UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA

E.T.S. DE INGENIERÍA INFORMÁTICA

Analysis of ultrasound images and clinical data in two clinical scenarios: prediction of failure of induction of labor and risk of preterm delivery

Author:
María Inmaculada García Ocaña

Supervisors:
Karen López-Linares Román
Francisco Javier Diez

MÁSTER UNIVERSITARIO EN I.A. AVANZADA: FUNDAMENTOS,MÉTODOS Y APLICACIONES

September 15, 2019

Abstract

This work explores the use of radiomics and machine learning to extract relevant biomarkers from ultrasound (US) images that can be used in obstetric practice. Two clinical applications are studied: the prediction of induction of labor (IOL) failure based on clinical data and US images obtained prior to IOL, and the estimation of risk of preterm birth based on routinely US images acquired in the 20th week of pregnancy. Several machine learning classifiers and feature selection techniques are tested and the results are compared.

The best model for the prediction of IOL failure was a random forest that model obtained an AUC of 0.75, with 69% sensitivity and 71% specificity. The best model for the prediction of preterm birth was a random forest that obtained an AUC of 0.77 AUC, with 71% sensitivity and 69% specificity .

These preliminary results suggest that features obtained from US images can be used to estimate risks in these two obstetric problems. Transvaginal US is cheap, widely available at hospitals, and performed routinely. Therefore these method can be easily implemented in clinical practice and help practitioners choose a most personalized treatment for each patient, improving the outcomes. Further validation with a largest and more diverse dataset is needed, especially to assess how the image analysis methods work with images from different US vendors.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Medical ultrasound (US) images are formed using an US probe to transmit mechanical wave pulses into tissue, which generates sound echoes at boundaries where different tissues exhibit acoustic impedance differences. The resulting image quality depends on the force exerted on the US transducer and the transducer location and orientation, and typical artifacts in US images include signal dropout, attenuation, speckle and shadows.

Different types of images can be formed using US equipment; the most commonly used in clinical practice is B-mode, which displays the acoustic impedance of a two-dimensional cross section of the tissue. Other types of US images are A-mode (amplitude), M-mode (motion over time) and D-mode (Doppler). Figure 1.1 shows examples of different types of US images.

US images are performed routinely in obstetrics to monitor pregnancy. Usually, they are inspected visually by the practitioners and stored in the hospital's PACS (Picture Archiving and Communication System). This process is qualitative, highly subjective. Furthermore, the quality of the images and the findings have high inter and intra observer variability.

In recent years, quantitative analysis of US images has been proposed. Several computer vision techniques have been used for the automatic segmentation, classification or registration of US images, sometimes incorporating artificial intelligence to the process. Machine learning and deep learning, which have been widely applied to other medical imaging modalities, such as Computer Tomography (CT) or Magnetic Resonance Imaging (MRI), have also been used for the automatic analysis of US. However, US imaging poses specific challenges and US image analysis techniques lag behind the other modalities [1].

Radiomics is a method that extracts a large amount of features from radiographic medical images using data-characterisation algorithms [2]. The successful application of radiomics to other imaging modalities has motivated studies using similar techniques for US image analysis.

In this project, we analyze B-mode US images. The process consists in extracting features from the image and then applying feature selection methods to find relevant those that can be used as imaging biomarkers with predictive value regarding a specific pathology, therefore providing an objective and quantitative way to extract information from the image. We also use convolutional neural networks (CNN) to

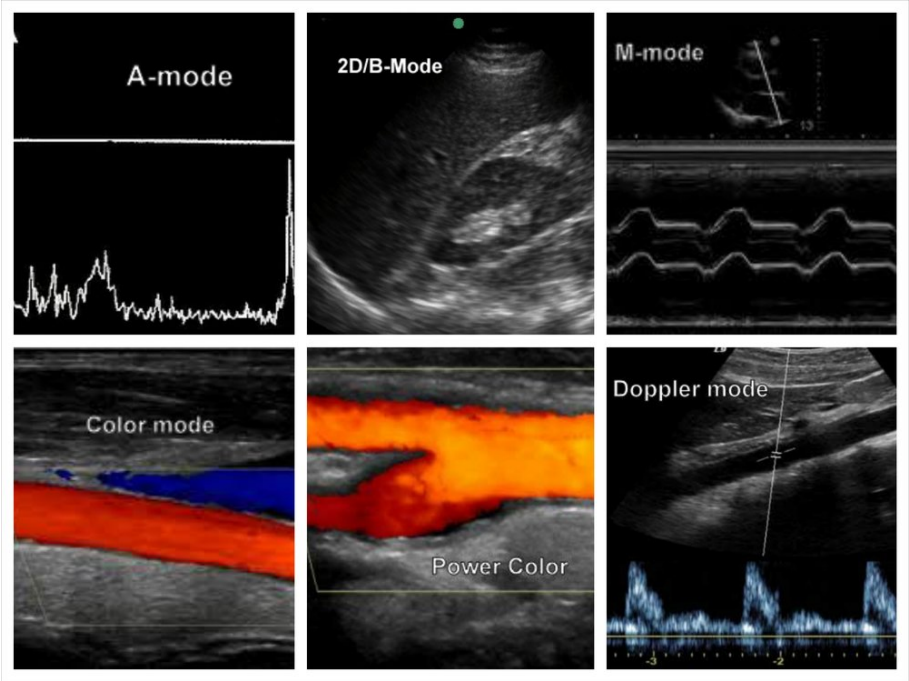

Figure 1.1: Example of images adquired with different US modes

extract features and compare the results.

We analyze two different clinical scenarios as different use cases of the same technology: prediction of induction of labor failure and prediction of preterm birth.

Induction of labor (IOL) is is the treatment that stimulates childbirth delivery. It is a very common procedure in current obstetrics; according to the American College of Obstetricians and Gynecologists (ACOG), between 20 and 40 percent of births are induced. About 20% of women who undergo IOL at term pregnancy end up needing a cesarean section (C-section), mainly due to the failure of induction, failure of progression of labor or fetal distress[3].

Bishop's Score has been the most widely used technique for the assessment of the cervical tissue prior to IOL. However, it is a subjective measure and has been found not to be consistent [4]. Inducing labor can be accomplished with pharmaceutical or non-pharmaceutical methods, including a mechanical instruments that promote cervical ripening and the onset of labor by stretching the cervix. Thus, a criterion for selecting of candidates for IOL as well as the most adequate IOL method is an open issue in obstetric practice.

During pregnancy and delivery, the cervix undergoes several changes. It transforms from a stiff, long and closed structure to a soft, short and dilated structure that allows delivery. Regarding the micro-structure of the tissue, collagen is aligned and organized in the cervix of non-pregnant women and more disorganized during the remodeling of the cervix. Water content of the cervical tissues is also increased in the process of preparation for delivery.

Changes in the cervical micro-structure and water content in the tissue are expected to be reflected in the image obtained from US since the consistency of tissues affect their interaction with US waves. A radiomic analysis could reveal these

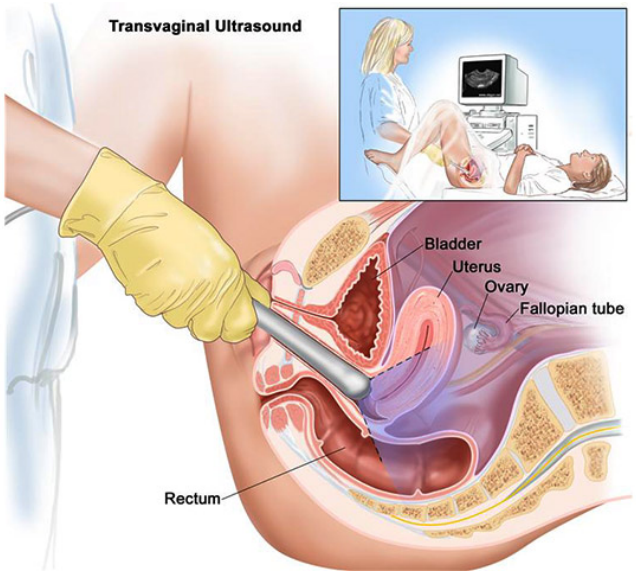

Figure 1.2: Acquisition of transvaginal US

changes, even when they are not apparent to a human observer. Transvaginal US (see Figure 1.2) allows obtaining an image of the cervix, which can be subsequently analyzed. Images of this type are already performed routinely during pregnancy to evaluate the state of the cervix, find complications, such as ectopic pregnancy or anomalies (cysts, fibrosis), and assess the risk of preterm birth, but are not yet included in the clinical guidelines for evaluation before IOL yet. However, several measurements that can be obtained from the transvaginal US have been correlated with the outcome of induction of labor[5, 6, 7], and some works have shown interesting preliminary results for analysis of texture as well [8, 9].

Preterm birth is the delivery before 37 weeks of pregnancy. About 15 millions of babies are affected by preterm birth complications every year, and almost 1 million die [10]. Children who survive a preterm birth can suffer life-long disabilities, learning problems and hearing and visual problems. Prediction of preterm birth is usually done by measuring the cervical length in the transvaginal US (Figure 3.4 shows the different anatomical structures that can be observed on the transvaginal US of the cervix), but it does not report any information about the compression or the structural and histological changes of the tissue. The aforementioned changes in the micro-structure of cervical tissue could be analysed from the US images and seem to have value to estimate the risk of preterm birth [11]. Some treatments to prevent it exist (progesterone, cervical cerclage) and are being developed, whih usually are only recommended to women at high risk. A correct evaluation of the risk of preterm delivery would help the clinicians to provide a more personalized treatment to the patients

The aim of our study is to evaluate the feasibility of applying artificial intelligence methods, such as machine learning and deep learning, to generate predictive models for the two clinical scenarios, using in both cases radiomic features from transvaginal cervical US images from the cervix. Furthermore, the possibility to combine other sources of information, such as clinical data from the patient's EHR (Electronic

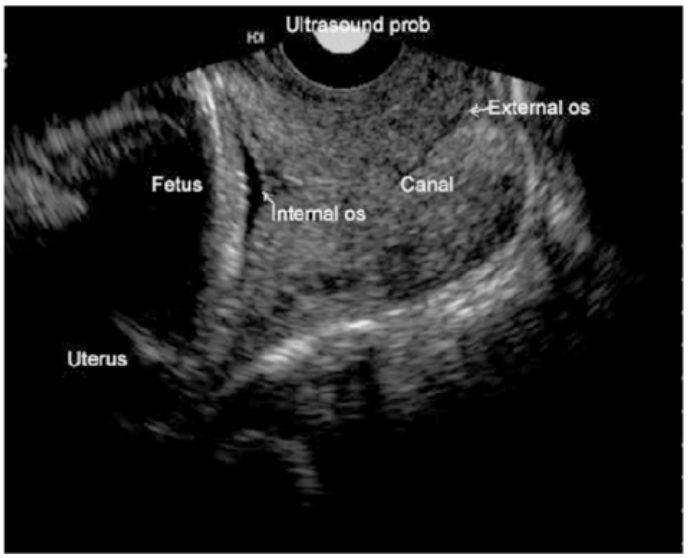

Figure 1.3: B-Mode transvaginal US of the cervix.

Health Record) are explored, as well as the use of CNNs for feature extraction.

1.1 Objectives

The objectives of this project can be summarized as follows:

- Exploring the use of radiomics analysis in obstetrics, particularly in transvaginal US.
- Building artificial intelligence models for decision support in two clinical scenarios: induction of labor and preterm birth.
- Exploring ways to combine information from the US images with other clinical relevant data. Exploring different methods to select relevant features.
- Performing an extensive evaluation of the models, to validate both the accuracy utility in a clinical environment.

1.2 Challenges

Working with medical images poses a series of domain specific challenges, many of them related to the protection of the patients' privacy, such as the need to obtain approval from ethical committees to perform the study, the difficulty of accessing the data, and the lack of big databases publicly available. Furthermore, working with US images poses a technical challenge as they present a lot of noise and artifacts, and lower resolution than other imaging modalities such as magnetic resonance imaging (MRI) or computerized tomography (CT). The main challenges of our project are:

- The difficulty to obtain a large enough database of US images. The database must contain not only images, but also labels about the image content, which requires an expert obstetrician to annotate the images.
- Obtaining a curated and structured database from the patient's electronic health records (EHRs) that can be additional information for the analysis of images.
- Working with US images, which are more noisy and have less resolution than other medical imaging modalities, making it very difficult to analyze them automatically.
- The novelty of the proposed method. There are very few similar studies in the literature and it is still not demonstrated that US images actually reflect changes in the micro-structure of the cervix.

During the project the ethical and legal issues of working with data from real patients have been taking into account. The data has been collected with the approval of the ethical committee and is fully anonymized.

Chapter 2

State of the art

This chapter summarizes the literature related to the project. We review first the literature on prediction of induction of labor (IOL) from US images, and then those concerning prediction of preterm delivery. The last section in this chapter presents a review focused on the technology rather than the clinical application. We provide a brief summary of the state of the art of machine learning methods for US imaging, covering different pathologies and clinical applications. Finally, we present a brief summary of the state of the art in deep learning for medical image analysis.

