Ultrasound Image Processing in the Evaluation of Labor Induction Failure Risk

Pablo Vásquez
Signal Theory and Communication Department
Mondragon University

A thesis submitted for the degree of
Doctor of Philosophy

Mondragon 2017

Acknowledgements

After these three years working on this thesis, there is a lot of people I would like to thank. First I'd like to give thanks to God, to my family who has been supporting me all this time and of course to my two thesis advisers Nestor and Tito. I really appreciate the time you devoted to me and the patience you had during this work. Finally, I would like to thank to our research project counterparts at the Bilbao Obstetrics and Gynecology Research Center, especially to Dr.Jorge Burgos who made the proposal for this research and to all my group mates at Mondragon University.

Abstract

Labor induction is defined as the artificial stimulation of uterine contractions for the purpose of vaginal birth. Induction is prescribed for medical and elective reasons. Success in labor induction procedures is related to vaginal delivery. Cesarean section is one of the potential risks of labor induction as it occurs in about 20% of the inductions. A ripe cervix (soft and distensible) is needed for a successful labor. During the ripening cervical, tissues experience micro structural changes: collagen becomes disorganized and water content increases. These changes will affect the interaction between cervical tissues and sound waves during ultrasound transvaginal scanning and will be perceived as gray level intensity variations in the echographic image. Texture analysis can be used to analyze these variations and provide a means to evaluate cervical ripening in a non-invasive way.

Contents

Chapter 1

Introduction

Labor induction is defined as the artificial stimulation of uterine contractions for the purpose of vaginal birth. It is one of the most commonly practiced procedures in obstetrics, occurring in over 20% of pregnancies [25].

Approximately 135,000 out of 450,000 deliveries occurred in Spain in 2012 were induced, and the rate is increasing compared to the last decade. This trend is not particular to Spain but worldwide.

Labor induction is indicated when the maternal or fetal benefits from delivery outweigh the risks of prolonging the pregnancy. Indications for induction vary in seriousness and may be for medical, obstetrical or elective reasons.

Labor induction carries various risks, including: cesarean section, premature birth, fetal low heart rate, infections, umbilical cord problems and bleeding after delivery. Nowadays in Spain about 30% of deliveries are induced and 18% end in cesarean section [4].

It is likely that some of these unwanted outcomes result from intervening when the uterus and cervix are not ready for labor. For this reason, the evaluation of cervical ripening is a crucial step when planning a labor induction procedure.

Currently digital examination (examination of the cervix with the hand) is the only standard method to assess spontaneous cervical ripening, usually indicated by the Bishop score. This method being manual is more or less subjective and prone to errors and inter observer variability.

A more accurate evaluation of cervical ripening is desirable before labor induction process is started. It is known that the cervix tissues goes through remarkable changes along pregnancy. Collagen, the most abundant element in the cervical micro structure (about 85 %), is aligned and organized in the cervix of non-pregnant women and becomes progressively more disorganized during remodeling of the cervix as the pregnancy progresses in preparation for the delivery [7]. Besides

collagen changes, water content of the cervical tissues is also increased.

The aforementioned changes in cervical microstructrure and tissue hydration are expected to be reflected in changes in the image obtained from transvaginal ultrasound since the consistency of tissues affect their interaction with the sound waves.

The importance of developing tools that help in the cervical evaluation is high and of the interest for our health-care system providers. In this research work, we aim to develop image processing algorithms capable to assess cervical changes based on transvaginal ultrasound B-mode images. The importance of these tools lies in the fact that they could influence the reduction in the cesarean rate along with the associated hospitalization costs.

In this chapter, our goal is to present several concepts that are important in the problem formulation. We are going to review the labor induction process, the reasons, the risks and the problems associated with it. One important step before starting an induction is to assess the state of the cervix. A ripe cervix (soft and distensible) is needed for a successful labor. We are going to describe the cervix and its different stages and transformations along pregnancy. We also review current methods for assessing cervical status and agents to promote cervical ripening.

1.1 Labor induction

Induction of labor is common in obstetric practice. According to the most current studies, the rate varies from 9.5 to 33.7 percent of all pregnancies annually [21]. The decision to start a labor induction procedure can be based on medical conditions of the mother and fetus, or sometimes influenced by other factors apart from obstetrical ones. Some important medical reasons occur when the mother is post term (42 weeks of pregnancy), suffering from renal disease, hypertension or diabetes.

A fetus with problems of growth restriction, or infection could also constitute a serious condition motivating labor induction, in these cases we talk about *indicated* labor induction. Motivations for labor induction apart from the obstetrical ones (e.g. for matter of preference or convenience) are termed *elective*. Among elective reasons, we can mention for example specialist services availability or psychosocial indications. Table 1.1 shows a summary of common reasons for labor induction.

A factor having a major influence on the success of a labor induction procedure

Table 1.1: Commonly quoted indications for inducing labor. Reproduced from reference [14]

Fetal reasons	Maternal reasons	Non MedicalReasons
Clinically Evident	Deteriorating health	Specialist services availability
Growth restriction	-renal	x-matched blood
Abruptio placentae	-hypertension	anaesthesia
Polyhydramnios	-psychological	fetal surgery
Red-cell alloimmunisation	-malignancy	Partners availability
Diabetes mellitus	-autoimmune diseases	
Unstable fetal lie	Diabetic fragility	
Fetal infection	Coagulopathy	
Macrosomia	Intra-uterine infection	
	Antepartum haemorrhage	
	Polyhydramnios	
	Discomfort	
Statistically anticipated		
Growth restriction	Hypertension	
Prolonged pregnancy	Feto-pelvic disproportion	
Previous obstetric history	Short maternal height	
Ruptured membranes	Intra-uterine fetal death	
Breech presentation	Prior caesarean section	
Diabetes mellitus	Ruptured membranes	
Antepartum haemorrhage		
Multiple pregnancy		
Red-cell alloimmunisation		

is the state of the uterine cervix [14].

Once labor induction has been decided, the next step is to evaluate the degree of readiness of the woman's uterus for labor. If the cervix is not ripe then the probability of a successful labor is small. In this context, successful means vaginal delivery and when it is not achieved despite the application of ripening agents, we talk about labor induction failure.

1.2 Cervical Ripening and Labor Induction

In this section we are going to review some concepts related to cervix, its ripening process and the methods currently in use to induce such a condition for improving outcomes from labor induction.

1.2.1 The Cervix.

The cervix or uterine cervix is the lower fibromuscular portion of the uterus that projects into the vagina (Figure 1.1) and is a unique part of the female anatomy of mammals. This opening or hole lets the blood out of the uterus during menstrua-

tion. Also let sperm enter the uterus and fallopian tubes

There are two main portions of the cervix: The part of the cervix that can be seen from inside the vagina during a gynecologic examination is known as the *ectocervix*. An opening in the center of the ectocervix, known as the *external Ostium or os* , opens to allow passage between the uterus and vagina.

The *endocervix*, or *endocervical canal*, is a tunnel through the cervix, from the external os into the uterus. The cervix also produces cervical mucus that changes in consistency during the menstrual cycle to prevent or promote pregnancy.

