



Prediction of Failure of Induction of Labor from Ultrasound Images Using Radiomic Features

María Inmaculada García Ocaña^{1,2(✉)}, Karen López-Linares Román^{1,2}, Jorge Burgos San Cristóbal³, Ana del Campo Real³, and Iván Macía Oliver^{1,2}

¹ Vicomtech, San Sebastián, Spain

{igarcia,klopez,imacia}@vicomtech.org

² Biodonostia Health Research Institute, San Sebastián, Spain

³ Obstetrics and Gynecology Service, Biocruces Bizkaia Health Research Institute, Cruces University Hospital, Osakidetza, UPV/EHU, Baracaldo, Spain

Abstract. Induction of labor (IOL) is a very common procedure in current obstetrics; about 20% of women who undergo IOL at term pregnancy end up needing a cesarean section (C-section). The standard method to assess the risk of C-section, known as Bishop Score, is subjective and inconsistent. Thus, in this paper a novel method to predict the failure of IOL is presented, based on the analysis of B-mode transvaginal ultrasound (US) images. Advanced radiomic analyses from these images are combined with sonographic measurements (e.g. cervical length, cervical angle) and clinical data from a total of 182 patients to generate the predictive model. Different machine learning methods are compared, achieving a maximum AUC of 0.75, with 69% sensitivity and 71% specificity when using a Random Forest classifier. These preliminary results suggest that features obtained from US images can be used to estimate the risk of IOL failure, providing the practitioners with an objective method to choose the most personalized treatment for each patient.

Keywords: Radiomics · Ultrasound · Induction of labor · Machine learning

1 Introduction

Induction of labor (IOL) is a very common procedure in current obstetrics; according to the American College of Obstetricians and Gynecologists, between 20% and 40% of births are induced. IOL is the treatment that stimulates child-birth and delivery. About 20% of women who undergo IOL at term pregnancy end up needing a C-section, mainly due to the failure of induction, failure of progression of labor or fetal distress.

Bishop Score is the most widely used method for the assessment of cervical tissue properties and aims at determining the readiness of the cervix for IOL. However, it is a subjective measure and has been found to be inconsistent [9].

© Springer Nature Switzerland AG 2019

Q. Wang et al. (Eds.): PIPPI 2019/SUSI 2019, LNCS 11798, pp. 153–160, 2019.

https://doi.org/10.1007/978-3-030-32875-7_17

Thus, proposing a method for the proper selection of candidates for successful IOL is an open issue in obstetric practice.

During pregnancy and delivery, the cervix transforms from a stiff, long and closed structure to a soft, short and dilated structure that allows delivery. While collagen is aligned and organized in the cervix of non-pregnant women, it is more disorganized during the remodeling of the cervix during pregnancy. Water content of the cervical tissues is also increased in the process of preparation for delivery. All these changes are expected to be reflected in the image obtained from a transvaginal ultrasound (US), since the consistency of tissues affects their interaction with US waves.

Therefore, an analysis of image features extracted from US images could reveal the cervical tissue properties before IOL, even when they are not apparent to a human observer. This idea has been applied to study the neonatal respiratory morbidity from fetal lung US [5], to assess the cervical structure in spontaneous preterm births [2] or to predict the fetus gestational age [3]. In [10,11], US image analysis is also used to predict failure of induction of labor [10,11]. In [10], local binary patterns were used to extract texture features from the image, while in [11], symmetric local binary patterns and Gabor filterbanks were used.

The aim of this study is to analyze the predictive value of radiomic features extracted from transvaginal US images to predict IOL failure, and to compare their performance against other sonographical features studied in the literature, such as cervical length and cervical angle [1,4,8], and clinical data. To the best of our knowledge, this is the first study that uses radiomics, in the sense of a large amount of imaging features, to predict IOL failure, since previous works were limited to a reduced set of texture features [10,11]. Furthermore, clinical data is included as complementary information to the radiomics to build a predictive model. Different combinations of imaging and clinical data and different machine learning classifiers are explored, and an extensive comparison of the results is provided.

2 Materials and Methods

The following subsections describe the employed data, including the imaging and clinical data available for each patient, as well as the extracted radiomic features, the experiments and the proposed machine learning classifiers used to build the predictive models.

2.1 Dataset Annotation

The database used in this study consists of images and clinical data from patients admitted for IOL at Cruces University Hospital (Bilbao, Spain). The patients underwent a transvaginal US before IOL. Images were acquired with a Voluson ultrasound scanner from General Electric by an expert obstetrician following the same protocol for all patients. All images were provided in DICOM format. Image resolution is 720×960 pixels with a pixel spacing of 0.11. An expert



Fig. 1. Transvaginal ultrasound images of three patients and the selected region of interest for radiomic analysis.

obstetrician manually selected a region of interest (ROI) delimiting the upper part of the cervix, which is thought to have the most relevant information and less noise for the analysis. Figure 1 shows examples of the input images and ROIs.