2.1 Clinical background

2.1.1 Predictive markers of the failure of induction of labour

IOL (IOL) is the treatment that stimulates childbirth and delivery. Inducing labor can be accomplished with pharmaceutical or non-pharmaceutical methods, such as a mechanical methods that promote cervical ripening and the onset of labour by stretching the cervix.

IOL is a common procedure in current obstetrics, according to the American College of Obstetricians and Gynecologists (ACOG) between 20 percent and 40 percent, but it can fail. 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 labours or fetal distress [3].

The Bishop scores has been the most widely used for the assessment of the cervical tissue prior to IOL, however, it is a subjective measure and it has been found not to be consistent [4]. Thus, a method for the proper selection of candidates for successul IOL as well as for the most adequate method for IOL is an open issue in obstetric practice.

The cervix is a complex structure composed by different cell types (muscle cells, fibroblasts, glandular cells, vascular cells) embedded in an extracellular matrix (ECM) which contains proteins -mostly collagen- and proteoglycans. The composition and strucutre of the ECM define the mechanical properties of the cervix. During pregnancy and delivery, the cervix transforms from a stiff, long, closed structure into a soft, short and dilated structure that allows delivery. Collagen is alligned and organized in the cervix of non-pregnant women, and more disorganized during

its remodelling for delivery. Water content of the cervical tissues is also increased in the process of preparation for delivery.

There are different ways to evaluate the composition of the cervix quantitatively, apart from Bishop's score.

Some techniques evaluate the deformability of the tissue, measuring the response of the tissue to a stimulation to quantify its softness or stiffness. Transvascular US is used for this purpose.

Cervical consistency index is the ratio of the anteroposterior diameter of the cervix under maximal compression with the transducer compared to the anteroposterior diameter of the cervix prior to compression. It is correlated with gestational age [12].

Strain elastography involves deforming the cervix and measuring the change in tissue displacement in a region of interest, computed from the US signals acquired before and after the deformation. The results are shown in a color map knows as elastogram.

Dynamic elastography is a similar technique that consists in displacing the tissue with a high-frequency US pulse and observing the reaction.

Texture analysis is based on the patterns and changes of brightness in the images. Changes in the cervical mirostructure and water content in the tissue are expected to be reflected in the image obtained from transvaginal US since the consistency of tissues affect their interaction with US waves. This idea has been applied to the study of neonatal respiratory morbidity from fetal lung US [13] and to assess cervical structure with spontaneous preterm birth [11] and to predict gestational age [14], and to predict IOL failure [9, 8].

To evaluate the current state of the art on prediction the failure of IOL using US images, we performed a methodological review on the PubMed database. The search term *«labor indcution failure prediction»* retrieved only 6 results, searching for publications in the last 5 years, so we explanded the time limit to 10 years, obtaining 10 results. The articles found can be classified according to the type of technique used to evaluate the cervical tissue:

- Prediction based on the deformation of the cervix, [15, 16].
- Texture analisys from the US images, [8, 9].
- Measurement of the cervical lenght, [6, 7, 17].

In [15] cervical tissue strain, assessed by elastography, was shown to be useful for the prediction of IOL failure. Mono and multivariate analysis was performed to prove that this measure was independent from cervical length, that there was no correlation with other outcomes, and that tissue strain performed better than Bishop score and cervical lenght, which were only found to predict early response to IOL. However, this was a preliminary study with a small dataset (77 patients). A recent study [16] also analyzed cervical deformation, but using the cervical consistency index instead of elastography. In their study, multivariate analylsis showed a lack of statistical association between cervical consistency index and failure of induction. Therefore the authors concluded that cervical consistency index cannot be associated with the risk of C-section delivery after IOL.

Table 2.1: Summary of the state of the art in methods for the prediction of IOL failure using transvaginal US imaging

Year	Reference	Number of patients / images	Method	Reported results
2014	[15]	77	Bishop score, cervical lenght, cervical strain	-
2015	[7]	131, (14 with failed induction)	Cervical length, funneling, position of cervix	0.90 AUC, 77% sensitivity, 93 % specifity
2016	[9]	53 patients, (9 C-section)	Local Binary Patterns and circular Gabor filters	AUC 0.83
2016	[18]	326	Simplified Bishop Score	AUC 0.88
2016	[6]	308, 187 vaginal, 58 C-section, 16 failure induction	Cervical length	R Pearson 0.237 $p<0.001$
2017	[8]	243, 22 C-section	Center Symmetric Local Binary Patterns and Gabor filters	84% accuracy
2018	[17]	70, 21 C-section	Cervical angle, cervical length	AUC 0.94 for cervical angle
2019	[16]	464	Cervical Consistency Index	CCI not associated with the risk of C-section delivery after IOL

Regarding cervical lenght, in [7] they found that the cervical score (based on cervical lenght, funneling, position of cervix, and distance of presenting part form external) was a better predictor of IOL success than Bishop's Score. In [6] shorter cervical length measurement was associated with short induction-to-delivery interval, and in [17] cervical length was found to be a predictor for failure of induction, although posterior cervical angle was found to outperform both Bishop score and cervical length. A review study published in 2016 [5] found that cervical length measurement is useful predictor for the outcome of IOL and might be better than Bishop's Score. Few studies have addressed the prediction of IOL outcome based on texture analysis. In [9], local binary patterns were used to extract texture features from the image and k-nearest neighbour and a neural network were used to classify according to IOL outcome. This study obtained good results, with an AUC of 0.88, but it was based on a sample of only 56 patients, among which just 9 had had a C-section. In [8], symmetric local binary patterns and Gabor filterbanks were used.

The main limitation of all these studies is the relatively small patient cohorts used, together with the fact that few induced labors ended in C-section compared to the ones that success to a vaginal delivery, that results in very few cases for training and validating the models. Table 2.1 summarizes the state of the art in prediction of IOL failure and the size of the database used for each study.

2.1.2 Predictive markers of preterm birth

preterm birth is defined as the birth before 37 weeks of gestation. According to the World Heath Organization (WHO) between 5% and 18% of the births (depending on the country) are preterm, resulting in 15 millions of preterm births per year [19]. It it estimated that about 1 million of those infants do not survive. Among the rest, many will suffer from some kind of disability during their lives, related to learning problems, or vision and hearing problems. Furthermore, the current tendency is an

increase in the percentage of preterm births in developed countries, due to factors as an increase in maternal age.

There are two main causes for preterm birth:

- Programmed preterm birth, due to clinical circumstances that make continuing with the pregnancy a risk for the mother or the fetus.

- Spontaneus preterm birth. It can be related to eclampsia, preeclampsia, some infections, diabetes, and other diseases. In many other cases, the cause is unknown.

In the case of spontaneus preterm birth, cervical length i the main sonographic measurement established for its prediction. During pregnancy, the length of the cervix is reduced due to the pressure of the amniotic sac. Therefore, a relation whitin preterm birth and the length of the cervix has been established [20].

However, the information obtained by cervical length is limited because it does not provide a real measurement of the compression on the cervical tissue or the structural changes in tissue composition. Elastography, a technique explained in Section 2.1.1, has been applied to the prediction of preterm birth, trying to leverage the structural changes in the tissue [21, 22]. The Cervical Consistency Index (see Section 2.1.1) has also been proposed as a way to evaluate the risk of preterm birth [23].

Finally, texture analysis has been applied to the prediction of preterm birth. In [11], cervix images from 310 women were analyzed using an algorithm based on local binary patterns, obtaining an AUC of 0.77. In [24], texture analysis was used to predict preterm birth, but analysing the fetus lungs in the US instead of the cervix.

2.2 Artificial intelligence for medical image analysis

2.2.1 Machine learning methods

Machine learning is a branch of artificial intelligence whose objective is to develop techniques that allow computers to learn. It is said that an agent learns when its performance improves with experience. The resulting models or programs must be able to generalize behaviors and inferences for a larger set of data.

It is, therefore, a process of induction of knowledge. In many cases, the field of action of machine learning overlaps with that of inferential statistics, since the two disciplines are based on data analysis. However, machine learning incorporates concerns about the computational complexity of problems.

Machine learning has a wide range of applications, including search engines, medical diagnostics, fraud detection, stock market analysis, classification of DNA sequences, recognition of speech and written language, games and robotics. We will offer a review of the specific application of machine learning to US image processing in Section 2.2.5.

Machine learning algorithms are usually divided into four main groups according to the training strategy:

- **Supervised learning**: The algorithm produces a function that establishes a correspondence between the desired inputs and outputs of the system. An example of this type of algorithm is the problem of classification, where the learning system tries to classify a series of vectors using one of several categories. The knowledge base of the system consists of examples with labels. This type of learning is the most commonly used in medical applications. However, the disadvantage is the need of annotated data to train the models. Large annotated databases of medical images or medical data are scarce, access to medical data is restricted by ethical and law requirements and clinical experts are needed to label the data, which is expensive and time consuming.

- **Unsupervised learning**: The entire modeling process is carried out on a set of examples formed only by inputs to the system. There is no information about the categories of these examples. Therefore, in this case, the system has to be able to recognize patterns in order to label the new entries. This type of learning is useful to find patterns in clinical data, but it cannot be applied directly to the development of predictive models for medicine since the target classes are unknown.

- **Semisupervised learning**: This type of algorithm combines the two previous algorithms to be able to classify adequately, both marked and unmarked data are taken into account. Semi-supervised strategies have been applied to medical image processing in the recent years to ease the burden of generating large annotated datasets; an example is the use of Generative Adversarial Neural Networks to pre-train a classifier that is later refined using annotated datasets.

- **Reinforcement learning**: The algorithm learns by observing the world around it, so its input information is the feedback it gets from the outside world in response to the actions. Therefore, the system learns based on trial-error. Reinforcement learning can be used in healthcare to refine the existing practices, analyzing the decisions which led to the maximum reward.

2.2.2 Machine learning classifiers

Several machine learning classifiers have been trained by supervised learning to build predictive models for the two clinical scenarios studied in this project. Their theoretical background is briefly explained in this section.

Decision Trees

A decision tree is a set of conditions or rules organized in a hierarchical structure, so that the final decision is given following the conditions that are fulfilled from the root node until any of the leaves. They can be used for regression or clustering, but they are best suited for classification tasks [25]. Decision trees present advantages such as being easy to use and to interpret, which is very important in the medical domain, where black box systems are less likely to be trusted; they have high tolerance to noise, non-significant attributes and missing values. However, they are *weak*

learners, having a strong dependency of the sampling of examples. Two different datasets extracted from the same underlying distribution can lead to very different trees.

There are different algorithms for learning a decision tree from data, they differ in the way to create and select partitions of the data at each node. There exist many methods to measure the similarity of the examples at each node, usually based in obtained measurements of the relative frequency of each class in the data subset generated for the child nodes with respect to the parent node. Some rules are Gini criteria, Gain criteria, Gain Ration, C4.5 and DKM. This is an important parameter that has to be chosen for building the decision tree [26].

Random Forest

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset, that is, it is a modification of *bagging* that builds a large number of trees and then and uses averaging to improve the predictive accuracy and to minimize over-fitting [27]. It is an efficient classifier for large databases and it is able to cope with high dimensional feature spaces giving an estimation of which variables are important for the classification [26]. However, in contrast to decision trees, random forests are difficult to interpret. A very important parameter for the random forest classifier is the number of estimators (number of trees) to be built. A higher number will lead to better performance, especially when the dataset has a lot of atributes, but also increases significantly the computational expense. Different similarity metrics can be used to build the trees, as explained in Section 2.2.2, which usually affects the performance of the classifier.

Extremely Randomized Trees

Extremely randomized trees are very similar to random forests, they are also built with bagging. However, instead of looking for the most discriminative thresholds, the thresholds are drawn at random for each candidate feature, and the best of those randomly generated thresholds is chosen as the splitting rule [28]. Therefore, extremely randomized trees is even more random than random forest [29].

Support Vector Machines

Support Vector Machines (SVM) are linear classifiers that induce hyper-planes in high dimensional feature spaces, trying to maximize the margin. That way, if the data is linearly separable, the SVM will find the hyper-plane which is at the same distance from the closest examples of each class. Therefore, it only considers the points that are at the boundary of the decision region. To learn non-linear SVM classifiers, the input feature space is transformed by a kernel function into a space of higher dimensionality, where the classes are linearly separable (see Figure 2.1) [30]. SVM classifiers have a good performance in high dimensional spaces and are robust to overfitting. However, a basic requisite for SVM is the correct choice of a kernel function. Common choices are Gaussian kernels, polynomic functions, or sigmoids [26].

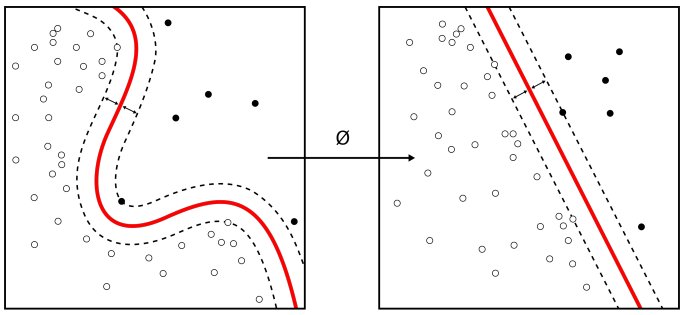

Figure 2.1: Change in the class boundaries after applying a kernel function to make the classes linearly separable.