When childbirth is approaching, the cervix begins to thin or stretch (efface) and open (dilate) in preparation for the passage of the baby through the birth canal or vagina (Figure 1.2).

1.2.2 Cervical Ripening

Cervical ripening is the term used to describe the transformation in tissue microstructure of the cervix occurring during pregnancy that leads to its progressive softening and distensibility. From a state of alignment and organization (collagen the most abundant component of cervical microstructure) goes to a progressively more disorganized state as pregnancy progress.

At the end of pregnancy the hyaluronic acid content is incremented in the cervix. As a result an increase in water molecules that intercalate among the collagen fibers occurs. The amount of dermatan sulfate decreases, leading to reduced bridging among the collagen fibers and a corresponding decrease in cervical firmness.

During the first month of pregnancy, a slow but progressive collagen reorganization phase begins. Near birth, a second phase includes a rapid and marked reorganization of the micro-structure [7] causing macro structural changes (including cervical shortening). The active dilatation during labor is the third phase

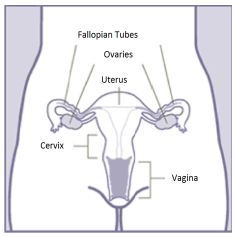


Figure 1.1: Female reproductive organs, showing the cervix.

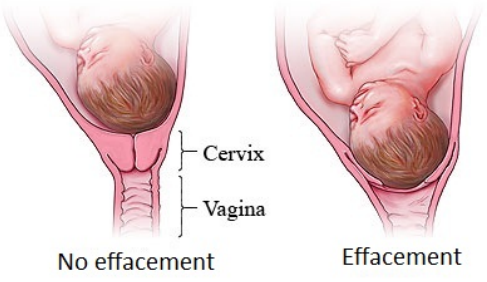


Figure 1.2: Cervical Effacement. Left: Cervix without effacement and right with effacement.

and the fourth includes micro-structure recovery. Cervical remodeling in short, consists of four overlapping phases [13]:

1. Softening alone. Reorganization of collagen.
2. Ripening (Softening with effacement, dilation and change in position)
3. Dilation in response to the contractions
4. Postpartum repair.

The microstructure of the human cervix during pregnancy is not known in depth due to the difficulty of performing invasive studies. There are however some works like [12, 22] in which using high resolution images of the micro-structure obtained by second harmonic generation imaging (SHG), has confirmed that the cervix possesses 3 layers of collagen, including a circumferential layer and two flanking longitudinal layers. It has been further discovered that cervix collagen (especially in non-pregnant state) behaves anisotropically (this means that tissue properties are not the same along different directions).

In summary the two main characteristics of cervical ripening are the increase in water content of the tissues and the disorganization of collagen both leading to an overall softening of the cervix.

1.3 Cervical ripening agents

To favor cervical changes needed for successful labor, several tools have been developed. Both pharmacological and mechanical methods are in use today as a part of clinical protocols for labor induction.

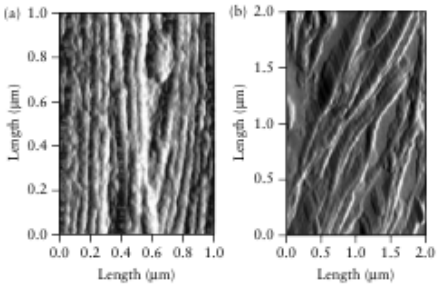


Figure 1.3: Atomic force microscopy cervical tissue images of a non pregnant rat (left), showing packages highly organized collagen bundles and a cervix on day 21 of pregnancy (right), showing disorganized collagen fibril spacing. The packages bundle shown in the images are < 0.1 microns. Rats typically give birth at day 21. From reference [15]

1.3.1 Mechanical Methods

Membrane Sweeping In this technique the physician insert his or her finger (wearing gloves) beyond the os and rotates it. This movement is aimed to separate the amniotic sac from the uterus wall.

It is known that this procedure promote the production of local prostaglandins which help in the ripening process of the cervix and have the potential to initiate labor and reduce pregnancy duration.

Intracervical balloon catheter placement Another procedure adopted for routine induction of labour involves transcervical application of a balloon catheter (Foley, Cook type). The balloon-tipped catheter is inserted beyond the cervical opening (figure 1.4) . Saline injected through the catheter expands the balloon, causing the cervix to widen. Such a catheter appears to induce labor not only through direct mechanical dilation of the cervix but also by stimulating endogenous release of prostaglandins

1.3.2 Pharmacological Agents

Prostaglandins Prostaglandins are hormones that helps to ripe the cervix, they are normally used when the cervix is not favorable. They offer the advantage to also promote myometrial conctractility. The most commonly employed prostaglandin in obstectrics is Dinoprostone (PGE2). Administration of prostaglandins can be done via intravaginal or intracervical. Some reported complications observed in patients treated with PGE2 have been

tachysystole and hyperstimulation of the uterus

Oxytocin Oxytocin is an agent that induce uterine contractions and must be administrated with care. It is administered when the cervix is favorable (Bishop Score > 6). Some risk associated wit Oxytocin are uterine hyper stimulation and fetal heart rate increase therefore monitoring of fetus state is always recommended.

1.4 Methods for Cervix Evaluation

For the purpose of determining whether cervix is ready for labor several methods have been developed. Directly or indirectly, these methods attempt to quantify the various changes that occur in the cervix during pregnancy: shortening, dilatation and softening.

1.4.1 Digital Examination

Digital examination is one of the oldest techniques, being the Bishop's score the most well known. The Bishop score is a quantitaive means to determine the inductibility or status of the uterine cervix in a pregnant woman, based on five parameters: dilatation, effacement, station, cervical consistency and position (see table 1.2).

A drawback of physical examination is that it suffers from a large variation between different examiners, do not provide information about internal os and has some risks such as infections due to its invasive nature. Several studies have also reported that the Bishop score could be a poor predictor of outcome of labor induction when the cervix is unfavorable [23].

Most authors defines an unfavorable cervix as having a Bishop score between

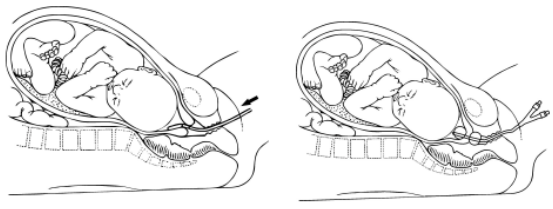


Figure 1.4: Catheter placement.

4 and 6, here we adopt BS < 6 as threshold. A Bishop score < 6 indicates an unfavorable cervix which may require a prelabor cervical ripening agent. According to the Modified Bishop's pre-induction cervical scoring system, effacement has been replaced by cervical length in cm, with scores as follows: 0 > 3cm, 1>2c m, 2> 1cm, 3 > 0cm.

Table 1.2: Bishop Score. A score > 6 indicates a good chance of vaginal delivery.