Data from a total of 182 patients with US images, annotation of the ROI and clinical data was available, from which 130 had a vaginal delivery and 52 needed a C-section. Only in 30 cases the cause of the C-section was related to a cervical motive. Figure 2 summarizes the database composition.

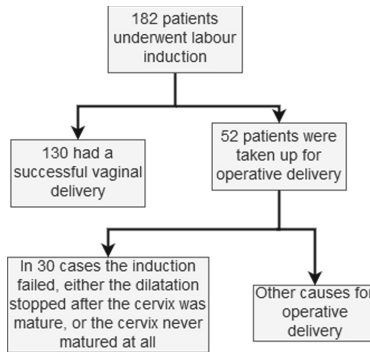


Fig. 2. Database composition.

Twenty relevant clinical attributes were selected from the database to be included in the study (such as age, weight, height, race, body mass index, number of abortions, weeks of gestation, information about previous pregnancies). Seven sonographic measurements, manually extracted from the transvaginal US, were also included. Sonographic features consist of different measures of the cervical anatomy: basal cervical length, compressed cervical length, basal anterior-posterior diameter, compressed anterior-posterior diameter, basal lateral diameter, compressed posterior diameter, compressed lateral diameter and segment.

2.2 Radiomic Feature Extraction

Radiomic features were extracted from the selected ROI using the PyRadiomics [7] Python software package. First order intensity-based features and texture-based features are included. To measure image texture, the following four matrices are calculated, from which descriptive values are computed:

Gray Level Co-occurrence Matrix (GLCM): GLCM describes the second-order joint probability function of an image region constrained by the mask. Each $(i, j)_{th}$ element of this matrix represents the number of times the combination of levels i and j occur in two pixels in the image that are separated by a given distance of pixels along a certain angle. We chose a distance of one pixel (angles are computed automatically).

Gray Level Run Length Matrix (GLRLM): GLRLM quantifies gray level runs, which are defined as the length in number of pixels of consecutive pixels that have the same gray value. In a GLRLM the $(i, j)_{th}$ element describes the number of runs with gray level i and length j that occur in the ROI along angle θ . The distance is 1 and angles are again computed automatically.

Gray Level Dependence Matrix (GLDM): GLDM quantifies gray level dependencies in an image. A gray level dependency is defined as a the number of connected voxels within a given distance that are dependent on the center voxel. A neighboring voxel with gray level j is considered dependent on center voxel with gray level i if $|i - j| \leq \alpha$. In a GLDM the $(i, j)_{th}$ element describes the number of times a voxel with gray level i with j dependent voxels in its neighborhood appears in image. The parameters used are distance = 1 and $\alpha = 0$.

Neighbouring Gray Tone Difference Matrix (NGTDM): it quantifies the difference between a gray value and the average gray value of its neighbours within a given distance (1 in this study).

By extracting mathematical descriptors from these matrices (mean, variance, entropy, uniformity ...), a total of 58 features were obtained. Moreover, 19 first-order features based on image intensity are included (energy, entropy, minimum, maximum, mean, median, interquartile range, skewness, kurtosis ...), resulting in a vector of 77 features.

2.3 Experimental Settings

Four experiments are proposed in order to find the best model for IOL failure prediction. In each experiment the following features are employed:

- **Experiment 1:** clinical data (20 features).
- **Experiment 2:** sonographic measurements (7 features).
- **Experiment 3:** sonographic measurements plus clinical data (27 features).
- **Experiment 4:** sonographic measurements, clinical data and radiomic features (104 features).

For some experiments, the feature vector is too long compared to the number of instances in the database, which can lead to poor performance of machine learning classifiers. Therefore, a filter method is applied to rank the features and select only the most relevant ones.

An additional problem when training machine learning classifiers for this task is class imbalance, which can lead the classifiers to have a bias towards the majority class. We used the *Synthetic Minority Over-sampling Technique-SMOTE* [6] to generate synthetic samples from the minority class (C-section deliveries), which are used to train the classifiers.

2.4 Machine Learning Classifiers

The following machine learning classifiers (from the Python *scikit learn* [12] library) are trained and validated to compare their performance: Gaussian Naive Bayes (GNB), Random Forest (RF), Multi Layer Perceptron (MLP), Support Vector Machine (SVM), Decision Tree (DT) and Extra Tree (ET). Table 1 summarizes the parameters used to train each classifier. Furthermore, to provide robustness to the results, we used a 10-fold cross-validation approach.