Multi Layer Perceptron

Artificial neural networks are a computational method that tries to emulate the human brain. The multilayer perceptron (MLP) is a neural network with one or more hidden layers. It is suitable for non-linear problems and can obtain good results for both classification and regression problems [31]. Figure 2.2 shows the architecture of an example MLP, with one hidden layer. Each neuron in the hidden layer is connected to the output of all the neurons in the first layer. The MLP is trained with backpropagation, so that the error due to the weights of the first layers is calculated using mean squared error, and part of it is propagated to the hidden layers, to take into account the error caused by those weights (assigning the proportional part to the weights that cause it). The gradient descent technique is used for optimization. Since the problem is not convex, convergence to a global minima is not guaranteed. Ususally, networks converge to local minima, but these are generally good enough solutions [26]. Weights are initialized with random values, poor initialization values, given that can lead to non-convergence, some heuristics can be applied to try to prevent this.

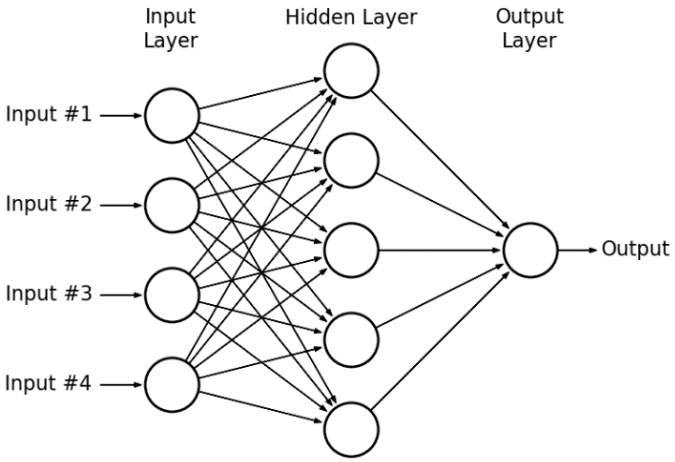

Figure 2.2: Example of a multi layer perceptron.

Naive Bayes

The Naive Bayes method consist in applying Bayes' theorem with the assumption of conditional independence between every pair of features:

$$P\left(A_i|B\right) = \frac{P\left(B|A_i\right)P\left(A_i\right)}{\sum_{k=1}^{n} P\left(B|A_k\right)P\left(A_k\right)} \dots$$

This way we can calculate the probability that the example belongs to a class given the value of its atributes [26].

2.2.3 Feature selection strategies

The training time and performance of a machine learning algorithm depends largely on the characteristics of the data set. Unnecessary and redundant features slow down the training time and also affect the performance of the algorithm. Thus, it is necessary to select the most appropriate features to train the machine models. There are different types of algorithms to perform feature selection, taking into account different characteristics of the data.

Filter methods select features regardless of the machine learning algorithm model. This is one of the biggest advantages of filter methods. Another advantage of filter methods is that they are very fast. Filter methods can be classified into: univariate filter methods and multivariate filter methods [32].

Univariate filter methods rank features according to specific criteria. Next, the N first features are selected. Thera are different types of classification criteria, for example, Fishermen's score, mutual information and characteristic variation. One of the main disadvantages of univariate filter methods is that they can select redundant features because the relationship between individual features is not taken into account when making decisions [26].

Multivariate filter methods are capable of eliminating redundant data features, since they take into account the mutual relationship between the characteristics. Multivariate filter methods can be used to eliminate duplicate and correlated features.

Wrapping methods are based on greedy search algorithms, as they evaluate all possible combinations of features and select the combination that produces the best result for a specific machine learning algorithm. A disadvantage of this approach is that testing all possible combinations of features can be computationally very expensive. Another disadvantage is that this feature set may not be optimal for any other machine learning algorithm [32].

2.2.4 Convolutional neural networks

Convolutional Neural Networks, CNNs, (LeCunn, 1989) are deep neural networks that use convolutional layers, which perform a convolution between the input data (or the output of the previous layer, it it is a hidden layer) and the filter composed by the weights of the layer. Another way to say this, is that each neuron is only connected to a locally related subset of the input and weights are shared across all the neurons from that layer. Figure 2.3 shows the difference between the fully

connected layers and convolutional layers. CNNs became very popular for computer vision applications in 2012 with the success of AlexNet[33].

In the last decades, deep learning has given a boost to medical image analysis allowing to efficiently learn features directly from the imaging data. Instead of relying on human-designed features for classification, deep learning techniques require only the datasets, from which the informative representations are directly inferred. Convolutional Neural Networks, have been applied to a wide range of medical image analysis tasks, including segmentation, regression and classification. The major limitation is the need for large annotated datasets, which has been dealt with using several strategies, such as semi-supervised training, weak labels or generative adversarial neural networks.

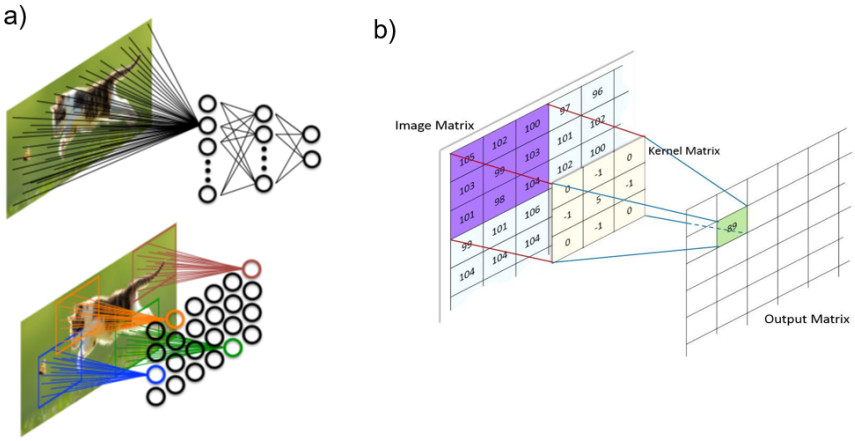

Figure 2.3: a) Difference between a fully connected layer and a convolutional layer, in the which the connections are spatially limited. (b) Example of the operation performed by convolutional layers.

2.2.5 Machine learning and deep learning applied to US image analysis

US imaging is low-cost, non-ionizing, widely available and capable of real time image acquisition and display. It is one of the most common diagnostic imaging modalities. However, the presence of noise, artefacts, high inter and intra-operator variability, and variability across different manufacturers can make it difficult to interpret the images. This motivates the use of computer vision and machine learning technologies to improve the quality of the information obtained by US imaging, to help practitioners interpret the findings and make decisions, and to reduce the time needed for those tasks. Machine learning has been used for noise reduction, segmentation, detection and classification.

The most common approach for **classification** is to extract handcrafted features, apply feature selection algorithms and then train a classifier. Features commonly used include morphological characteristics and texture features. The existence of a public database for breast cancer US (Breast US Image) has motivated the application of machine learning to this task, using classifiers based on neural networks,

support vector machines or random forests to assess the malignancy of the lesion [34, 35, 36, 37]. Machine learning has also been used for classification of US images on other organs, as liver, lung, kidney or heart [1]. A common limitation in most of the studies is the reduced number of patients or images. Furthermore, the classification usually relies on a manual selection of the region of interest (ROI), which is time consuming and operator dependent.

Machine learning algorithms have been applied for **segmentation** of anatomical and pathological structures from US images. Automatic segmentation from US images is a very challenging task since the data is affected by speckle, shadow, and missing boundaries. The process usually consists in a pixel-wise classification followed by a post processing to smooth the segmentation. Machine learning has been used for prostate segmentation from 3D US [38, 39], breast lesion segmentation [40] and carotid artery segmentation [41].

In addition to classification and segmentation, machine learning has been used for **registration** of US images. Examples include the registration of CT and US images of the spine [42] and the registration of longitudinal US images of the prostate [43].

To overcome the limitations of machine learning based methods, (e.g. manual region of interest selection) and use of handcrafted sets of features, **deep learning** based methods, such as convolutional neural networks, have been recently applied to US image processing [44, 45, 46]. However, a major disadvantage for the application of deep learning to US images is the lack of large annotated image datasets that can be used for supervised training. The generation of large US databases faces ethical issues (need for patient consent, ethical committee approval, anonimization and compliance with data protection regulation) as well as practical issues (it is very time consuming and requires highly qualified personal, such as doctors and radiologists).

In general, the application of machine learning and deep learning to US imaging is in an early stage, behind other modalities such as CT and MRI, but it rapidly progressing.

Chapter 3

Materials and Methods

This chapter presents the materials and methods used in this project.

The materials consist of two databases for each of the two clinical problems. For IOL, both US images and clinical data from the electronic health record are available, while for preterm birth we work only with US images. Details about the size of the database and acquisition protocols are given in the following section.

Regarding the methods, several steps are needed to build models using US images and clinical data. Firstly, features have to be extracted from the images, using radiomics or CNNs. Then, all the features, either clinical or image-based, are pre-processed, applying feature selection strategies to reduce the dimensionality of the feature vector. We also had to solve problems such as missing values and class imbalance. After pre-processing, we trained and built the models, and finally we evaluated the results in terms of AUC. More details about each of the steps are given in Section 3.2.

3.1 Materials

The following sections explain the data that was used in this project for the two clinical scenarios: prediction of IOL failure and prediction of preterm birth. It must be noted that for both studies, all the images as well as the clinical data have been anonimized, and that the studies got the approval of the ethical committee of the hospital (Hospital Universitario de Cruces, Bilbao, Spain).

3.1.1 Data about IOL

The database used in this study consists of images and clinical data from patients admitted for IOL at Hospital Universitario de Cruces. The patients underwent a transvaginal US before IOL. Images were acquired with a Voluson US scanner from General Electric by the expert obstetrician, following the same protocol for all patients. Image resolution is 720960 pixels with a pixel spacing of 0.11. All images were provided in DICOM format. A region of interest (ROI) was drawn manually on the images by an expert obstetrician, so only a relevant region of the cervix was further analyzed. Figure 3.2 shows examples of the input images and ROIs.

Data was available from a total of 182 patients, from which 130 had a vaginal delivery and the rest, 52, needed a C-section. The data included US images, annotation of the ROI and clinical data In 30 cases the cause of the C-section was known to be failure of induction related to a cervical motive. Figure 3.1 summarizes the database composition.

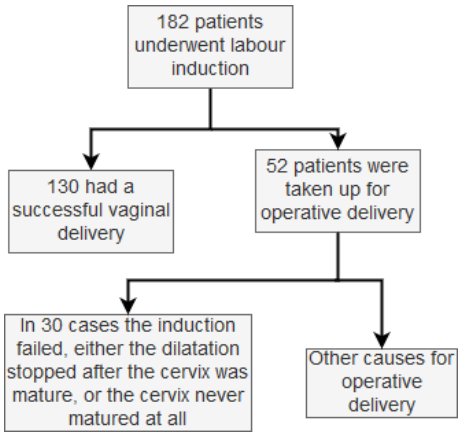

Figure 3.1: Dataset composition.

Twenty relevant clinical attributes were selected from the database to be included in the study, (including age, weight, height, race, body mass index, number of abortions, weeks of gestation, and information about previous pregnancies). Seven sonographic measurements, manually extracted from the trans vaginal US, were also included: 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.

3.1.2 Data about preterm birth

Week of delivery	**Number of images**
preterm	**178**
24 - 34 weeks	93
35 - 36 weeks	85
Term	**205**
37 - 40 weeks	86
41 - 42 weeks	119

Table 3.1: Distribution of the patients available for the preterm birth study in different classes.

The database used in this study consists of images from patients admitted for IOL at the same hospital. The patients underwent a transvaginal US during a routine examination in the 20th week of pregnancy . Images were acquired with a Voluson US scanner from General Electric (the same as for IOL) by an expert obstetrician, following the same protocol for all patients. A total of 383 patients with US images were available, 178 of which had a preterm delivery, as shown in

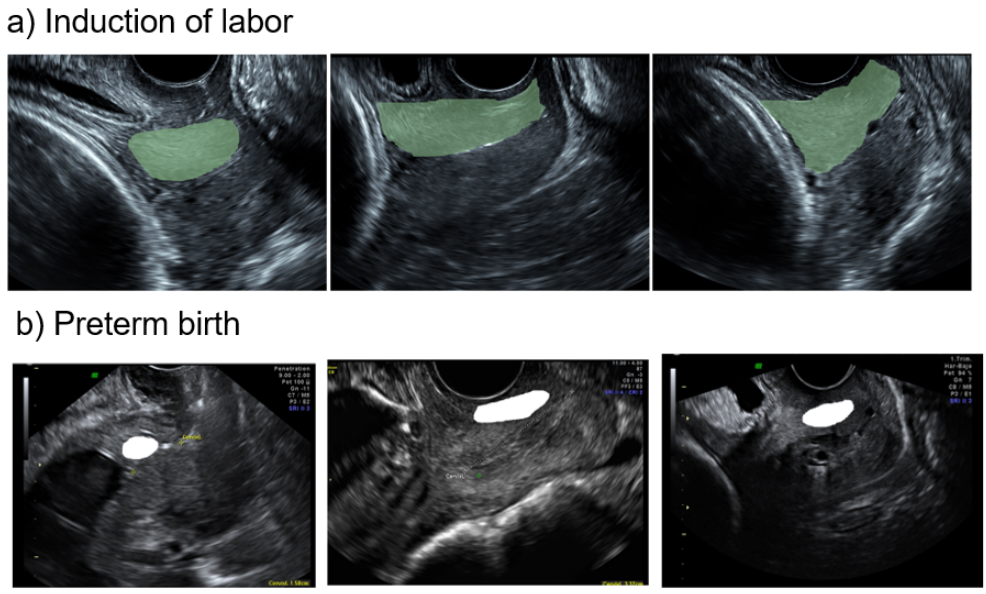

Figure 3.2: Examples of the transvaginal US images and the selected region of interest for radiomic analysis.