	0	1	2	3
Dilation (cm)	0	1-2	3-4	5-6
Effacement (%)	0-30	40-50	60-70	80
Station	-3	-2	-1 to 0	+1 to +2
Cervical Consistency	Firm	Medium	Soft	
Position of Cervix	Posterior	Mid	Anterior	

1.4.2 Chemical Markers

Fetal fibronectin Fetal fibronectine (FFN) is a glycoprotein that binds the amniochorion to the decidua (i.e fetal sac to the uterine lining) and is released into cervicovaginal fluid in response to inflammation or separation of amniochorion from the decidua.

The concentration of fetal fibronectin in cervical transudate correlates with the result of the induction of labor in concentrations of more than 50 mg/ml associated with a favorable cervix [14].

The presence of fetal fibronectin in vaginal secretion of women undergoing labor induction has been associated with a lower cesarean rate. The absence of this protein in cervicovaginal secretions predicts prolongation of pregnancy.

1.4.3 Ultrasound Evaluation

Several studies involving cervix evaluation by means of clinical ultrasound derived parameters have been published. Metrics derived from ultrasound scanning analyses the different aspects of ripening, i.e softening, shortening and dilation.

There is however controversy about the utility of such parameters. Some studies claim that cervical length is useful in predicting the likelihood of vaginal delivery within 24 hours of induction [18] and that it was related to type of delivery in women with Bishop score ≤ 5, [10] however in [6], it is concluded that neither

fetal fibronectine nor transvaginal ultrasound examination has been shown to be superior to Bishop score. More recently, in [9] it is stated that for cervical ripeness assessment cervical length showed higher reliability than the Bishop score. This incongruence is probably due to a lack of consensus regarding definitions like induction failure, or patient inclusion criteria.

1.4.3.1 Transvaginal Ultrasound

Cervix assessment by ultrasound is usually performed by means of transvaginal ultrasound (TVU). A transvaginal ultrasound transducer is cylindrical in shape (see Figure 1.5) and is inserted through the vaginal canal. This type of ultrasound provides better images with higher quality and more detail and is often used to confirm the diagnosis of lesions found with conventional abdominal ultrasound.

Transvaginal ultrasound allows clearly and consistently the visualization of the cervix and the internal os (internal Ostium or Os) providing an advantage over transabdominal ultrasound evaluation. This last method may not be reliable due to the mother habitus (constitution or body build), cervix position, degree of fullness of the bladder and the darkening effect of the fetus.

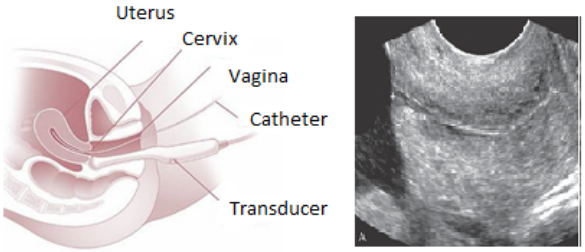


Figure 1.5: Left: Examination with transvaginal ultrasound, right: B-mode image of a normal cervix showing an internal T-shaped closed os.

1.4.3.2 Ultrasound Parameters to Assess Cervical ripening

Cervical length The most commonly used parameter is the measurement of cervical length. This method first described in 1996 [2] showed an inverse relationship between cervical length measured by ultrasound during pregnancy and the frequency of preterm birth.

It has been shown that measurement of cervical length by ultrasound (usually transvaginal) is more effective than digital examination of the cervix [24] for predicting preterm birth. The length is measured from the internal

os along the endocervical canal until the external os and is the most reproducible and reliable measurements. If the channel is curved, cervical length can be measured in a straight line between the inner and outer os or as the sum of two straight lines that follow the curve.

Dilation of internal os (Funneling) The dilation of internal os in the uterine cervix has been also proposed as a indicator of uterus being prepared for labor, although mostly used in preterm birth risk analysis [3]. The funneling percentage (figure 1.6) is defined as:

$$\%F = \frac{A}{A+B} \tag{1.1}$$

Where A is the funnel length and B is the functional length of the cervix.

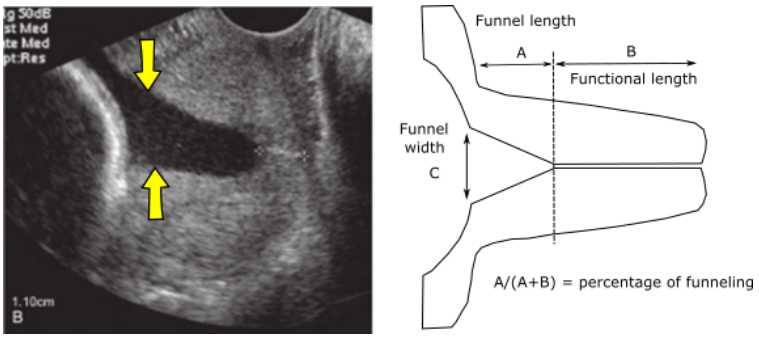


Figure 1.6: Measuring of funneling. Left: US image showing a dilated internal os (arrows). To the right (A) is the funnel length (B) is the functional cervix length and (C) is the funnel width.

Cervical gland area (CGA) As pregnancy progresses the cervical tissue becomes full of mucus producing glands. This mucus block the external os (known as mucus plug) and is useful in the prevention of infections. The gland area has been described as the sonographically hyperechoic or hypoechoic zone surrounding the cervical canal.

The presence of these glands and mucus has been introduced as a new marker. In [1] it was shown that the decrease or absence of glandular area (figure 1.7) may be an early indicator of cervical insufficiency.

Cervical Consistency Index The cervical consistency index (CCI) described in [19] is another parameter designed recently to evaluate the ripening of the cervix. For the calculation of the index, the length of the cervix is measured in the

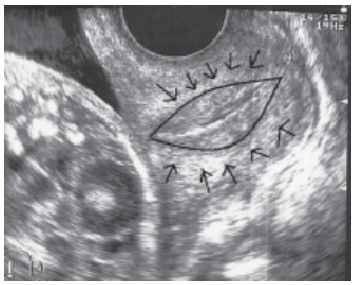


Figure 1.7: The presence of cervical glands is a normal finding during pregnancy; their absence may constitute a predictive sign of preterm birth. The glandular area is the region pointed by arrows. Reproduced from [13]

usual way (distance AP) and then is measured after applying gently pressure with the transducer up to the point where the cervix is not shorter (AP'), the ratio of two lengths, multiplied by one hundred is the value of the CCI (equation 1.2).

$$CCI = \frac{AP'}{AP} 100. \tag{1.2}$$

The CCI according to its authors, has a high sensitivity for the prediction of spontaneous preterm birth before 34 weeks and can be easily determined, showing a high level of repeatability

1.4.3.3 Additional US based Methods

There exist several methods that are in use for cervical evaluation purposes but not in the context of labor induction and are mostly experimental. Most of these methods have been proposed for the diagnosis of cervical incompetence or insufficiency ,which is a condition that happens when the uterine cervix ripes too early and can be cause for prematurity.

