table III).

Table III. Results of the of preterm and non-preterm (false positives) prediction (percentage) and the values of Area Under the Curve for the different models.

Models*	Value Positive Predictive	Negative Predictive Value	False Positives	Area Under the Curve
glm	12	98	1	0.841
glmup	92	71	25	0.846
glmnetup	92	72	25	0.837
glmglmnetup	88	75	21	0.840

*glm: logistic regression on original data, glmup: logistic regression up-sample data,
glmnetup: LASSO regression on up-sample data,
glmglmnetup: Logistic regression with selected LASSO variables on up-sample data.

The comparison of the distribution of the PTB risk estimated by each model to the original data (Fig. 2), showed that logistic regression before up-sampling (glm) and the Lasso model (glmnet) generally underestimate the probabilities in comparison to the other models. Even the last logistic model using the lasso selected variables slightly under-estimated those probabilities. However, both logistic regression with up-sampling before or after lasso regularization gave a closer risk or probability distribution to the original data than the other models (Fig. 2).

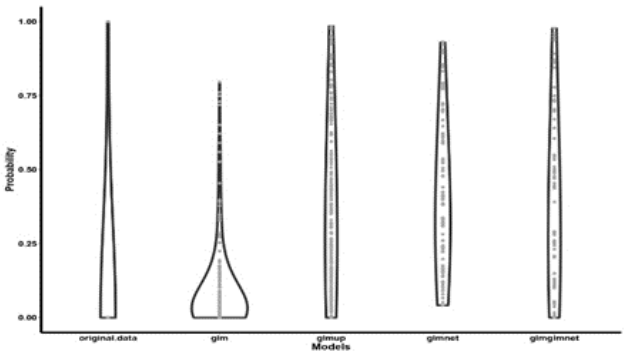

Figure2. Predicted probability distribution for each model (glm: logistic regression on original data, glmup: logistic regression up-sample data,glmnetup: LASSO regression on up-sample data and glmglmnetup: Logistic regression with selected LASSO variables on up-sample data)

.

Along with the improvement of preterm prediction the number of statistically significant covariates (at least at the level 5%) also increased from 5 for glm, to 10 in glmup but the glmnetup reduced this number to 6 (Table II). The regression model using the selected Lasso variables (glmglmnetup) was used to develop a nomogram (Fig. 3). The validation of this nomogram using the data of this study showed the possibility of having a reasonably accurate risk of PTB given the levels of Social status, Residence, Pre-hemorrhage, Age, BMI, and Weight gain for a multiparous woman.

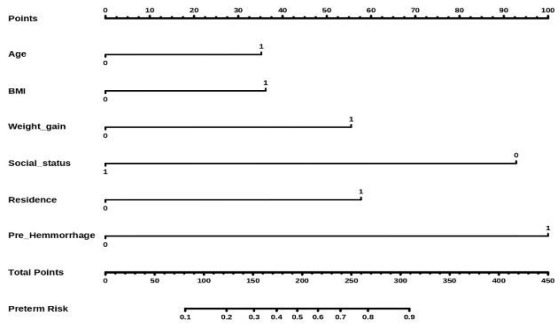

Figure 3. Nomogram for the screening of nulliparous women at risk of preterm birth.

IV. Discussion

The results of this work improved detection of high risk PTB among multiparous women using routinely collected social, demographic, and health indicators. The model that led to the best result of 88% PTB correctly predicted along with the lowest number of false positives, was used to draw a graphical nomogram that could be easily used by physicians to screen for high-risk PTB. Nevertheless, the physicians will need to put on stricter medical surveillance about 21% (at risk of PTB + false positives) of the total number of multiparous women.

In comparison Mehta-Lee et al. (2016) [10] have reported a significantly lower prediction of 51.5 % for PPV vs 88% in our study and 76.7 % for NPV versus 75% in this work. To achieve this improved level of PTB prediction, data augmentation of the initial sample through up-sampling algorithms was used. Hence, it is probable that the low PTB prediction of the logistic regression model based on the original data was at least partially due to the low prevalence of preterm birth. Furthermore, using logistic regression to predict low prevalence events may lead to meaningless outcomes [17]. Data augmentation by up-sampling randomly increases the number of positive preterm birth profiles in the newly generated dataset without changing the other class comprising women not presenting PTB [18]. This technique has been successfully used in investigations with low or very low prevalence, including some machine learning techniques such as convolutional neural networks [9].