Table 3.1. All images were provided in TIF format. Image resolution is 974×660 pixels. A ROI was drawn manually on the images by an expert obstetrician, so only a relevant region of the cervix was further analyzed (see figure 3.2).

The clinical data related to these cases was not available at the time of this study, so it was not incorporated.

3.2 Methods

The code for the analysis has been developed in Python, using some of the most popular Python libraries for machine learning: Scikit-learn [28], Pandas [47] and Keras [48].

The following sections give more insight about each of the algorithms, which include functions for data pre-processing, feature selection strategies, radiomics and CNNs to extract features from the US images and machine learning classifiers to build predictive models. Figure 3.3 depicts the workflow of the project.

3.2.1 Machine learning models

Within this project, different **supervised classifiers** we have trained to generate predictive models and compared their results. A theoretical description of each classifier is given in Section 2.2.2, and Table 3.2 summarizes the classifiers used and their more important parameters.

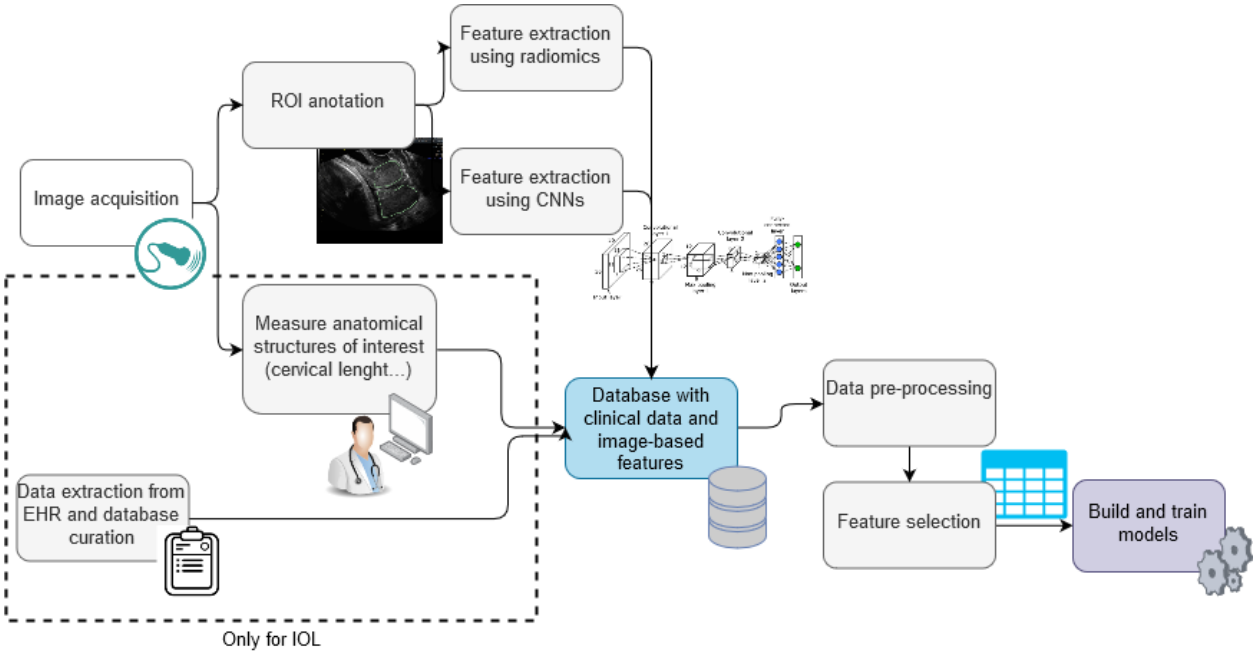

Figure 3.3: General workflow of the project.

Table 3.2: Relevant parameters for the classifiers used in the study: Gaussian Naive Bayes (GNB), Random Forest (RF), Multi Layer Perceptron (MLP), Support Vector Machine (SVM), Decision Tree and Extremely Randomized Trees (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
Decision tree	Impurity: Gini, minimum samples per split: 2
Extra Tree	Number of estimators: 150, minimum samples per split: 2, impurity: Gini

3.2.2 Radiomics

Radiomics consist in extracting large amount of features from radiographic medical images using a data-characterisation algorithm [2]. The basic process common to every application that uses radiomics consists in: (1) acquiring the images, (2) segmenting the region of interest (often using semi-automatic or automatic segmentation), (3) extracting the features, which may include volume, shape, texture, or other information, and (4) analysis.

Radiomic features can be divided into:

- size and shape-based features,
- descriptors of the image intensity histogram, and,
- texture features, such as gray-level co-occurrence matrix (GLCM), run-length matrix (RLM), size-zone matrix (SZM), and neighborhood gray-tone difference matrix (NGTDM) derived features.

Within this project, only texture features and intensity features are considered. The open-source Python package PyRadiomics [49] is used for feature extraction.

The following matrices are computed and some mathematical descriptors are extracted from them:

Gray Level Co-occurrence Matrix (GLCM): It 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 1 pixel; angles are computed automatically.

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

Gray Level Dependence Matrix (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 neighbouring voxel with gray level j is considered dependent on the center voxel with gray level i if $|i - j| \leq \alpha$. In a gray level dependence matrix the $(i, j)_{th}$ element describes the number of times a voxel with gray level i having j dependent voxels in its neighborhood appears in the image. The parameters we 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).

Nineteen first order features based on image intensity are included as well (such as energy, entropy, minimum, maximum, mean, median, interquartile range, skewness, kurtosis...) To illustrate the process of texture features computation, Figure **??** shows an example of the GLCM computation from a very simple gray level image, with distance 1 and angle 0 (horizontal). Table 3.3 presents the total of radiomic features used.

3.2.3 Convolutional neural networks

CNNs can be trained end-to-end for regression, classification, detection or segmentation. They can also be used to extract features that are then fed into a machine learning classifier; for instance, the first approach to object detection using CNNs [50] worked that way.

In this project, a popular CNN architecture for image classification is used to extract features from the US images. We used ResNet50 [51], the model available in Keras[48], which is pre-trained on ImageNet[52]. Keras is a high-level neural

	Features	Total number
First-order features	energy, total energy, entropy, minimum, maximum, 10th percentile, maximum, mean, median, interquartile range, range, mean absolute deviation, robust mean absolute deviation, root mean squared, standard deviation, skewness, kurtosis, variance, uniformity	19
GLCM	autocorrelation, joint average, cluster prominence, cluster shade, cluster tendency, contrast, correlation, difference average, difference entropy, difference variance, joint energy, joint entropy, informational measure of correlation,inverse difference moment, maximal correlation coefficient, inverse difference moment normalized, inverse difference, inverse difference normalized, inverse variance, maximum probability, sum average, sum variance, sum entopry, sum of squares	24
GLRLM	short run emphasis, long run emphasis, gray level non-uniformity, gray level non-uniformity normalized, run length non uniformity, run length non uniformity normalized, run percentage, gray level variance, run variance, run entropy, low gray leven run emphasis, high gray level run emphasis, short run low gray level emphasis, short run high gray level emphasis, long run low gray level emphasis, long run high gray level emphasis	16
GLDM	small dependence emphasis, large dependence emphasis, gray level non uniformity, dependence non uniformity, dependence non uniformity normalized, gray level variance, dependence variance, dependence entropy, low gray level emphasis, high gray level emphasis, small dependence low gray level emphasis, small dependence low gray level emphasis, large dependence high gray level emphasis	14
NGTDM	coarseness, contrast, busyness, complexity, strength	5

Table 3.3: Radiomic features used in this study.

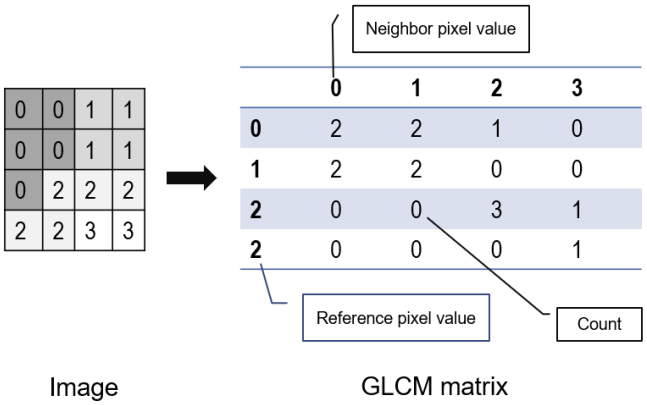

Figure 3.4: Example of calculation of GLCM matrix for a very simple 2D grayscale image, with distance=1 and angle=0.

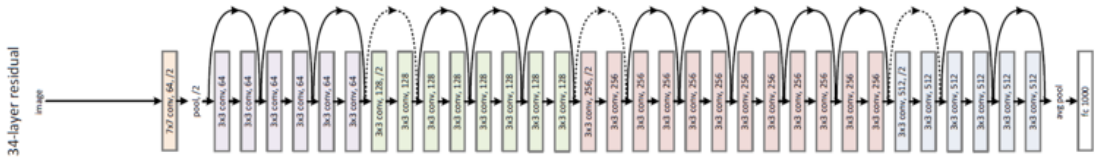

Figure 3.5: ResNet architecture, with skip-connections to allow training deeper networks.

networks API, written in Python. In this project it is used with as a TensorFlow backend.

Figure 3.5 presents the architecture of ResNet. Features are extracted from the last average pooling layer, obtaining a vector of 2048 features, that is reduced later on.

3.2.4 Pre-processing of the data

Before feeding the data to the machine learning classifiers or to the feature selection process, we had to address problems with the dataset:

1. **Missing values:** Different patients have different clinical data available, thus resulting in missing values for some of the attributes. When an attribute value was missing for more than 50% of the cases, that column was removed. In the rest of cases, the missing values were filled with the mean, using the functions available in the Python package Pandas.

2. **Class imbalance:** For both clinical scenarios, the number of instances in the classes imbalanced, however, this problem is especially relevant in the case of prediction of failure of IOL, with 130 and 52 cases per class. To overcome this problem, we used *SMOTE - Synthetic Minority Over-sampling Technique* [53] to generate artificial samples that are used during training.

3.2.5 Feature selection strategies

Working with radiomics leads to a high dimensional feature space, as shown in Table 3.3, where a total of 78 features are presented. Furthermore, we add clinical features to the radiomic features vector. In the case of features extracted from a CNN, the vector has 2048 features.

While there is not a mathematical rule about the number of features to select, some common heuristics are that 10 data points are needed for each model parameter and 3 to 5 independent cases are needed per class and feature. Taking into account that the sample size for the studies in this project is 182 and 383 cases, we can assume that the classifiers will have difficulties learning (problem known as *curse of dimensionality*). Therefore, we have applied feature selection strategies to reduce the feature vector and the results with and without feature selection have been compared. The Python packages Scikit-learn[28] and mlxtend [54] have been used.

Sequential Forward Feature Selection

Sequential forward feature selection is a wrapper method for feature selection. **Wrapper methods** are based on greedy-search algorithms that evaluate all the possible combinations of the features and select the combination that produces the best result. Sequential feature algorithms are a suboptimal solution to the exhaustive search of all the possible feature combinations, since finding the optimal solution is not computationally feasible in many cases.

In this project we use the Sequential Forward Feature Selection implemented in http://rasbt.github.io/mlxtend/. The algorithm is initialized with tan empty set

and in each iteration an additional feature is added to the selection. The feature added is the one that maximizes the criterion function, that is, the feature associated with the best classifier performance when it is added to the subset. In each iteration, a new feature is added. The process goes on until the subset has the desired size. We can plot the performance of the classifier as we add features to choose the optimal number of features to use.

Random Forest Feature Selection

Random forest classifiers naturally assign an importance to the features when building trees. As explained in Section 2.2.2, the nodes of the tree are built based on a metric of the quality of the generated subsets are (e.g. Gini impurity). Random forest performs bagging, so different trees are built with different subsets of features. For each tree, the nodes will be created according to Gini impurity; therefore, nodes with the greatest decrease in impurity will be at the beginning of the trees. Random forest classifiers can be used to obtain the importance of each feature, order the features according to that, and to select the desired number of features according to the ranking. Feature selection using random forest can be considered as a **filter method** as it ranks the features according to their importance given by Gini impurity and then select the top N features.