Elastography Elastography is an ultrasound-based technique used to evaluate the consistency of soft tissues by means of compression applied in vivo using conventional ultrasound systems with specialized software. It is a noninvasive method in which stiffness images of soft tissue are used to detect or classify mass. The main issue with elastography of the cervix is the lack of reference tissue for comparison, also elastographic image is sensitive to pressure changes. Some studies have reported the use of elastography for cervix evaluation but it is acknowledged that this method is still in its infancy [17].

Quantitative Ultrasound The need to develop measurable parameters or numerical descriptors for characterization of body tissue by ultrasound has led to what is known as quantitative ultrasound (QUS). The QUS parameters are proposed to characterize the conditions of tissue based on the assumption that such a disease changes the acoustic properties of the medium (tissue). The most common parameters in the literature are the speed of sound propagation, attenuation coefficients and scattering, the average spacing of scatterers (periodicity) and the size of the scatterers.

In [16] an increase in collagen content of the cervix, is reported as pregnancy progress in a experiment with rats. While collagen increases, the concentration decreased as fibers disorganized and more space was created between the collagen fibers.

In their study it was concluded that scatterer diameter varied little during pregnancy whereas acoustic concentration decreased what suggest that the scattering size did not markedly change, but the concentration of the scatterers in the cervix tissue do so as pregnancy progressed.

Recently there have been several experiments to study the changes in cervical tissues to ultrasound, for example in [15], 40 women were treated with ultrasound cervical recognition for the purpose of developing a method for evaluating the attenuation of ultrasound in the cervix during pregnancy using a clinical ultrasound system.

Summarizing this section we can say that evaluation of cervical ripening by ultrasound is focused on the various changes experienced by the cervix during pregnancy. Cervical ripening can be described as a 3-step process that should occur in sequence: softening, effacement and dilation of the cervix. The different methods can be sorted according to the parameter evaluated:

- **Cervical shortening:** Cervical length.
- **Glands and cervical mucus plug:** Glandular area (CGA)
- **Softening:** CCI, attenuation coefficient, QUS parameters, elastography.
- **Dilation:** Funneling.

1.5 Predictive Value of Bishop score and Ultrasound derived methods.

Bishop's score (BS) remains, according to our review, as the standard method for cervix evaluation. A predictive value outperforming that of BS is required for this tool to be useful for clinical use. Making a comparison regarding error performance of the different methods for cervix evaluation is difficult because the research works reviewed so far do not use exactly the same settings, same population, objectives and definitions. Usually, in medical application, to express the ability of a classifying scheme researchers use parameters like sensitivity, specificity, accuracy, positive predictive value, etc.

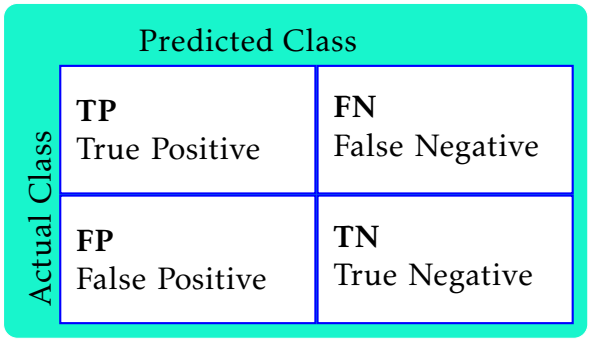


Figure 1.8: Confusion matrix: TP is also known as hit, TN as correct rejection, FP as false alarm or Type I error and FN miss or Type II error. P_{PV} stands for Positive Predictive Value (also Precision), T_{PR} True Positive Rate (Sensitivity) T_{NR} , True Negative Rate (Specificity)and Negative Predictive Value (False Positive Rate)

The definition of these parameters depends upon the different entries of what is known as a confusion matrix in predictive analysis. That matrix is an array with two rows and two columns (see figure 1.8) that reports the number of false positives, false negatives, true positives, and true negatives regarding to a classification task where the goal is to predict an outcome from a process. For example in the screening process for a disease, the test outcome can be positive (predicting that the person has the disease) or negative (predicting that the person does not have the disease).

Receiver operating characteristic (ROC) curves are also frequently used in algorithm performance evaluation. These are 2D plots where sensitivity is plotted against 1-specificity (False Positive Rate). When we require a unique value the F-score can be used as a single measure of performance of the test. The F-score is the harmonic mean of precision (also called PPV or Positive Predictive Value) and

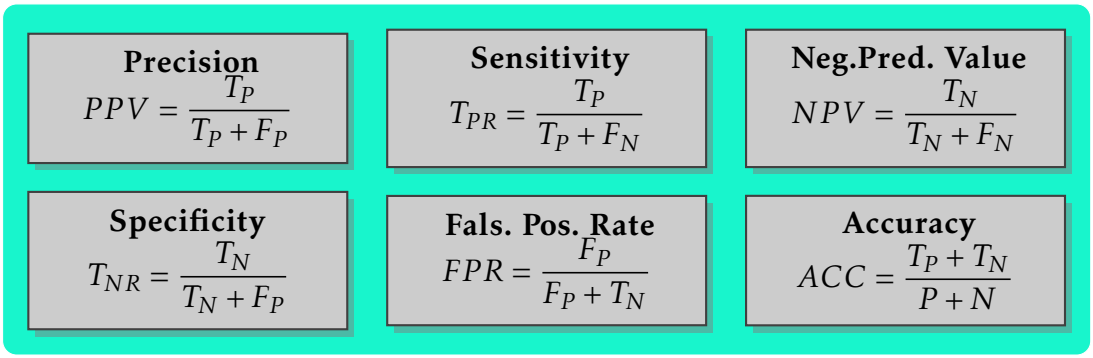


Figure 1.9: The different parameters used to describe error performance. P_{PV} stands for Positive Predictive Value (same as Precision), True Positive Rate T_{PR} (Sensitivity), True Negative Rate T_{NR} (same as Specificity), Negative Predictive Value N_{PV} (False Positive Rate)

Table 1.3: Successful induction predictive values reported in several works.

Method	Condition	Sensitivity	Specificity	P_{PV}	N_{PV}
Bishop score [8]	>5	0.66	0.49		
Cervical length	>26mm	0.62	0.61	–	
Bishop score [18]	>3	0.58	0.77	–	
Cervical length	>28mm	0.87	0.71	—	
Bishop score [23]	>4	0.87	0.45	–	
Cervical length[11]	< 20mm	0.64	0.70	0.57	0.76
Cervical Area	< 100mm^2	0.64	0.70	0.57	0.76
Bishop score	> 5	0.62	0.57	0.46	0.71
Mean elastographic index	< 100	0.76	0.56	0.51	0.79
Cervical hard area	< 200mm^2	0.86	0.60	0.51	0.87

recall (Sensitivity,TPR):

$$F = 2 * \frac{PPV * TPR}{PPV + TPR}. \tag{1.3}$$

The current prediction capabilities of clinical ultrasound parameters and Bishop's score (BS) have been studied in several articles. In [8] a study of 179 women, BS is compared to cervical length (CL) they concluded that the Bishop score was not predictive of the delivery mode, but cervical length was. ROC curves were also constructed showing optimal values for BS (BS >5) and CL>25mm. Cervical elastography and several other parameters are compared to BS in [11] cervical length, cervical area, mean elastographic index, and cervical hard area using delivery within 24 hours to predict successful labor induction.