The logistic regression model on low prevalence data clearly under-estimates the general probability [19]. A similar phenomenon was also observed for the Lasso based model, albeit with significantly smaller under-estimation. Furthermore, the regressions on up-sampled data included a higher number of significant variables to explain the model.

The number of significant variables by logistic regression almost exactly corresponded to the variables selected by Lasso regularization. However, the final model using the 6 selected variables from Lasso regularization decreased the number of false positives and hence gave the best results for PTB prediction.

The most statistically significant covariates that seem to affect PTB in this study were Social status, Pre-hemorrhage, Residence, Weight gain, BMI, and Age. Hence, it seems that the possibility of access to adequate medical care through a high income, residing in the city, and avoiding weight problems are key factors to decrease PTB incidence for this group of multiparous women. Nevertheless, urban women presented a slightly higher PTB risk. In China, a similar result has also been reported with higher PTB risk in urban areas [20] related to a higher stress and anxiety. Indicators of excess weight in terms of BMI or during pregnancy weight gain especially coupled to older pregnancy age correlated well with higher preterm risk. These last factors have been identified in other investigations [21, 22] as risk factors for PTB. It is noteworthy that besides the social status, the high incidence of hemorrhage in this group of women was relatively high. Indeed, 29 % of the multiparous women presented hemorrhage during their pregnancy. This incidence is higher even in comparison to some countries of lower national income [23] led to the highest adjusted odds ratio for PTB of 6.88 to 10.24 (95% interval).

Despite showing promising results of PTB prediction in multiparous women in Northern Lebanon, this study presents many limitations. It would be improved with a higher number of women in the sample. On top of the low number of cases, the sample was fairly homogeneous because data are better kept in hospitals treating a bigger number of high social status patients. We are hoping that this type of work will encourage health authorities to establish public databases on births in this type of low to middle-income countries. Pre-eclampsia and Diabetes were not used in the models because of the very low prevalence affecting the interpretation of the models. More variables could be added such as the number of children, stressful work, anxiety, and planned pregnancies among others. Measurements such as cervical length and cervicovaginal fetal fibronectin should be added in the screening model or at least carried out on the group of screened women by the nomogram.

V. CONCLUSION

Using readily available information from past pregnancy along with social and weight indicators, we developed a nomogram that can be used to screen for PTB risk in multiparous women. The nomogram uses the binary response to six covariates including Social status, Pre-hemorrhage, Residence, Weight gain, BMI, and Age. The nomogram could identify about 88% of the high PTB risk women.

In order to achieve a reasonably high prediction for PTB, the logistic regression was trained on a data augmented sample using up sampling. LASSO penalization helped select the final covariates in the model. These methods could improve analysis and prediction of diseases or health complications that present low or very low prevalence.

ACKNOWLEDGMENTS

The authors thank the Hospitals that helped us collect the data for this work.

References

[1] World Health Organization. “Born too soon: the global action report on preterm birth”, 2012.

[2] S. Chawanpaiboon, J.P. Vogel, A.-B. Moller, P. Lumbiganon, M. Petzold, D. Hogan, S. Landoulsi, N. Jampathong, K. Kongwattanakul, M. Laopaiboon, et al. “Global, regional, and national estimates of levels of preterm birth in 2014, a systematic review and modelling analysis”, Lancet Glob. Health 7: e37–e46. 2018. Available: https://pubmed.ncbi.nlm.nih.gov/30389451/.