Principal Component Analysis (PCA)

PCA is not strictly a feature selection method, but it is included here because it has been used to reduce the number of features used during training. PCA reduces the dimensionality of the feature space by applying a transformation that converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables. The new features are defined in a way that the first has the largest possible variance and the resulting vectors are an uncorrelated orthogonal basis set. The main disadvantage of this method with respect to the previous ones is that, as it transform the features, the newly generated features will not have a clinical meaning and therefore the models will lack interpretability.

3.2.6 Evaluation of the results

Area under the curve (AUC)

To evaluate the performance of the classifiers, the Area Under the Curve (AUC) is reported for every experiment. This is the area under the Receiver Operating Characteristic (ROC) curve, which is a graphical representation of sensitivity versus specificity for a binary classifier system as the discrimination threshold is varied. Another meaning of this graph is the representation of the true positive ratio versus the false negative ratio, also as the discrimination threshold is varied.

To provide a more detailed analysis of the results, for some of the most interesting experiments the values of sensitivity, specificity, false positive ratio and false negative ratio are provided as well.

$$\text{sensitivity} = \frac{\text{TP}}{\text{P}} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$
$$\text{specificity} = \frac{\text{TN}}{\text{N}} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$
$$\text{FNR} = \frac{\text{FN}}{\text{P}} = \frac{\text{FN}}{\text{FN} + \text{TP}}$$
$$\text{FPR} = \frac{\text{FP}}{\text{N}} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

Statistical significance tests

Within the project, hypothesis testing is used to determine whether the difference found in the results of some experiments are statistically significant or not and p-values are reported. The Python package Scipy [55] is used. We use the T-test for means of two independent samples from descriptive statistics. It is a two-sided test for the null hypothesis that two independent samples (the results for the two different techniques) have identical average values and the difference observed is due to random variations. That is, the null hypothesis is that there is no difference in the performance of the two models.

Chapter 4

Results

This chapter explains the results obtained withing the project, for the two clinical scenarios studied: prediction of failure of IOLand estimation of the risk of preterm birth. The subjective discussion about the results and the conclusions will be presented in the next chapter, reserving this one for the objective data.

4.1 Prediction of IOL

Heterogeneous data is available for this problem, as explained in Section 3.1.1. The data can be divided into: (1) clincal data from the EHR (electronic health record). (2) sonographic measurements from anatomical structures such as cervical lenght, cervical angle, etc., (3) radiomic features from the US image, and (4) CNN features from the US image. Several experiments with different combinations of data have been designed. Thus, in this section we will present results of the experiments that used:

- only clinical data,
- only sonographic measurements,
- both clinical data and sonographic measurements,
- only radiomic features,
- results combining radiomic features, clinical data, and sonographic measurements.
- only using the features extracted from the CNN.

For each combination of features we have applied different feature selection techniques.

We provide a summary of the most relevant results from the best performing experiments is provided. An extensive evaluation of the results and comparison between the different experiments is given on Chapter 5.

It is worth noting that for every experiment, machine classifiers have been trained using a 10 k-fold cross validation strategy, so the results reported are always the mean of the 10 folds.

4.1.1 Prediction of IOL using only clinical data

Table 4.1.1 shows the mean Area Under the Curve (AUC) for the models created training machine learning classifiers (as explained in Section 3.2.1) using 20 clinical features. The top-left plot in Figure 4.1 shows the correlation between these clinical features, as well as the correlation between each feature and the outcome. This correlation is very low as it can observed in the last row and the last column.

Classifier	AUC(mean)	AUC(std)
Gaussian NB	0.622	0.029
Random Forest	0.621	0.014
MLP	0.512	0.0575
SVM (rbf)	0.515	0.035
SVM (sigmoid)	0.499	0.0401
Decision Tree	0.572	0.030
Extra Trees	0.614	0.020

Table 4.1: Results (mean and std AUC) using only clinical data

4.1.2 Prediction of IOL using only sonographic measurements

Seven sonographic measurements (of Section 3.1.2) have been used to train machine learning classifiers. The AUC obtained for every classifier is shown on table 4.2.

Classifier	AUC (mean)	AUC (std)
Gaussian NB	0.682	0.009
Random Forest	0.652	0.012
MLP	0.678	0.012
SVM (rbf)	0.507	0.029
SVM (sigmoid)	0.494	0.030
Decision Tree	0.525	0.036
Extra Trees	0.639	0.0165

Table 4.2: Results (mean and std AUC) using sonographic measurements

4.1.3 Prediction of IOL with a combination of sonographic measurements and clinical data

The combination of clinical data and sonographic measurements yields a feature vector of size 27. This vector can be too big for the number of instances in the dataset, so we have explored several feature selection techniques. Three different experiments have been performed:

- Training machine learning classifiers using all the 27 features,
- using sequential forward feature selection (SFFS, see Section 3.2.5) to reduce the feature vector and then training machine learning classifiers, and

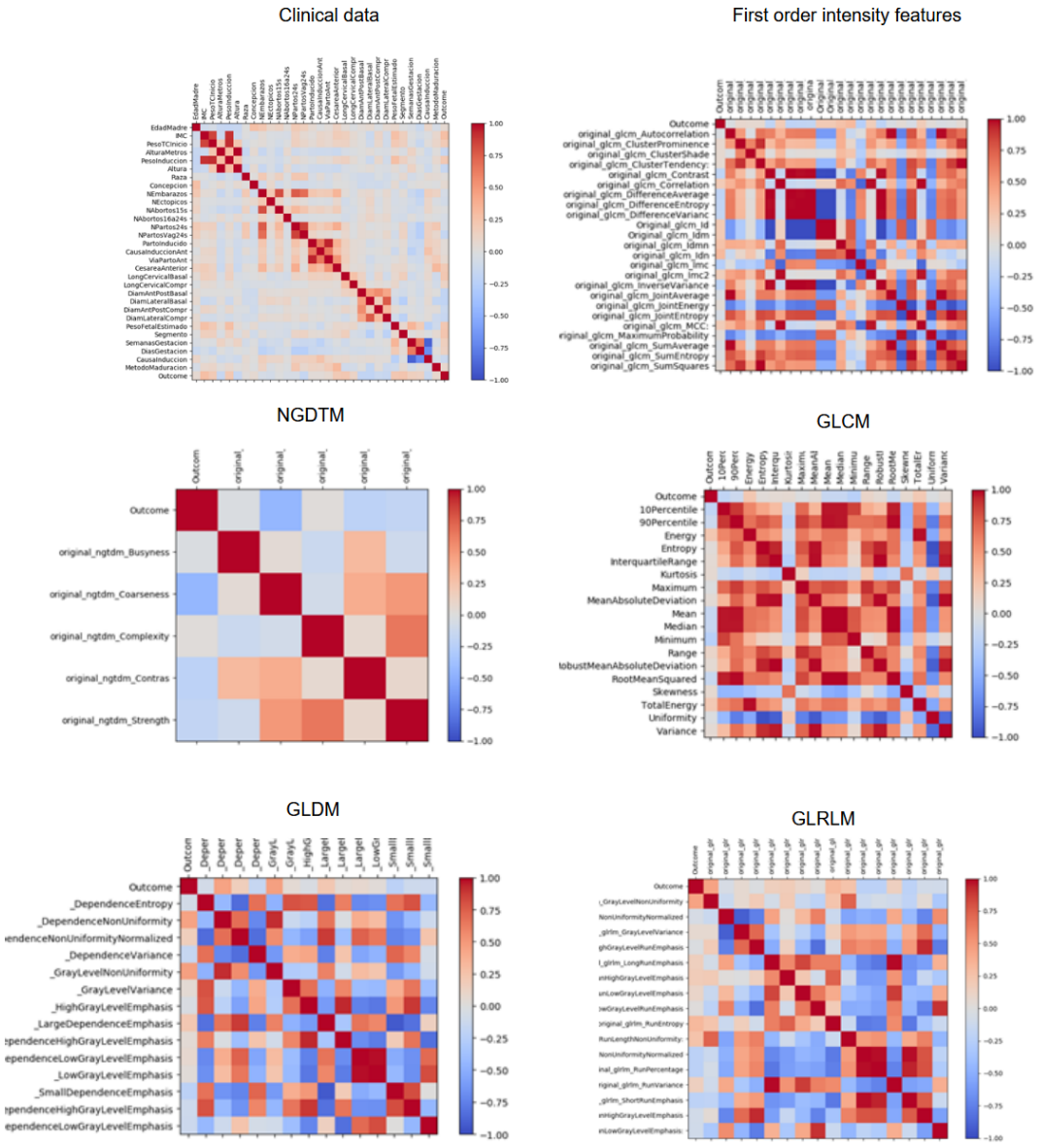

Figure 4.1: Plot of the correlation between different sets of features (clinical , first order intensity features and radiomic features).

- using random forest importance (RFFS, see Section 3.2.5) to reduce the feature vector and then training machine learning classifiers.

Table 4.3 summarizes the results of the three experiments. For both feature selection algorithms we have implemented a voting strategy, so that the algorithm is run several times with different subsets of data and only the features that are consistently selected are kept. This helps to prevent over-fitting by choosing only the most robust features.

The features eventually kept are different for the two algorithms, as shown in Table 4.4. Figure 4.3 shows the results of running sequential forward feature selection on different subsets of the data. It can be seen that the selected features are different

	All the features		Selected features (RFFS)		Selected features (SFFS)	
	AUC	**std**	**AUC**	**std**	**AUC**	**std**
Gaussian Naive Bayes	0.643	0.029	0.705	0.01	0.670	0.018
Random Forest	0.754	0.021	0.765	0.011	0.608	0.007
Multi Layer Perceptron	0.599	0.048	0.582	0.04	0.680	0.015
SVM (RBF)	0.486	0.054	0.48	0.046	0.650	0.020
SVM (sigmoid)	0.509	0.047	0.5	0.048	0.495	0.034
Decision Tree	0.631	0.046	0.637	0.018	0.580	0.019
Extra Trees	0.719	0.012	0.747	0.012	0.608	0.015

Table 4.3: Results (mean ± AUC) of the models for IOL prediction using sonographic measurements and clinical data with:(1) all the features, (2) features selected using random forest for feature selection (RFFS), (3) features selected using sequential forward feature selection (SFFS).

SFFS	**RFFS**
previous C-section, number of abortions, number of ectopic pregnancies, number of previous deliveries, number of previous induced deliveries, race, weeks of gestation, previous way of delivery (vaginal or C-section), body mass index	age, body mass index, weight, height, number of previous vaginal deliveries, basal cervical lenght, compressed cervical lenght, basal anterior posterior diameter, basal lateral diameter, compressed anterior posterior diameter, compressed lateral diameter, estimated fetal weight, weeks of gestation, segment

Table 4.4: Selected features using sequential forward feature selection and random forest feature selection

on every run, but the number of optimal features (arrow) is always around 10 to 13.

4.1.4 Prediction of IOL using only radiomic features vs. combining features and radiomic features

As explained in Section 3.2.2, radiomic features have been extracted from the transvaginal US images. To test their predictive value as biomarkers of failure of IOL, several experiments have been peformed:

- Training using only the radiomic features,
- combining radiomic features with the clinical and sonographic features (used in the previous sections),
- combining radiomic, sonographical, and clinical features, applying feature selection strategies, with either SFFS or RFFS.

Table 4.5 shows the results for all the experiments, excetp the results for feature selection with SFFS becasue that selection algorithm was not consistent. As it can

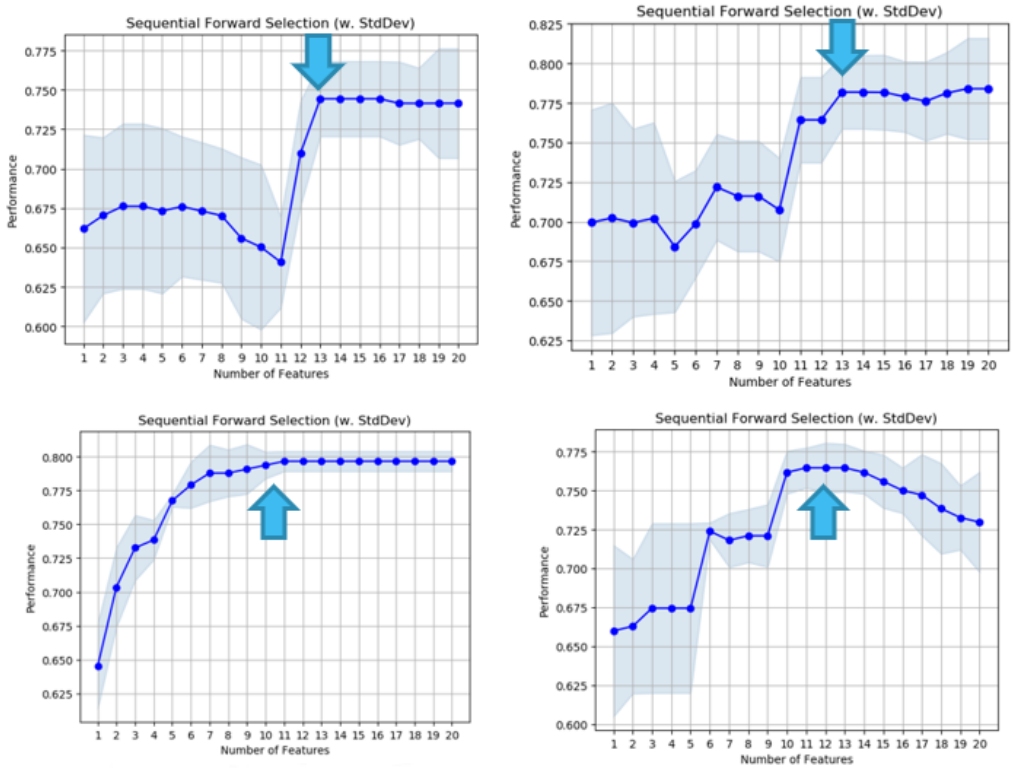

Figure 4.2: Example of four plots obtained in different runs of the sequential feature selection algorithm. Performance is plotted against number of features. The arrows indicate the maximum points, which correspond to the optimal number of features.

be seen on figure 4.3, the number of selected features was very different between different runs.