A study involving [10], 177 women who underwent induced labor, compared Bishop score to CL. The total cesarean rate was significantly lower in the group of women with short cervical length, and BS did not predict the type of deliv-

ery. In a similar study [20] transvaginal sonography examination at 37 weeks in 1571 singleton low-risk pregnancies was carried out. It was observed that in the pregnancies requiring induction for post-term the incidence of cesarean section for failed induction or failure to progress increased with cervical length.

In [18], 240 women participated in a study to examine the measured cervical length and the Bishop score and to compare the two measurements in the prediction of successful vaginal delivery within 24 h of induction. Examination of the different components of the Bishop score showed that only cervical length provided a significant contribution in the prediction of the likelihood of vaginal delivery.

To compare transvaginal ultrasound and digital cervical examination in predicting successful induction in posterm pregnancies, a group of 122 women at 41 or more weeks' gestation, immediately before labor were examined with ultrasound to measure cervical length, dilatation and cervical funneling in [5]. The study conclusion is that no ultrasound characteristic predicted outcomes, this conclusion agrees with the one in [23] that states that no improvement in outcome prediction is gained by ultrasound features.

The values of the different parameters for the above mentioned research studies are summarized in table 1.3. A comparison is presented regarding to the predictive value of the various parameters from ultrasound and Bishop Score as reported in references. For an easy interpretation in figure 1.10 a scatter plot using the reported specificity and sensitivity values is shown. The triangular shaded area, represents the target predictive performance aimed at in this thesis work.

1.6 Chapter Summary

In this chapter a review of the labor induction procedure is presented along with complication that may arise and the current methods used to evaluate the cervix status prior to labor. Several findings have shown that tissues can be evaluated by parameters calculated from B-Mode ultrasound images. The echogenicity (related to gray level intensities) of tissues may vary in response to structural changes. These variations of gray level can convey important information about tissues and can be used as a classification tool for diagnosis. The texture concept is related to intensity variations and patterns in an image. Cervix maturation process is susceptible to be evaluated by texture analysis that could help to build a tool for non-invasive, quantitative evaluation of the cervix.

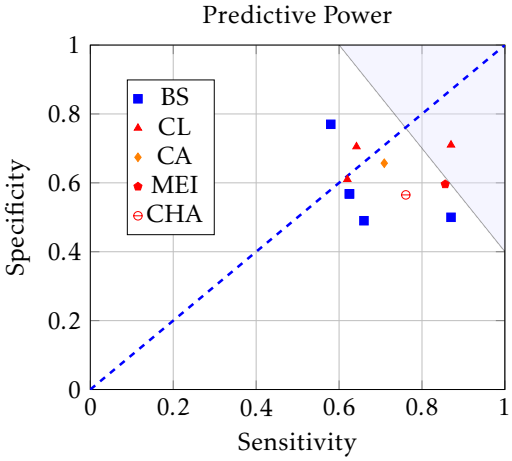


Figure 1.10: Scatter plot of sensitivity vs specificity for the various works discussed above for labor induction outcome prediction. Legend: BS: Bishop Score, CL: cervical length, CA: cervical area ,CHA: cervical hard area, and MEI: mean elastographic index.

References

[1] Nargess Afzali et al. "Cervical gland area: A new sonographic marker in predicting preterm delivery". In: *Archives of Gynecology and Obstetrics* 285.1 (2012), pp. 255–258.

[2] J A Y D I Ams et al. "the Length of the Cervix and the Risk of Spontaneous Premature Delivery". In: *The New England Journal of Medicine* 334.9 (1996), pp. 567–572.

[3] V. Berghella et al. *Cervical funneling: Sonographic criteria predictive of preterm delivery*. 1997.

[4] Jorge Burgos et al. "Induction at 41 weeks increases the risk of caesarean section in a hospital with a low rate of caesarean sections." In: *The Journal of Maternal-Fetal & Neonatal Medicine* 25.9 (2012), pp. 1716–8.

[5] S Chandra and JMG Crane. "Transvaginal ultrasound and digital examination in predicting successful labor induction". In: *Obstetrics & ...* (2001).

[6] Joan M G Crane. "Factors predicting labor induction success: a critical analysis." In: *Clin Obstet Gynecol* 49.3 (2006), pp. 573–584.

[7] Helen Feltovich, Kibo Nam, and Timothy J Hall. "Quantitative ultrasound assessment of cervical microstructure." In: *Ultrasonic imaging* 32.3 (2010), pp. 131–142.

[8] R. Gabriel et al. "Transvaginal sonography of the uterine cervix prior to labor induction". In: *Ultrasound in Obstetrics and Gynecology* 19.3 (2002), pp. 254–257.

[9] Raquel Garcia-Simon et al. "Cervix assessment for the management of labor induction: Reliability of cervical length and Bishop score determined

by residents". In: *Journal of Obstetrics and Gynaecology Research* 41.3 (2015), pp. 377–382.

[10] Ana Maria Gomez-Laencina et al. "Sonographic cervical length as a predictor of type of delivery after induced labor". In: *Archives of Gynecology and Obstetrics* 285.6 (2012), pp. 1523–1528.

[11] H S Hwang, I S Sohn, and H S Kwon. "Imaging analysis of cervical elastography for prediction of successful induction of labor at term". In: *J Ultrasound Med* 32.6 (2013), pp. 937–946.

[12] Tassilo Johannes Klein. "Statistical Image Processing of Medical Ultrasound Radio Frequency Data". PhD thesis. Technische Universitat Munchen, 2012.

[13] A Kurjak and FA Chervenak. *Donald School Textbook of Ultrasound in Obstetrics and Gynecology*. Ed. by Jaypee Brothers Medical Publishers. 2011.

[14] I. Z. MacKenzie. "Induction of labour at the start fo the new millennium". In: *Reproduction* 131.6 (2006), pp. 989–998.

[15] B. L. McFarlin et al. "Ultrasonic attenuation estimation of the pregnant cervix: A preliminary report". In: *Ultrasound in Obstetrics and Gynecology* 36.2 (2010), pp. 218–225.

[16] Barbara L McFarlin et al. "Quantitative ultrasound assessment of the rat cervix." In: *Journal of ultrasound in medicine : official journal of the American Institute of Ultrasound in Medicine* 25.8 (2006), pp. 1031–40.

[17] F. S. Molina et al. "Quantification of cervical elastography: A reproducibility study". In: *Ultrasound in Obstetrics and Gynecology* 39.6 (2012), pp. 685–689.

[18] G. K. Pandis et al. "Preinduction sonographic measurement of cervical length in the prediction of successful induction of labor". In: *Ultrasound in Obstetrics and Gynecology* 18.6 (2001), pp. 623–628.