Table 1. Relevant meta-parameters for the classifiers used in the study. GNB: Gaussian Naive Bayes, RF: Random Forest, MLP: Multi Layer Perceptron, SVM: Support Vector Machine, DT: Decision Tree, ET: Extra Tree.

Classifier	Parameters
GNB	Variance smoothing: 1e-9
RF	Number of estimators: 150, impurity: Gini; minimum samples per split: 2
MLP	Maximum number of iterations: 1000; number of layers: 100
SVM	Kernel: radial basis function
DT	Impurity: Gini; minimum samples per split: 2
ET	Number of estimators: 150; minimum samples per split: 2; impurity: Gini

3 Results and Discussion

Table 2 summarizes the results obtained for each of the experiments described in Sect. 2.3. Clinical data alone (experiment 1) does not seem to have high predictive value, while the predictive value of the sonographical measurements alone (cervical length, cervical angle ...) is shown to be better, with a maximum AUC of 0.682 ± 0.009 when using a GNB classifier. This is consistent with previous studies in the literature which have shown a correlation between cervical length and other anatomical measurements and the outcome of IOL [8].

Regarding experiment 3, which combined clinical data and sonographic measurements, and experiment 4, including radiomic features, similar maximum AUC values are achieved. The RF and ET classifiers perform the best in both

Table 2. Mean AUC (area under the ROC, Receiver Operating Characteristic, curve) and standard deviation for the 10 k-fold cross validation for every classifier: GNB: Gaussian Naive Bayes, RF: Random Forest, MLP: Multi Layer Perceptron, SVM: Support Vector Machine, DT: Decision Tree, ET: Extra Tree. Results for all the experiments are shown.

Classifier	Experiment 1		Experiment 2		Experiment 3		Experiment 4	
	AUC	Std	AUC	Std	AUC	Std	AUC	Std
GNB	0.622	0.029	0.682	0.009	0.705	0.010	0.501	0.026
RF	0.621	0.014	0.652	0.012	0.763	0.011	0.750	0.017
MLP	0.512	0.058	0.678	0.012	0.582	0.040	0.505	0.051
SVM (RBF)	0.515	0.035	0.506	0.029	0.480	0.046	0.500	0
DT	0.572	0.030	0.525	0.037	0.637	0.018	0.62	0.031
ET	0.614	0.020	0.639	0.016	0.747	0.012	0.769	0.014

cases, yielding AUC values between 0.747 and 0.769. When comparing the highest AUC values from both experiments (experiment 3, random forest: 0.763 ± 0.011 ; experiment 4, extra trees: 0.769 ± 0.014) the obtained difference is not statistically significant (T-test, p value = 0.1934).

Nonetheless, differences in the performance of the classifiers can be seen by analyzing their behaviour with respect to false negatives and false positives. A more detailed analysis of the results is provided in Table 3, which shows the sensitivity, specificity, false positive rate and false negative rate for the classifiers that yielded the best AUCs for each experiment. Overall, false negative rate is higher than false positive rate. From a clinical point of view, this situation (sending a patient for IOL that ends up in C-section) is better than the opposite (performing a C-section when IOL would have succeeded).

Thus, comparing the models from experiment 4 with experiment 3, it can be observed that adding radiomic features increases the sensitivity and, consequently, reduces the false negative rate of the classifiers. We obtain useful information from the radiomic analysis that makes the model more balanced, with a sensitivity of 69% and a specificity of 71%, in contrast to the rest of models which achieve good specificity but have too low sensitivity. This means that radiomics contain relevant information and may help avoiding unnecessary C-sections.

It is worth noting that the same feature selection procedure was applied to experiment 3 and experiment 4, to make the results comparable. The feature selection process allows us to understand which features are more important. For experiment 3 many of the high-ranked features are sonographical (5 features), but a few clinical data are kept as well: weight, height, age, estimated fetal weight, number of previous pregnancies, previous vaginal births, weeks of pregnancy. While using only clinical data is not enough to predict IOL (experiment 1), adding this information to sonographical measurements helps improving the models, according to the results from experiments 2 and 3. In experiment 4, five

Table 3. Sensitivity, specificity, false positive rate (FPR) and false negative rate (FNR) for the best performing classifiers for every experiment