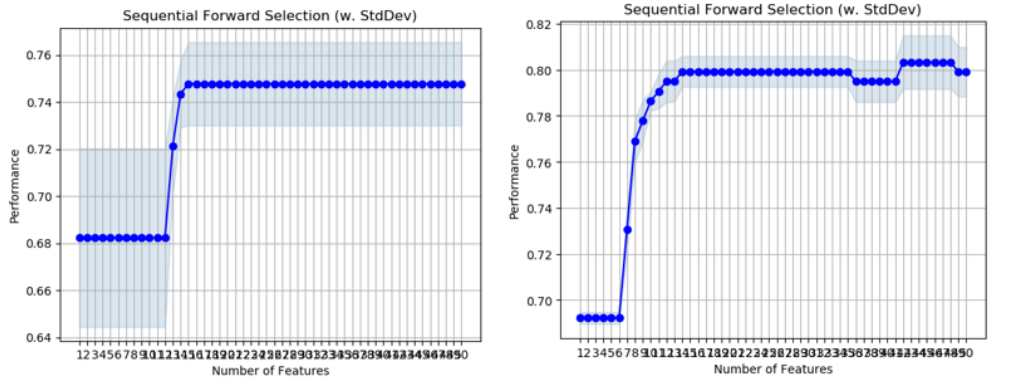

Figure 4.3: Example of two plots obtained in different runs of the sequential feature selection algorithm. Performance is plotted against the number of features.

	Only radiomics		Radiomics + other		Selected features	
	AUC	**std**	**AUC**	**std**	**AUC**	**std**
Gaussian Naive Bayes	0.471	0.022	0.487	0.017	0.505	0.026
Random Forest	0.521	0.034	0.642	0.019	0.746	0.017
Multi Layer Perceptron	0.471	0.033	0.501	0.049	0.505	0.052
SVM (RBF)	0.500	0.000	0.497	0.008	0.500	0.000
SVM (sigmoid)	0.459	0.039	0.542	0.056	0.510	0.041
Decision Tree	0.511	0.042	0.582	0.033	0.620	0.032
Extra Trees	0.495	0.018	0.626	0.026	0.769	0.014

Table 4.5: Mean AUC and std for all the experiments: (1) radiomic features, combination of radiomic features, (2) clinical data and (3) sonographic measurements and features selected using random forest feature selection.

4.1.5 Prediction of IOL with features extracted from the CNN

In the last experiment, a feature vector of 2048 features is extracted from the images using a Convolutional Neural Network (CNN). Because the feature vector is too big, before feeding the features into the classifiers they are transformed using Principal Component Analysis (as explained in Section 3.2.5) to obtain a vector of 10 features.

Two different experiments have been carried out: (1) extracting features from the whole image, (2) extracting features only from the ROI (the same region used for the radiomic feature extraction). Table 4.6 shows the results for both experiments.

	Whole image		ROI	
Classifier	**AUC(mean)**	**AUC(std)**	**AUC(mean)**	**AUC(std)**
Gaussian NB	0.554	0.017	0.594	0.021
Random Forest	0.534	0.023	0.579	0.019
MLP	0.525	0.020	0.524	0.025
SVM (rbf)	0.456	0.041	0.585	0.025
SVM (sigmoid)	0.471	0.027	0.509	0.044
Decision Tree	0.499	0.050	0.515	0.026
Extra Trees	0.513	0.020	0.585	0.018

Table 4.6: Results (mean AUC and std) for the classifiers trained with image features extracted using a CNN (ResNet50), from the whole image and from a ROI.

Finally, in order to compare the results using only radiomics and only CNN features in the same conditions, PCA has been applied to radiomic features. Table 4.7 summarizes the results.

Classifier	**AUC(mean)**	**AUC(std)**
Gaussian NB	0.471	0.022
Random Forest	0.521	0.034
MLP	0.471	0.033
SVM (RBF)	0.500	0.000
SVM (sigmoid)	0.459	0.039
Decision Tree	0.511	0.042
Extra Trees	0.495	0.018

Table 4.7: Radiomic features after PCA feature selection (mean and standar deviation).

4.1.6 Summary of the 7 experiments

Many different experiments have been done. To facilitate the comparison between them and the further analysis of the results, these section provides a summary of the best performing experiments. Furthermore, for the best performing classifiers, sensitivity, specificity, false positive rate, and false negative rate are provided as well as AUC, to give more insight about their behavior and performance. Table 4.8 summarizes the data used in each experiment. Table 4.9 shows the AUC(mean) and std for all the classifiers, while table 4.10 presents sensitivity, specificity, false positive rate and false negative rate for the best classifiers.

Exp.	**Data description**	**Total**
1	Clinical	20 features
2	Clinical and sonographic	27 features
3	Clinical and sonographic, after feature selection	16 features
4	Sonographic	7 features
5	Clinical and sonographic + radiomic	104 features
6	Clinical and sonographic + radiomic after feature selection	16 features
7	Features extracted from CNN, after PCA	10 features

Table 4.8: Summary of the data used for each experiment

Classifier	**Exp. 1**	**Exp. 2**	**Exp. 3**	**Exp. 4**	**Exp. 5**	**Exp. 6**	**Exp. 7**
Gaussian NB	0.622±0.029	0.643±0.029	0.705±0.01	0.682±0.010	0.487±0.017	0.505±0.026	0.584±0.014
Random Forest	0.621±0.014	0.754±0.021	0.765±0.011	0.652±0.012	0.642±0.019	0.746±0.017	0.566±0.02
MLP	0.512±0.058	0.599±0.048	0.582±0.04	0.678±0.012	0.501±0.049	0.505±0.052	0.531±0.019
SVM (RBF)	0.515±0.035	0.486±0.054	0.48±0.046	0.507±0.029	0.497±0.008	0.5±0	0.603±0.014
SVM (sigmoid)	0.499±0.04	0.509±0.047	0.5±0.048	0.495±0.031	0.542±0.056	0.51±0.041	0.49±0.03
Decision Tree	0.572±0.03	0.631±0.046	0.637±0.018	0.525±0.037	0.582±0.033	0.62±0.032	0.526±0.021
Extra Tree	0.614±0.02	0.719±0.012	0.747±0.012	0.639±0.016	0.626±0.026	0.769±0.014	0.571±0.018

Table 4.9: Mean AUC and std for all the experiments

	Classifier	Sensitivity	Specificity	FPR	FNR	AUC
Exp. 1	Gaussian Naive Bayes	0.808	0.338	0.662	0.192	0.622
	Random Forest	0.423	0.885	0.223	0.635	0.621
Exp. 2	Random Forest	0.462	0.831	0.169	0.538	0.754
	Extra Trees	0.423	0.885	0.115	0.577	0.719
Exp. 3	Random forest	0.596	0.825	0.175	0.404	0.765
	Extra trees	0.451	0.858	0.142	0.549	0.747
Exp. 4	Gaussian Naive Bayes	0.632	0.682	0.318	0.368	0.682
	MLP	0.600	0.677	0.323	0.400	0.678
Exp. 5	Random forest	0.365	0.838	0.162	0.635	0.642
	Extra trees	0.308	0.900	0.100	0.692	0.626
Exp. 6	Random Forest	0.692	0.715	0.285	0.308	0.746
	Extra Trees	0.481	0.846	0.154	0.519	0.769
Exp. 7	Naive Bayes	0.523	0.571	0.429	0.477	0.584
	Random Forest	0.462	0.674	0.326	0.538	0.566

Table 4.10: Sensitivity, specificity, false positive rate and false negative rate for the two best classifiers for each of the experiments.

	All features		Selected features	
	AUC (mean)	**std**	**AUC (mean)**	**std**
Gaussian NB	0.755	0.003	0.773	0.002
Random Forest	0.757	0.005	0.768	0.005
MLP	0.504	0.021	0.500	0.016
SVM (rbf)	0.500	0.000	0.500	0.000
SVM (sigmoid)	0.505	0.018	0.486	0.032
DecisionTree	0.632	0.021	0.524	0.023

Table 4.11: Results (mean AUC and std) for the experiment with the 77 radiomic features and the 54 features selected based on their correlation.

4.2 Prediction of preterm birth

The results for the second clinical case, the prediction or estimation of the risk of preterm birth are presented in this section. Because clinical data was not available, only features extracted from the US images are used (radiomic features and CNN features). Methods for feature selection have been explored as well.

4.2.1 Prediction of preterm birth with radiomic features

Table 4.11 shows the results for the models built using the 77 radiomic features. The correlation between the features was analyzed as well, and the features that were very correlated were removed, as they are are redundant. Consequently, a vector of 54 features was kept.

4.2.2 Feature selection for the prediction of preterm birth

The same feature selection process of the previous section (prediction of failure of IOL) has been applied to this case, that is, SFFS and random forest importance feature selection. Sequential forward feature selection resulted in 12 features, while only 8 were kept after RFFS. Figure 4.4 depicts the plots of two runs of the sequential feature selection algorithm on different subsets of the data. Table 4.12 compares the results with both methods.

	SFFS		RFFS	
	AUC (mean)	**std**	**AUC (mean)**	**std**
Gaussian NB	0.692	0.006	0.767	0.009
Random Forest	0.739	0.009	0.762	0.006
MLP	0.715	0.040	0.596	0.043
SVM (rbf)	0.593	0.010	0.484	0.019
SVM (sigmoid)	0.504	0.031	0.497	0.030
DecisionTree	0.631	0.021	0.640	0.022

Table 4.12: Results for the two feature selection algorithms, sequential forward feature selection (SFFS) and random forest feature selection (RFFS).

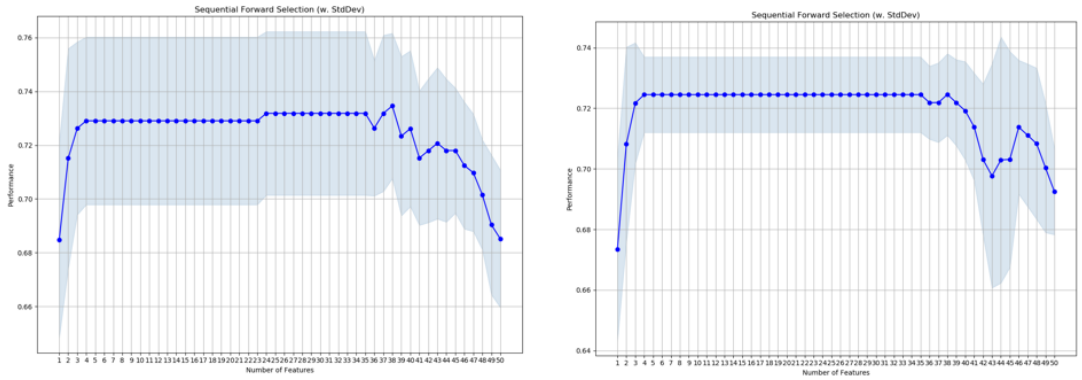

Figure 4.4: Example of two plots obtained in different runs of the SFFS algorithm for the dataset for preterm birth prediction. Performance is plotted against number of features.

4.2.3 CNN features for the prediction of preterm birth

As we did for IOL failure, we extracted features using a CNN (ResNet50) and reduced the feature vector using PCA. Two experiments were performed, using the whole image and using only the selected ROI. Table 4.13 presents the results.

	Whole image		**Region of interest**	
	AUC (mean)	**std**	**AUC (mean)**	**std**
Gaussian NB	0.679	0.006	0.756	0.004
Random Forest	0.686	0.005	0.753	0.008
MLP	0.639	0.006	0.674	0.009
SVM (rbf)	0.544	0.017	0.521	0.012
SVM (sigmoid)	0.530	0.015	0.571	0.024
DecisionTree	0.584	0.023	0.638	0.023

Table 4.13: Results for the models built using CNN features, extracted from the whole image and from the manually chosen ROI.

4.2.4 Summary of the 4 experiments

To provide more insight about the performance of different classifiers, sensitivity, specificity, false positive rate and false negative rate are provided for the best classifiers. Table 4.14 summarizes the results for experiment 1 (54 features selected after removing the ones with higher correlation), experiment 2 (12 features selected after sequential forward feature selection), experiment 3 (8 features selected with random forest feature selection) and experiment 4 (CNN features after PCA).