[19] M. Parra-Saavedra et al. "Prediction of preterm birth using the cervical consistency index". In: *Ultrasound in Obstetrics and Gynecology* 38.1 (2011), pp. 44–51.

[20] G. Ramanathan et al. "Ultrasound examination at 37 weeks' gestation in the prediction of pregnancy outcome: The value of cervical assessment". In: *Ultrasound in Obstetrics and Gynecology* 22.6 (2003), pp. 598–603.

[21] Larry MD Rand et al. "Post Term Induction of Labor Revisited". In: *Journal of Obstetrics & Gynecology* 9.5 (2000), pp. 779–783.

[22] Lisa M Reusch et al. "Nonlinear optical microscopy and ultrasound imaging of human cervical structure." In: *Journal of biomedical optics* 18.3 (2013), p. 031110.

[23] H. Roman et al. "Does ultrasound examination when the cervix is unfavorable improve the prediction of failed labor induction". In: *Ultrasound in Obstetrics and Gynecology* 23.4 (2004), pp. 357–362.

[24] I Tekesin, L Hellmeyer, and G Heller. "Evaluation of quantitative ultrasound tissue characterization of the cervix and cervical length in the prediction of premature delivery for patients with spontaneous preterm labor". In: *American Journal of Obstetrics and Gynecology* 189.2 (2003), pp. 532–539.

[25] DA Wing. "Induction of labor". In: *Protocols for High-Risk Pregnancies* (2010).

Chapter 2

Theoretical framework for cervical evaluation.

2.1 Ultrasound Systems

Ultrasound is the term used to describe sound above 20,000 Hertz (Hz), beyond the human hearing frequency range. Ultrasonography or ultrasound is an imaging technique based on ultrasound. In this modality the image depends on the computer analysis of waves that in a noninvasive manner create images of internal body structures.

The World Health Organization (WHO) recognizes the ultrasound as an important form of medical imaging [14]. Its appeal comes from a variety of important reasons: It is non-ionizing, non-invasive, produce real time images and is available in most clinics and hospitals. The low cost of an US scanner (compared with other imaging modalities), makes it one of the preferred tools for monitoring, tracking and medical diagnosis.

In addition to the traditional fields of cardiology and obstetrics, in which has been widely used for a long time, it has also become very useful in the diagnosis of diseases of the prostate, liver, and atherosclerosis of the coronary and carotid arteries (deposits of yellowish plaques of cholesterol lipids and cellular debris in the inner layers of the arterial walls of medium and large diameter).

2.1.1 Image construction.

For imaging, an ultrasound scanner goes through a three-step procedure: first produces sound waves, then receives the echoes and finally processes that information to create a 2D gray scale image. The transmission and reception of ultrasound waves is usually accomplished by use of a piezoelectric transducer enclosed in a

Figure 2.1: Two types of ultrasound scanners and a 2D sonographic image showing a fetus.

housing of a variety of shapes, (see figure 2.2).

Transducers usually contain a large number of piezoelectric elements aligned next to each other along the transducer face to perform 2D scanning or arranged in a matrix for 3D scanning. These elements are generally made of thin wafers of artificial ceramic material such as lead zirconate titanate. The thickness (usually 0.1-1 mm) determines the frequency of ultrasound.

Typically up to five of these elements firing simultaneously generate a short pulse of ultrasound that travels in a narrow column away the probe. The transmitters then act as receivers and record the intensity of the reflected sound.

The process is repeated sequentially along the length of the probe. The time taken for an echo to return is used to determine the distance from the probe and is calculated assuming a constant sound speed of 1,540 m/s. This value is the average of the measurements obtained from normal tissues. The choice of the optimum ultrasonic frequency is determined by the resolution and the required penetration depth. The strength of returning echoes from any point is represented by the brightness of that point on the screen.

Figure 2.2: Different types of transducers used in diagnostic ultrasound.

2.1.2 Ultrasound -Tissue Interaction

As US waves travel through tissues, they are partly transmitted to deeper structures, partly reflected back to the transducer as echoes, partly scattered, and partly transformed to heat. The amount of echo returned after hitting a tissue interface is determined by a tissue property called acoustic impedance. This is a measure of the degree to which the medium opposes to the movement constituting the sound wave. Acoustic impedance z depends on the density ρ of the medium and the speed of sound c as expressed in the following equation $z = \rho c$. The sound speed in turn depends on medium properties compressibility κ and density ρ as $c = \sqrt{\frac{1}{\kappa * \rho}}$

Table 2.1: Acoustic impedance of different body tissues organs.

Body Tissue	Acoustic impedance (10^6 Rayls)
Air	0.0004
Lung	0.18
Fat	1.34
Liver	1.65
Blood	1.65
Kidney	1.63
Muscle	1.71
Bone	7.8

2.1.2.1 Attenuation

Sound energy is attenuated or weakened as it passes through tissue because parts of it are reflected, scattered, absorbed, refracted or diffracted. Attenuation and sound speed in tissues are important parameters related to the consistency of body parts. Both has been proposed in the literature to assess the state of the cervix as described in the following sections. In an early work [17] the authors studied the feasibility of using sound velocity to predict the cervical changes that could diagnose the structural development of cervical incompetence.

Sources of attenuation

Absortion Tissue absorption of sound energy contributes the most to the attenuation of an ultrasound wave in tissues.

Refractions Refraction is the change of direction experienced by the sound beam

being incident upon a tissue interface at an oblique angle. Angles of the incident wave and transmitted wave obeys Snell's law.

Reflections Organs containing gas (such as the lung or intestines) have the lowest acoustic impedance, while dense organs such as bone have very high acoustic impedance. The intensity of a reflected echo is proportional to the difference (mismatch) in acoustic impedance between two mediums. No echo is generated if two tissues have identical acoustic impedance.

Scattering Scattering occurs when echoes are scattered in all directions in a non-uniform manner. This is especially true when the sound wave hits an object whose size is much smaller than the sound wavelength. The part of the scattering that goes back to the transducer and generate images is called backscatter. The scattering phenomenon gives raise to the typical *speckle noise* present in sonographic images. The location, number and size of the objects being scanned in a particular region influence the statistical distribution of the speckle noise.