	Data	Classifier	Sensitivity	Specificity	FPR	FNR
Exp. 1	Clinical data	RF	0.808	0.338	0.662	0.192
		ET	0.423	0.885	0.223	0.635
Exp. 2	Sonographic measurements	GNB	0.632	0.682	0.318	0.368
		MLP	0.600	0.677	0.323	0.400
Exp. 3	Clinical data and sonographic measurements	RF	0.596	0.825	0.175	0.404
		ET	0.451	0.858	0.142	0.549
Exp. 4	Clinical data, sonographic measurements and radiomics	RF	0.692	0.715	0.285	0.308
		ET	0.481	0.846	0.154	0.519

sonographic measurements are again selected, as well as some clinical variables (height, estimated fetal weight, body mass index, number of vaginal births) and 4 radiomic values (energy, long run high gray level emphasis, run length non uniformity, run entropy). The fact that radiomic features are high-ranked in the feature selection process and that the results in experiment 4 are better than in experiment 3 suggests that radiomic features from transvaginal US can be useful for the prediction of IOL failure.

4 Conclusion

Correctly evaluating the probability of successful IOL is still an open issue in modern obstetrics, since 20% of the induced women have a C-section and the current evaluation method, Bishop Score, has been found to be subjective and inconsistent. Transvaginal US is cheap and widely available at hospitals, and it is performed routinely in other stages of pregnancy. The results presented in this paper agree with previously reported results [4,8] in that cervical length and cervical angle measured from the transvaginal US are useful for the prediction of IOL failure (0.682 AUC, 63% sensitivity, 68% specificity).

Furthermore, a novel methodology for IOL failure prediction based on radiomics has been applied to these images for the first time. We have shown how a combination of radiomic features with cervical measurements and clinical data can be used to build a predictive model that achieves an AUC of 0.75 with 69% sensitivity and 71% specificity. These preliminary results indicate that US can provide the clinicians with useful information prior to the IOL.

An important limitation of our study is the size of the patient cohort, with 182 patients from which only 52 had a C-section. Furthermore, only in 30 cases the IOL failure was related to a cervical motive. All the images come from the same hospital and have been acquired following the same protocol with US devices from the same vendor. Poor generalization is a common problem

working with radiomics, as different protocols or vendors could result in different image properties, which implies that the selected features and models could be overfitted for the current available data. Further validation should be performed with a larger and more diverse database to assess the robustness of the proposed method. Future works should also develop a technique to obtain measurements as cervical length and cervical angle automatically from the US images.

References

1. Al-Adwy, A.M., et al.: Diagnostic accuracy of posterior cervical angle and cervical length in the prediction of successful induction of labor. *Int. J. Gynecol. Obstet.* **141**(1), 102–107 (2018)
2. Baños, N., Perez-Moreno, A., Julià, C., Murillo-Bravo, C., Coronado, D., Gratacós, E., Deprest, J., Palacio, M.: Quantitative analysis of cervical texture by ultrasound in mid-pregnancy and association with spontaneous preterm birth: cervical texture associated with spontaneous preterm birth. *Ultrasound Obstet. Gynecol.* **51**(5), 637–643 (2018)
3. Baños, N., et al.: Quantitative analysis of the cervical texture by ultrasound and correlation with gestational age. *Fetal Diagn. Ther.* **41**(4), 265–272 (2017)
4. Brik, M., Mateos, S., Fernandez-Buhigas, I., Garbayo, P., Costa, G., Santacruz, B.: Sonographical predictive markers of failure of induction of labour in term pregnancy. *J. Obstet. Gynaecol.* **37**(2), 179–184 (2017)
5. Burgos-Artizzu, X.P., Perez-Moreno, A., Coronado-Gutierrez, D., Gratacos, E., Palacio, M.: Evaluation of an improved tool for non-invasive prediction of neonatal respiratory morbidity based on fully automated fetal lung ultrasound analysis. *Sci. Rep.* **9**(1), 1950 (2019)
6. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002)
7. van Griethuysen, J.J., et al.: Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* **77**(21), 104–107 (2017)
8. Kehila, M., Abouda, H., Sahbi, K., Cheour, H., Chanoufi, M.B.: Ultrasound cervical length measurement in prediction of labor induction outcome. *NPM* **9**(2), 127–131 (2016)
9. Kolkman, D., et al.: The bishop score as a predictor of labor induction success: a systematic review. *Am. J. Perinatol.* **30**(08), 625–630 (2013)
10. Obando, V.P., Arana, A.N., Izaguirre, A., Burgos, J.: Labor induction failure prediction based on B-mode ultrasound image processing using multiscale local binary patterns. In: *International Conference on Optoelectronics and Image Processing*, pp. 25–29 (2016)
11. Obando, V.P., Arana, A.N., Izaguirre, A., Burgos, J.: Labor induction failure prediction using Gabor filterbanks and center symmetric local binary patterns. In: *IEEE 37th Central America and Panama Convention*, pp. 1–5 (2017)
12. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)