[3] J. Katz, A.C. Lee, A, N. Kozuki, J.E. Lawn, S. Cousens, H. Blencowe, M. Ezzati, Z. A. Bhutta, T. Marchant, B.A.Willey, L. Adair, F. Barros, A.H. Baqui, P. Christian, W. Fawzi, R. Gonzalez, J. Humphrey, L. Huybregts, P. Kolsteren,, A. Mongkolchati,, CHERG. “Mortality risk in preterm and small-for-gestational-age infants in low-income and middle-income countries: a pooled country analysis”, Lancet (London, England), 382(9890), 417–425, 2013.
Available: https://pubmed.ncbi.nlm.nih.gov/23746775/.

[4] G. U. Eleje, J. I. Ikechebelu, A. C. Eke, P. C. Okam, I. U. Ezebialu, & C. P. Ilika, “Cervical cerclage in combination with other treatments for preventing preterm birth in singleton pregnancies”, The Cochrane Database of Systematic Reviews, (11)2017.
Available:
https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD012871.pub2/full.

[5] Z. A. Oskovi Kaplan, & A. S. Ozgu Erdinc, “Prediction of Preterm Birth: Maternal Characteristics, Ultrasound Markers, and Biomarkers: An Updated Overview”, Journal of Pregnancy, 1-8, 2018.
Available:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6199875/.

[6] L. J. E. Meertens, P. van Montfort, H. C. J. Scheepers, S. M. J. van Kuijk, R. Aardenburg, J. Langenveld, I. M. A. van Dooren, I. M. Zwaan, M. E. A. Spaanderman, L. J. M. Smits. “Prediction models for the risk of spontaneous preterm birth based on maternal characteristics: a systematic review and independent external validation”, Acta Obstet.Gynecol Scand ;97(8):907-920, Epub 9, PMID: 29663314; PMCID: PMC6099449, 2018.
Available: https://pubmed.ncbi.nlm.nih.gov/29663314/.

[7] R. L. Goldenberg, J. F. Culhane, J. F. Iams, R. Romero, “Epidemiology and Causes of Preterm Birth”, Lancet 371 :75-84, 2018.
Available: https://pubmed.ncbi.nlm.nih.gov/18177778/.

[8] C. E. Kleinrouweler, F. M. Cheong-See, G. S. Collins, A. Kwee, S. Thangaratinam, K. S. Khan, B. W. Mol, E. Pajkrt, K. G. Moons, E. Schuit. “Prognostic models in obstetrics:

available, but far from applicable", Am J Obstet. Gynecol, 214(1):79-90, e36, 2016.
Available: https://pubmed.ncbi.nlm.nih.gov/26070707/
[9] T. Włodarczyk, S. Płotka, P. Rokita, N. Sochacki-Wójcicka, J. Wójcicki, M. Lipa, T. Trzciński. "Spontaneous Preterm Birth Prediction Using Convolutional Neural Networks", In: Hu Y. et al. (eds) "Medical Ultrasound, and Preterm, Perinatal and Pediatric Image Analysis", vol 12437. Springer, Cham. Lecture Notes in Computer Science ASMUS 2020, PIPPI 2020.
Available: https://link.springer.com/chapter/10.1007/978-3-030-60334-2_27.
[10] S. S. Mehta-Lee, A. Palma, P. S. Bernstein *et al.* "A Preconception Nomogram to Predict Preterm Delivery", Matern Child Health *J,* **21,** 118–127, 2017. Available: https://pubmed.ncbi.nlm.nih.gov/27461021/.
[11] B. Koullali, M. D. van Zijl, B. M. Kazemier *et al* "The association between parity and spontaneous preterm birth: a population-based study", BMC Pregnancy Childbirth, 20, 233, 2020. Available: https://pubmed.ncbi.nlm.nih.gov/32316915/.
[12] M. Chabachib, R. H. Kusmaningrum, H. Hersugondo, I. D. Pamungkas, "Financial Distress Prediction in Indonesia, WSEAS Transactions on Business and Economics", ISSN / E-ISSN: 1109-9526 / 2224-2899, Volume 16, Art. #28, pp. 251-260, 2019.
Available:
https://www.wseas.org/multimedia/journals/economics/2019/a505107-730.php.
[13] Y. Alsaawy, A. Alkhodre, M. Benaida, R. A. Khan, "A Comparative Study of Multiple Regression Analysis and Back Propagation Neural Network Approaches on Predicting Financial Strength of Banks: An Indian Perspective, WSEAS Transactions on Business and Economics", ISSN / E-ISSN: 1109-9526 / 2224-2899, Volume 17, Art. #60, pp. 627-637, 2020.
Available:
https://www.wseas.org/multimedia/journals/economics/2020/b225107-978.pdf.
[14] World Health Organization (WHO), "Global Strategy on Diet, Physical Activity and Health", Cited 2020.
[15] D. Koniak-Griffin & C. Turner-Pluta, "Health risks and psychosocial outcomes of early childbearing: a review of the literature", The Journal of perinatal & neonatal nursing, 15(2), 1-17, 2001.
Available: https://pubmed.ncbi.nlm.nih.gov/12095025/.
[16] M. Jolly, N. Sebire, J. Harris, S. Robinson, L. Regan. "The risks associated with pregnancy in women aged 35 years or older", Human Reproduction, Volume 15, Issue 11, Pages 2433–2437, 2000.
Available: https://pubmed.ncbi.nlm.nih.gov/11056148/.
[17] S. Doerken , M. Avalos, E. Lagarde, M. Schumacher, "Penalized logistic regression with low prevalence exposures beyond high dimensional settings", PLOS ONE, 14(5): e0217057,2019.
Available:
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0217057.
[18] G. Cheng, S. Osmundson, D. R. Velez Edwards, G. Purcell Jackson, B. A. Malin, Y. Chen, "Deep learning predicts extreme preterm birth from electronic health records", Journal of Biomedical Informatics Volume 100, 103334, ISSN 1532-0464,2019.
Available: https://pubmed.ncbi.nlm.nih.gov/31678588/.
[19] G. Francesco, M. Niglio & M. Restaino. "A new procedure for variable selection in presence of rare events". Journal of the Operational Research Society, 2020. Available: https://www.tandfonline.com/doi/abs/10.1080/01605682.2020.1740620.
[20] L. Li, J. Ma, Y. Cheng, *et al.* "Urban–rural disparity in the relationship between ambient air pollution and preterm birth", Int J Health Geogr 19, 23 2020. Available: https://ij-healthgeographics.biomedcentral.com/articles/10.1186/s12942-020-00218-0.
[21] S.W Masho, D.L. Bishop & M. Munn, "Pre-pregnancy BMI and weight gain: where is the tipping point for preterm birth?", BMC Pregnancy Childbirth, 13, 120, 2013. Available: https://pubmed.ncbi.nlm.nih.gov/23706121/.
[22] F. Fuchs, B. Monet, T. Ducruet., N. Chaillet & F. Audibert, "Effect of maternal age on the risk of preterm birth: A large cohort study", *PloS one* 2018, *13*(1), e0191002, 2018. Available:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5791955/.
[23] B. A Kebede, R. A. Abdo, A. A Anshebo & B. M Gebremariam, "Prevalence and predictors of primary postpartum hemorrhage: An implication for designing effective intervention at selected hospitals, Southern Ethiopia", *PloS one*, *14*(10), e0224579, 2019. Available: https://pubmed.ncbi.nlm.nih.gov/31671143/.

Contribution of individual authors to the creation of a scientific article (ghostwriting policy)

Mrs. Traboulsi Mayssa: is a Ph.D. candidate that collected the data, participated in the design and write up of this work.

Pr. Zainab E. El Alaoui- Talibi: is the Ph.D. main advisor, participated in the design and write up of this work.

Pr. Boussaid Abdellatif: is the Ph.D. co-advisor, participated in the design and write up of this work. Executed and helped in the interpretation of the statistical analyses.

Data from this study will be available on request to the corresponding author:
Mayssa A. Traboulsi, E-mail : Mayssatr@gmail.com

Sources of funding for research presented in a scientific article or scientific article itself

No special funds were used in this study.

Competing interests

The authors declare that they don't have any conflict of interest regarding the data published in this work

List of abbreviations

PTB: Pre-Term Birth
AUC: Area Under the Curve
BMI: Body Mass Index
LASSO: Least Absolute Shrinkage and Selection Operator