	Classifier	Sensitivity	Specificity	FPR	FNR
Exp. 1	**Gaussian NB**	0.572	0.834	0.165	0.427
	Random Forest	0.701	0.697	0.302	0.298
Exp. 2	**Random Forest**	0.706	0.668	0.331	0.293
	MLP	0.671	0.640	0.360	0.328
Exp. 3	**Gaussian NB**	0.830	0.525	0.474	0.169
	Random Forest	0.711	0.662	0.337	0.288
Exp. 4	**Random Forest**	0.661	0.752	0.247	0.338
	Gaussian NB	0.706	0.732	0.2673	0.293

Table 4.14: Sensitivity, specificity, false positive rate (FPR) and false negative rate (FNR) for the two best classifiers that had the best performance for prediction of preterm birth.

Chapter 5

Discussion

This section presents the discussion of the results obtained and the conclusions we can draw from them. First, the results for both experiments (prediction of failure of IOL and prediction of preterm birth) are discussed independently, then they are compared and global conclusions are obtained.

5.1 Prediction of failure of IOL

The results for the prediction of failure of IOL (IOL) are detailed in Section 4.1. Clinical data alone (as seen in Table 4.1.1) does not seem to have high predictive value (the best classifier, Gaussian naive Bayes, achieves 0.62 ± 0.02 AUC), 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 Gaussian naive Bayes classifier, the complete results are in Table 4.2. 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 [**?**]. In the next experiments (tables 4.3 and 4.5) we combine clinical data, sonographic measurements and radiomic features. The results obtained only with clinical data and sonographic measurements are similar to the results obtained adding radiomic features. Similar maximum AUC values are achieved. The random forest and extreme trees classifiers perform the best in both cases, yielding AUC values between 0.747 and 0.769. When comparing the highest AUC values from both methods (experiment 3, random forest: 0.763 ± 0.011; experiment 4, extra trees: 0.769 ± 0.014) the 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 behavior with respect to false negatives and false positives. A more detailed analysis of the results, provided in table 4.10, shows the sensitivity, specificity, false positive rate and false negative rate for these classifiers. Overall, false negative rate is higher than the 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).

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, 0.75

AUC with a sensitivity of 69% and a specificity of 71% using a Random Forest classifier, 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. However, as seen in Table 4.5, radiomic features alone are not enough to predict failure of IOL (the best AUC is 0.52 ± 0.03 when using a Random Forest classifier).

Table 4.9 highlights the importance of feature selection. Comparing experiment 2 with experiment 3 and experiment 5 with experiment 6, it can be seen that for most of the classifiers the AUC improves after feature selection, even though for some classifiers (MLP, SVM) it remains similar. Furthermore, the feature selection process allows us to understand which features are more relevant. When clinical and sonographic features are combined, 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, and weeks of pregnancy. While using only clinical data is not enough to predict IOL (table 4.1.1), adding this information to sonopraphical measurements helps improving the models. In the next experiment, combining clinical data, sonographic measurements and radiomics, the same five sonographic measurements are selected again, 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 suggests that radiomic features from transvaginal US can be useful for the prediction of IOL failure. However, it must be noted that the selected features are not necessarily the most important ones. When the dataset has two or more correlated features, from the point of view of the model any of them can be used as the predictor, with no preference for one over the other. But, once one of them is added to the set of selected features, the the importance of the other correlated features is significantly reduced since they do not add new information, and they will not be selected. Therefore, features that may actually have a high correlation with the outcome can be removed if they also have high correlation to other features that have already been selected. This is not an issue when we want to use feature selection to reduce the dimension of the dataset, since it makes sense to remove features that are mostly duplicated by other but can lead to the incorrect conclusion that one of the variables is a strong predictor while the others in the same group are unimportant. Figure 4.1 shows that some features have a very high correlation to other, so some of them will be removed and consider unimportant by the feature selection process.

The next experiment consisted in using a CNN to extract features from the image and from the selected ROI. The best performing classifier is a random forest when using the ROI and yields 0.57 ± 0.01 AUC. In general, it can be seen that classifiers perform better with the features extracted from the ROI than when using the entire image (table 4.6).

The low performance obtained from classifiers trained using only radiomic features and from classifiers trained using CNN features suggest that data extracted from the image alone is not predictive of failure of IOL. However, the results may improve by adding more training data. The hypothesis for the image analysis is

that changes in the micro-structure of the cervical tissue could result in changes in the speckle of the US image, and those changes are correlated with IOL failure. These changes in the pattern in the speckle are not even visible to a human observer and should be found by the classifiers, inferred from the training data. As seen in figure 3.1, from the 52 patients that a C-section, only in 30 cases the induction failed because the cervix was not mature or because the dilatation stopped after the cervix was mature. In other patients the causes for operative delivery were not related to a cervical problem. Thus, it is probable that there are no changes in the cervical tissue of these patients. That would mean that a different speckle pattern would be present only in 30 cases, and that is a very small number for the machine learning classifiers to learn. Therefore, recruiting more patients with a failure of induction related to a cervical motive might improve the results from the image-based analysis, and would also reduce the class imbalance.

Regarding the difference in performance of the machine learning classifiers, in general meta-estimators based on trees (random forests, extra trees) show the best performance. These classifiers perform better in noisy datasets and with high-dimensional feature spaces; random forest classifiers have been widely used for medical image analysis for that reason. The main disadvantage is the difficulty to interpret the results, as compared with a traditional decision tree. On the other hand, the worst performing classifiers for most of the experiments are SVM classifiers. This is probably because only two kernels have been used (RBF and sigmoid). Finding a good kernel for SVM classifiers is very crucial, otherwise the data will not be separable in the transformed space, so probably a more exhaustive parameter selection process should be performed to find the adequate settings for SVM classifiers to work.

The dataset used is small (182 patients) and highly imbalanced. We tried to alleviate it using SMOTE. We also dealt with the problem known as the *curse of dimensionality*, which in the context of machine learning means that the number of features is very high compared to the number of instances in the database (104 features in one of the experiments). Learning is very difficult for some classifiers in these conditions. While the results obtained are good (we achieve a maximum AUC of 0.75, with 69% sensitivity and 71% specificity when using a random forest classifier) with radiomic, clinical, and sonographic features, these values are still far away from the accuracy needed in clinical practice. Higher AUCs for the prediction of IOL failure using transvaginal US are reported in the state of the art (see Table 2.1) but the datasets used are small and imbalanced, containing very few examples of failure of induction. For instance, [9] report an AUC of 0.83 using local binary patterns, but their dataset only has 9 patients that ended up needing a C-section. A bigger study [56], which did not incorporate imaging data, only clinical data extracted from clinical reports, achieved an AUC of 0.867 with a patient cohort of 10,487 cases. Obtaining a large dataset for an image-based analysis is more challenging since many times the US images are not saved or even performed before IOL, but it would be necessary in order to compare the results of clinical data alone with image-based analysis to establish whether a transvaginal US should be recommended before IOL.

5.2 Prediction of preterm birth

Results for the prediction of preterm birth are presented in Section 4.2. In this case, the clinical data was not available, so the analysis is only image-based. The feature vectors used are smaller (77 radiomic features, 10 CNN features) than for the previous clinical scenario and the database is bigger (383 patients) and more balanced. Therefore, the results are better in general (except for SVM), and the differences between experiments and classifiers are smaller. Results above or close to 0.7 AUC are achieved for many of the models.

An AUC of 0.768 ± 0.005, with 71% sensitivity and 69% specificity, is achieved using a random forest classifier with radiomic features, after removing redundant features based on correlation. This can be considered as the best result for radiomic analysis, since the model is more balanced. In this case, the results are very similar with and without feature selection.

Using CNN features, 0.753 ± 0.008 AUC with 70% sensitivity and 73% specificity is achieved using a Random Forest classifier. Even though the Gaussian naive Bayes classifier obtains better results in terms of AUC, the RF is chosen as the best one because it has more balanced results in terms of sensitivity and specificity, which is preferred from a clinical point of view.

These results are close to the state of the art (0.77 AUC reported by [11] using local binary patterns) but still far from being useful in a clinical practice.

Causes for preterm birth are urinary infections, diabetes, renal or cardiac disease, eclampsia and preeclampsia, streess, smoking... However for many patients (about 50% of the cases) that end up in spontaneous preterm birth, the cause cannot be established. Therefore, it is difficult to assess whether in all the training and test samples of our database the cause of the preterm birth is related to alterations that appear in the US image.

5.3 Conclusions and future work

Good results have been obtained for image-based analysis, both with radiomics and CNN features (0.768 ± 0.005 and , 0.753 ± 0.008 respectively), with transvaginal US from the cervix for the prediction of preterm birth, without adding any other clinical information to the model, which is consistent with the hypothesis that transvaginal US can reflect changes that occur in the cervical tissue. In the case of IOL, an AUC of 0.68 ± 0.009 is obtained with the sonographic features such as cervical lenght, diameter, etc., proving that transvaginal US images provide valuable information for the prediction of IOL failure. While radiomics and CNN features alone do not have a good performance, the combination with clinical data and sonographic measurements yields 0.75 ± 0.02 AUC. In this scenario, image-based analysis alone is not enough, but the difference in the performance of these features between the two clinical scenarios could be explained by the difference in the datasets, given that the preterm birth dataset is bigger and more balanced than IOL dataset.

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 techniques for obtaining measurements, such as cervical length and cervical angle, automatically from the US images.

Regarding prediction of preterm birth, future research should address the combination of clinical data with image-based analysis. A more extensive database should be created and curated, highlighting the cause of preterm delivery. If enough number of images is available, the multi-label classification could be used instead of binary classification to account for the difference between late and moderate preterm birth (between 32 and 36 weeks of gestation), extremely preterm (less than 28 weeks) and very preterm (28 to 32 weeks). Another approach could be a regression to estimate the week of delivery from the images or clinical data. More extensive validation with a bigger database is needed as well to study the generalization of the models to data sets acquired by different obstetricians with different acquisition protocols or vendors.

Another limitation of the study is that for the image-based analysis the region of interest has been drawn manually. This requires a lot of work from the radiologists and introduces inter observer variability. Future works should explore automatic segmentation techniques to obtain the ROI.

End-to-end deep learning networks could be applied to the classification of transvaginal US images for both clinical scenarios if a larger database is available, allowing the network to automatically extract the most relevant features from the images instead of using a set of manually designed radiomic features or a pretrained network, which has been trained on natural images (ImageNet) instead of US images.

The results of our work have been accepted for publication in the *MICCAI Workshop on Perinatal, Preterm and Paediatric Image analysis (Shenzhen, China)*, the author accepted manuscript version of the paper is included in Appendix I.

Chapter 6

Conclusion

In this project we have applied machine learning and image analysis techniques to build predictive models in two clinical scenarios: preterm delivery and induction of labor. Transvaginal US images are widely available in hospitals and clinics, not expensive and fast, however, they are usually only visually examined by the radiologists and then stored in the hospital PACS, sometimes, they are not even stored. In this preliminary study, we assess whether applying artificial intelligence to these images could provide valuable information to the clinicians, improving patient care without significant additional costs.

Two different approaches have been explored for image analysis: extraction of radiomic features and CNN extracted features. Several machine learning classifiers have been trained to build models and the results obtained are close to the state-of-art.

For prediction of failure of induction of labor, clinical data and sonographic measurements manually extracted from the images were included as well, leveraging the imaging and clinical data.

Prediction of preterm birth is usually done by measuring the cervical length in the transvaginal US, but it does not report any information about the compression or structural and histological changes of the tissue. Transvaginal US images are performed in routine examinations for pregnancy follow up, but no further analysis is made. Our results show that cervical US images can help can help estimating the risk of preterm birth using texture analysis and machine learning. We achieved 0.768 $\pm$ 0.005 AUC with 71% sensitivity and 69% specificity. The correct identification of women who are at risk of preterm delivery could help the practitioners give a more personalized treatment, perform additional follow up. or recommend interventions to prevent preterm birth.

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's Score, has been found to be subjective and inconsistent. The results presented in this project agree with previously reported results [6, **?**] 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 com-

bination 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.

Future work should perform an extensive technical evaluation of the models with a larger and more diverse database, including US images from different vendors. Eventually, a clinical evaluation should establish the relevance of the results for the clinical practice.

Chapter 7

Appendix I: Author accepted manuscript version of the paper for MICCAI Workshop on Perinatal, Preterm and Paediatric Image analysis

Prediction of failure of induction of labor from ultrasound images using radiomic features

M. Inmaculada García Ocañ[1,2], Karen López-Linares Román[1,2], Jorge Burgos San Cristóbal[3], Ana del Campo Real[3], 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

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 childbirth 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]. 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 [11, 10], US image analysis is also used to predict failure of induction of labor [11, 10]. 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, 8, 4], 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 [11, 10]. 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 720x960 pixels with a pixel spacing of 0.11. An expert obstetrician manually selected a region of interest (ROI) delimiting the upper part of the

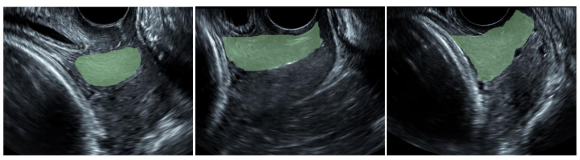

Fig. 1. Transvaginal ultrasound images of three patients and the selected region of interest for radiomic analysis.