Resolution and Attenuation. The achievable resolution is greater with shorter wave lengths, being the wavelength inversely proportional to frequency. The propagation speed of sound c is related to the frequency f and wavelength λ by the equation $c = f\lambda$:

Table 2.2: Attenuation coefficients and propagation velocities of sound waves. From reference [6]

Tissue	Average attenuation coefficient in dB/cm	Propagation velocity of sound in m/s	Average acoustic impedance in $(g/cm^2/s) \times 10^5$
Fat	0.6	1450	1.38
Soft Tissue	0.7-1.7	1540	1.7
Liver	0.8	1549	1.65
Kidney	0.95	1561	1.62
Brain	0.85	1541	1.58
Blood	0.18	1570	1.61
Skull and bone	3-10 and higher	3500-4080	7.8
Air	10	331	0.0004

However, the use of high frequencies is limited by its greater attenuation in the tissue and therefore shorter penetration depth. The ultrasonic waves are attenuated in human tissues, typically by 0.2 -0,5 dB/cm/MHz. Thus the reflected echoes must be amplified by a factor which depends on the transversal depth, pro-

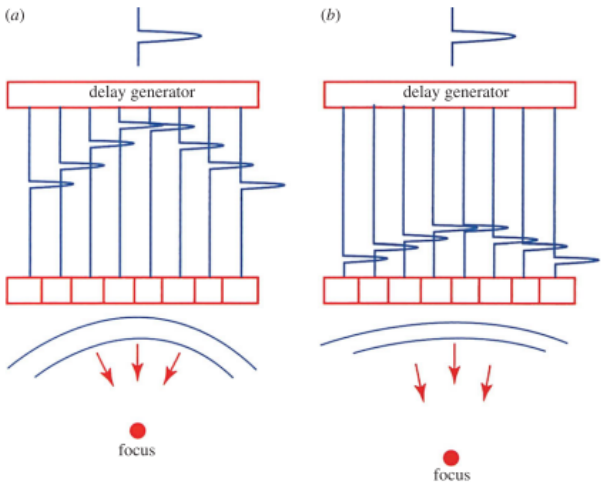


Figure 2.3: Array focusing showing delays for close focus (a) and far focus (b)

cess denoted as *time-gain compensation* (TGC). For this reason, different frequency ranges are used for the examination of various body parts:

- 3-5 MHz for abdominal area.
- 5-10 MHz for superficial and small parts and
- 10 to 30 MHz for the skin and eyes.

2.1.2.2 Array Beamforming

Beamforming is the process of providing the resulting wave from the elements being fired in a transducer array at certain time, with direction and focusing. Steering and focusing of the ultrasonic beam is nowadays achieved by careful electronic gating of the elements for both, transmission and reception of the resulting echoes. An example illustrating the kind of delays used for focusing is shown in figure 2.3.

2.1.2.3 Artifacts

Artifacts in ultrasound imaging are structures appearing in the image that have the following conditions:

- They are not really present.
- They are missing
- Represented in the screen with improper brightness

- Represented with improper shape or size

The origin of artifacts is diverse, some have origin in the scanner user (e.g not operating the equipment properly, wrong settings) and some are inherent of the ultrasound physics (e.g. shadowing). Commonly encountered artifacts include:

1. Reverberation
2. Shadowing
3. Mirror artifacts
4. Range Distortion
5. Side lobe Artifacts
6. Partial Volume

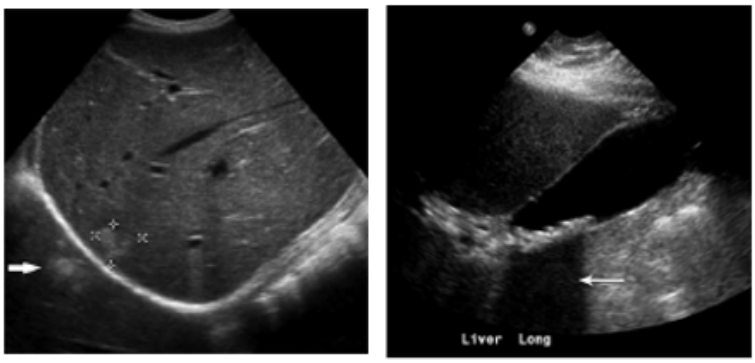


Figure 2.4: Ultrasound artifacts. To the left, mirroring artifact pointed by an arrow and right, shadowing.

2.2 Image Texture Analysis and its Applications in Medicine

Texture is an important tool for the analysis of many types of image features, including natural scenes and medical images. Although not yet defined, texture is an important spatial property related to patterns and changes on brightness in images. It has been used in many topics of image analysis such as segmentation tasks or classification.

2.3 Approaches to Texture Analysis

According to the methods used to evaluate the interrelationships between pixels (picture elements) or voxels (volume elements) methods for texture analysis have been classified in various ways. In [6] three main approaches are presented to represent the texture: statistical, structural and spectral. A more widely shown in [1] where different methods are classified as.

1. *Structural Methods* Texture is represented by the use of well defined primitives or texels (from **Tex**ture **El**ements), providing a good image symbolic description

2. *Model based methods* Mathematical models are used to represent the texture (stochastic (Markov Random Fields), fractal models).

3. *Spectral Methods* The properties of the image are analyzed in a different space as frequency or scale. These methods are based on some type of transform such as Fourier, Gabor and Wavelets, the latter being the most used.

4. *Statistical Methods* These are based on the representation of the texture using the properties that govern the interrelation and distribution of gray levels in the image. This distribution is analyzed by computing local features at each point in the image.

   Depending on the number of pixels defining the local feature, statistical methods can be further classified into first order (one pixel), second order (two pixels) and higher-order (three or more pixels) statistics.

2.3.1 Texture Attributes Derived from pixel statistics.

2.3.1.1 First Order Statistics

These statistics are calculated on one pixel which attribute is the gray level. i.e relationship with neighboring pixels are not considered. The gray level can be described by first order statistics such as mean, variance, dispersion, average energy, entropy , skewness and kurtosis estimated from a histogram computed from this distribution. For an image with n pixels, its histogram can be calculated as $P(i) = \frac{h(i)}{n}$, where $h(i)$ represents the number of occurrences of the i^{th} gray level and $P(i)$ is the probability of finding that particular gray level value. Some important first order statistic derived parameters are the following:

$$\text{Mean} \quad \mu = \frac{1}{N}\sum_{i=0}^{L} h(i) \tag{2.1}$$

$$\text{Variance} \quad \sigma^2 = \frac{1}{N-1}\sum_{i=0}^{L} (h(i)-\mu) \tag{2.2}$$

$$\text{Skewness} \quad s = \frac{1}{n\sigma^3}\sum_{i=0}^{L} (h(i)-\mu)^3 \tag{2.3}$$

$$\text{Kurtosis} \quad k = \frac{1}{n\sigma^4}\sum_{i=0}^{L} (h(i)-\mu)^4 - 3 \tag{2.4}$$

$$\text{Energy} \quad E = \sum_{i=0}^{L} (P(i))^2 \tag{2.5}$$

$$\text{Entropy} \quad H = -\sum_{i=0}^{L} P(i)\log(P(i)) \tag{2.6}$$

Where L is the highest of gray levels present in the image and N is the number of pixels in the image.

2.3.1.2 Second order statistics

Second order statistic consider relationship among pixels or groups of pixels (usuallly two). Among the most well known second order methods for texture analysis we can mention:

Gray level co-ocurrence matrix In the work of Haralick (1973) [9], he proposed the use of a gray level co-ocurrence matrix (GLCM) which has since then become a very popular method for texture analysis. In the GLCM each entry $GLCM_{(\delta_x.\delta_y)}(i,j)$ represent a probability estimate of the co-ocurrence of the gray levels i and j at two arbitrary locations separated by the displacement $(\delta_x.\delta_y)$. Different matrices are obtained by modifying the spatial relationship, orientation (angle) or distance between pixels. The number of rows and columns of the GLCM thus, depends only on the gray levels in the texture and not on the image size.