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.

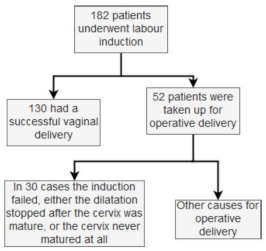

Fig. 2. Database composition.

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 $|ij| \leq \alpha$. 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 *sci-kit 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 2.4 summarizes the parameters used to train each classifier. Furthermore, to provide robustness to the results, we used a 10-fold cross-validation approach.

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

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.

3 Results and discussion

Table 2 summarizes the results obtained for each of the experiments described in Section 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 $\pm$ 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

	Experiment 1		Experiment 2		Experiment 3		Experiment 4	
Classifier	**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

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.

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 sonopraphical measurements helps improving the models, according to the results from experiments 2 and 3. In experiment 4, five sonographic measurements are again selected, as well as some clinical variables

	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

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

(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., Sobh, S.M., Belal, D.S., Omran, E.F., Hassan, A., Saad, A.H., Afifi, M.M., Nada, A.M.: 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., Juli, C., Murillo-Bravo, C., Coronado, D., Gratacs, 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., Perez-Moreno, A., Migliorelli, F., Triginer, L., Cobo, T., Bonet-Carne, E., Gratacos, E., Palacio, M.: 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. Scientific Reports **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., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R.G., Fillion-Robin, J.C., Pieper, S., Aerts, H.J.: Computational radiomics system to decode the radiographic phenotype. Cancer Research **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., Verhoeven, C., Brinkhorst, S., van der Post, J., Pajkrt, E., Opmeer, B., Mol, B.: The Bishop Score as a Predictor of Labor Induction Success: A Systematic Review. Amer 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: 2016 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., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. J Mach Learn Res **12**, 2825–2830 (2011)

Bibliography

[1] Laura J Brattain, Brian A Telfer, Manish Dhyani, Joseph R Grajo, and Anthony E Samir. Machine learning for medical ultrasound: status, methods, and future opportunities. *Abdominal Radiology*, 43(4):786–799, 2018.

[2] Philippe Lambin, Emmanuel Rios-Velazquez, Ralph Leijenaar, Sara Carvalho, Ruud GPM Van Stiphout, Patrick Granton, Catharina ML Zegers, Robert Gillies, Ronald Boellard, André Dekker, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer*, 48(4):441–446, 2012.

[3] Labor induction - american college of obstetricians and gynecologists, *https://www.acog.org/Patients/FAQs/Labor-Induction*.

[4] Diny Kolkman, Corine Verhoeven, Sophie Brinkhorst, Joris van der Post, Eva Pajkrt, Brent Opmeer, and Ben Mol. The Bishop Score as a Predictor of Labor Induction Success: A Systematic Review. *Amer J Perinatol*, 30(08):625–630, January 2013.

[5] M. Kehila, H.S. Abouda, K. Sahbi, H. Cheour, and M. Badis Chanoufi. Ultrasound cervical length measurement in prediction of labor induction outcome. *NPM*, 9(2):127–131, 2016.

[6] Maia Brik, Silvia Mateos, Irene Fernandez-Buhigas, Paloma Garbayo, Gloria Costa, and Belen Santacruz. Sonographical predictive markers of failure of induction of labour in term pregnancy. *Journal of Obstetrics and Gynaecology*, 37(2):179–184, February 2017.

[7] Neha Bajpai. Manipal Cervical Scoring System by Transvaginal Ultrasound in Predicting Successful Labour Induction. *JCDR*, 2015.

[8] Pablo Vasquez Obando, A. Nestor Arana, Alberto Izaguirre, and Jorge Burgos. Labor induction failure prediction using gabor filterbanks and center symmetric local binary patterns. In *2017 IEEE 37th Central America and Panama Convention (CONCAPAN XXXVII)*, page 1–5. IEEE, Nov 2017.

[9] Vasquez O. Pablo, Nestor Arana, Alberto Izaguirre, and Jorge Burgos. Labor induction failure prediction based on b-mode ultrasound image processing using multiscale local binary patterns. In *2016 International Conference on Optoelectronics and Image Processing (ICOIP)*, page 25–29. IEEE, Jun 2016.

[10] Hannah Blencowe, Simon Cousens, Doris Chou, Mikkel Oestergaard, Lale Say, Ann-Beth Moller, Mary Kinney, and Joy Lawn. Born too soon: the global epidemiology of 15 million preterm births. *Reproductive health*, 10(1):S2, 2013.

[11] N. Baños, A. Perez-Moreno, C. Julià, C. Murillo-Bravo, D. Coronado, E. Gratacós, J. Deprest, and M. Palacio. 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, May 2018.

[12] Helen Feltovich and Lindsey Carlson. New techniques in evaluation of the cervix. *Seminars in Perinatology*, 41(8):477–484, December 2017.

[13] Xavier P. Burgos-Artizzu, Álvaro Perez-Moreno, David Coronado-Gutierrez, Eduard Gratacos, and Montse Palacio. Evaluation of an improved tool for non-invasive prediction of neonatal respiratory morbidity based on fully automated fetal lung ultrasound analysis. *Scientific Reports*, 9(1):1950, Dec 2019.

[14] Núria Baños, Alvaro Perez-Moreno, Federico Migliorelli, Laura Triginer, Teresa Cobo, Elisenda Bonet-Carne, Eduard Gratacos, and Montse Palacio. Quantitative Analysis of the Cervical Texture by Ultrasound and Correlation with Gestational Age. *Fetal Diagn Ther*, 41(4):265–272, 2017.

[15] A. Fruscalzo, A. Londero, C. Fröhlich, M. Meyer-Wittkopf, and R. Schmitz. Quantitative Elastography of the Cervix for Predicting Labor Induction Success. *Ultraschall in Med*, 36(01):65–73, February 2014.

[16] F. Migliorelli, C. Rueda, M. A. Angeles, N. Baños, D. E. Posadas, E. Gratacós, and M. Palacio. Cervical consistency index and risk of cesarean delivery after induction of labor at term: Cervical consistency index and induction of labor. *Ultrasound in Obstetrics Gynecology*, 53(6):798–803, Jun 2019.

[17] Akram M. Al-Adwy, Sherin M. Sobh, Doaa S. Belal, Eman F. Omran, Amr Hassan, Ahmed H. Saad, Mai M. Afifi, and Adel M. Nada. 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, April 2018.

[18] Joanna Ivars, Charles Garabedian, Patrick Devos, Denis Therby, Sabine Carlier, Philippe Deruelle, and Damien Subtil. Simplified Bishop score including parity predicts successful induction of labor. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, 203:309–314, August 2016.

[19] Christopher Paul Howson Mary Kinney Joy Lawn Althabe, Fernando and World Health Organization. *orn Too Soon: The Global Action Report on Preterm Birth*. 2012.

[20] Liran Hiersch, Nir Melamed, Amir Aviram, Ron Bardin, Yariv Yogev, and Eran Ashwal. Role of cervical length measurement for preterm delivery prediction in women with threatened preterm labor and cervical dilatation. *Journal of Ultrasound in Medicine*, 35(12):2631–2640, Dec 2016.

[21] Syun-ichi Yamaguchi, Yoshimasa Kamei, Shiro Kozuma, and Yuji Taketani. Tissue elastography imaging of the uterine cervix during pregnancy. *Journal of Medical Ultrasonics*, 34(4):209–210, Dec 2007.

[22] E. Hernandez-Andrade, S. S. Hassan, H. Ahn, S. J. Korzeniewski, L. Yeo, T. Chaiworapongsa, and R. Romero. Evaluation of cervical stiffness during pregnancy using semiquantitative ultrasound elastography: Cervical elastography during pregnancy. *Ultrasound in Obstetrics Gynecology*, 41(2):152–161, Feb 2013.

[23] M. Parra-Saavedra, L. Gómez, A. Barrero, G. Parra, F. Vergara, and E. Navarro. Prediction of preterm birth using the cervical consistency index. *Ultrasound in Obstetrics Gynecology*, 38(1):44–51, Jul 2011.

[24] Sleiman R. Ghorayeb, Luis A. Bracero, Matthew J. Blitz, Zara Rahman, and Martin L. Lesser. Quantitative ultrasound texture analysis for differentiating preterm from term fetal lungs: Fetal lung texture analysis with ultrasound. *Journal of Ultrasound in Medicine*, 36(7):1437–1443, Jul 2017.

[25] S Rasoul Safavian and David Landgrebe. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3):660–674, 1991.

[26] Cesar Ferri Ramirez Jose Hernandez Orallo, M.Jose Ramirez Quintana. *INTRODUCCIÓN A LA MINERÍA DE DATOS*. 2004.

[27] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[29] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

[30] Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[31] Sankar K Pal and Sushmita Mitra. Multilayer perceptron, fuzzy sets, and classification. *IEEE Transactions on neural networks*, 3(5):683–697, 1992.

[32] Mark A Hall and Lloyd A Smith. Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper. In *FLAIRS conference*, volume 1999, pages 235–239, 1999.

[33] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[34] Afsaneh Jalalian, Syamsiah BT Mashohor, Hajjah Rozi Mahmud, M Iqbal B Saripan, Abdul Rahman B Ramli, and Babak Karasfi. Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review. *Clinical imaging*, 37(3):420–426, 2013.

[35] Neil Joshi, Seth Billings, Erika Schwartz, Susan Harvey, and Philippe Burlina. Machine learning methods for 1d ultrasound breast cancer screening. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 711–715. IEEE, 2017.

[36] Anton S Becker, Michael Mueller, Elina Stoffel, Magda Marcon, Soleen Ghafoor, and Andreas Boss. Classification of breast cancer in ultrasound imaging using a generic deep learning analysis software: a pilot study. *The British journal of radiology*, 91(xxxx):20170576, 2018.

[37] Mazin Abed Mohammed, Belal Al-Khateeb, Ahmed Noori Rashid, Dheyaa Ahmed Ibrahim, Mohd Khanapi Abd Ghani, and Salama A Mostafa. Neural network and multi-fractal dimension features for breast cancer classification from ultrasound images. *Computers & Electrical Engineering*, 70:871–882, 2018.

[38] Xiaofeng Yang, Peter J Rossi, Ashesh B Jani, Hui Mao, Walter J Curran, and Tian Liu. 3d transrectal ultrasound (trus) prostate segmentation based on optimal feature learning framework. In *Medical Imaging 2016: Image Processing*, volume 9784, page 97842F. International Society for Optics and Photonics, 2016.

[39] Raman Preet Singh, Savita Gupta, and U Rajendra Acharya. Segmentation of prostate contours for automated diagnosis using ultrasound images: A survey. *Journal of computational science*, 21:223–231, 2017.

[40] Hui Xiong, Laith R Sultan, Theodore W Cary, Susan M Schultz, Ghizlane Bouzghar, and Chandra M Sehgal. The diagnostic performance of leak-plugging automated segmentation versus manual tracing of breast lesions on ultrasound images. *Ultrasound*, 25(2):98–106, 2017.

[41] Rosa-María Menchón-Lara and José-Luis Sancho-Gómez. Fully automatic segmentation of ultrasound common carotid artery images based on machine learning. *Neurocomputing*, 151:161–167, 2015.

[42] Fang Chen, Dan Wu, and Hongen Liao. Registration of ct and ultrasound images of the spine with neural network and orientation code mutual information. In *International Conference on Medical Imaging and Augmented Reality*, pages 292–301. Springer, 2016.

[43] Xiaofeng Yang and Baowei Fei. 3d prostate segmentation of ultrasound images combining longitudinal image registration and machine learning. In *Medical Imaging 2012: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 8316, page 83162O. International Society for Optics and Photonics, 2012.

[44] Fausto Milletari, Seyed-Ahmad Ahmadi, Christine Kroll, Annika Plate, Verena Rozanski, Juliana Maiostre, Johannes Levin, Olaf Dietrich, Birgit Ertl-Wagner, Kai Bötzel, et al. Hough-cnn: deep learning for segmentation of deep brain regions in mri and ultrasound. *Computer Vision and Image Understanding*, 164:92–102, 2017.

[45] Y Gao, Mohammad Ali Maraci, and J Alison Noble. Describing ultrasound video content using deep convolutional neural networks. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 787–790. IEEE, 2016.

[46] Christian F Baumgartner, Konstantinos Kamnitsas, Jacqueline Matthew, Sandra Smith, Bernhard Kainz, and Daniel Rueckert. Real-time standard scan plane detection and localisation in fetal ultrasound using fully convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 203–211. Springer, 2016.

[47] Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX, 2010.

[48] François Chollet et al. Keras. `https://keras.io`, 2015.

[49] Joost JM Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo JWL Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107, 2017.

[50] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[51] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[52] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[53] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[54] Sebastian Raschka. Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. *The Journal of Open Source Software*, 3(24), April 2018.

[55] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. [Online; accessed <today>].

[56] Cristina Pruenza, Mari´a Teulón, Luis Lechuga, Julia Díaz, and Ana González. Development of a predictive model for induction success of labour. *International Journal of Interactive Multimedia & Artificial Intelligence*, 4(7), 2018.