In figure 2.5 an 3x3 image and its co-ocurrence matrix is shown. The angle θ in this example is 0^o , or the pixel to the right of the considered pixel ($d = 1$). The 2 in the GLCM indicates that there are two occurrences of a pixel with gray level 3 immediately to the right of pixel with gray level 1.

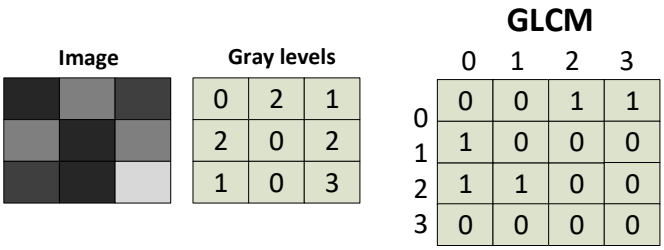


Figure 2.5: Original image and Co-ocurrence matrix. Here we assume only 4 gray levels, so the matrix is four by four,

Several parameters are computed from GLCM entries. Haralick proposed a total of 14 statistical measures: angular second moment, contrast, correlation, entropy, energy are among the most used. The defining equations for these texture features are:

$$\text{Energy} = \sum_i \sum_j N_d(i,j)^2 \tag{2.7}$$

$$\text{Entropy} = -\sum_i \sum_j N_d(i,j) log_2 N_d(i,j) \tag{2.8}$$

$$\text{Contrast} = \sum_i \sum_j (i-j)^2 N_d(i,j) \tag{2.9}$$

$$\text{Homogeneity} = \sum_i \sum_j \frac{N_d(i,j)}{1+|i-j|} \tag{2.10}$$

$$\text{Correlation} = \frac{\sum_i \sum_j (i-\mu_i)(j-\mu_j) N_d(i,j)}{\sigma_i \sigma_j} \tag{2.11}$$

Where μ_i and μ_j are the means and σ_i and σ_j are the standard deviations of the row and column sums. These summations are denoted $N_d(i)$ and $N_d(j)$ and are defined by,

$$N_d(i) = \sum_j N_{(i,j)} \tag{2.12}$$

$$N_d(j) = \sum_i N_{(i,j)} \tag{2.13}$$

Local Binary Patterns In general, LBP measures the local structure at a given pixel using P samples on a circle of radius R around the pixel and summarizes this

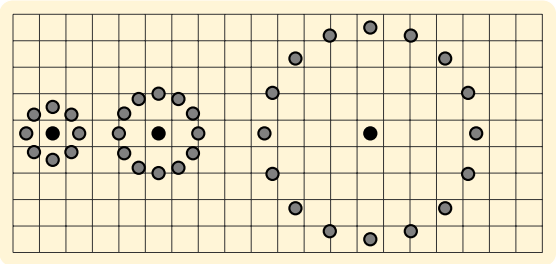


Figure 2.6: Different resolution LBPs using a circular neighborhood. Left $P = 8, R = 1$, Center: $P = 12, R = 1.5$ and right $P = 16, R = 4$

information with a unique code for each local structure or pattern, some examples are shown in figure 2.6. To calculate the code, g_P the intensity of neighboring samples is compared with the intensity of the center pixel (g_C) and a sign function is applied using this value as a threshold. A sample are thus assigned one if its gray value is larger than the threshold or zero when the opposite is true. By choosing a fixed sample position on the circle as the leading bit, the samples can be turned into a binary number. Thus each pattern has an associated unique binary number calculated using equation 2.14.

$$LBP_{P,R} = \sum_{p=0}^{P-1} t(g_p - g_c)2^{-p} \tag{2.14}$$

where $t(.)$ represents a thresholding operation.

Sample position not on pixel location requires interpolation. The LBP histogram thus combines structural and statistical information of an image. A complementary contrast measure $VAR_{P,R}$ was also developed in [7] that together with LBP can describe an image by their pattern and contrast aspects. The VAR operator is calculated according to equation 2.15. Contrast measure has a continuous-valued output; hence, quantization of its feature space is needed.

$$VAR_{P,R} = \frac{1}{P}\sum_{p0}^{P-1}(g_p - \mu)^2 \text{ where } \quad \mu = \frac{1}{P}\sum_{p=0}^{P-1} g_p \tag{2.15}$$

Local Binary Pattern Mappings Since the original LBP was not rotation invariant, some especial coding strategies have been introduced. The *rotation invariant LBP* mapping consist on performing bit shifting so that image rotation do not alter the calculated binary code. The rotation invariant LBP ($LBP^{ri}_{P,R}$) is defined as,

$$LBP^{ri}_{P,R} = \min\{ROR(LBP_{P,R}i)| \quad i = 0,1,..,P-1\} \tag{2.16}$$

where $ROR(x,i)$ is a circular bit-wise shift on the P-bit number. Uniform patterns were also developed to reduce the number of possible bins. An LBP code is called *uniform* if the binary pattern consists of at most two bitwise transitions from 0 to 1 or vice versa. This coding scheme reduces the number of bins of a LBP histogram. In a neighborhood of P bits there are by definition $P+1$ uniform codes , the remaining non-uniform codes are included in an additional bin so that the complete histogram consist of $P+2$ bins. The rotation invariance with uniform patterns $LBP^{riu2}_{P,R}$ is defined as follows:

$$LBP^{ri}_{P,R} = \begin{cases} \sum_{p=0}^{P-1} \text{sign}(g_p - g_c) & \text{if} \quad \text{U}(LBP_{P,R}) \leq 2 \\ p+1, & \text{Otherwise} \end{cases} \tag{2.17}$$

where,

$$\begin{aligned} \text{U}(LBP_{P,R}) &= |s(g_{P-1} - g_c) - s(g_0 - g_c)| \\ &\quad + \sum_{p=1}^{P-1} |s(g_P - g_c - s(g_{p-1} - g_c)| \end{aligned} \tag{2.18}$$

The U value of an LBP pattern is defined as the number of spatial transitions (bitwise 0/1 changes) in that pattern. It was verified that only "uniform" patterns are fundamental patterns of local image texture.

Gray Level Difference Method The *Gray Level Difference Method* (GLDM) is based on the occurrence of two pixels which have a given absolute difference in gray level and which are separated by a specific displacement $\delta = (\Delta x, \Delta y)$.

If I is a intensity image, for a given displacement δ, let $I_\delta(x,y) = |I(x,y) - I(x + \Delta x, y + \Delta y)|$ and $f(i|\delta)$ be the estimated probability-density of $I_\delta(x,y)$. The value of $f(i|\delta)$ is obtained from the number of times $I_\delta(x,y)$ occurs for a given δ , i.e

$$f(i|\delta) = Prob[I_\delta(x,y) = i] \tag{2.19}$$

Other well known methods are the *Autocorrelation function* and the *Gray Level Run Length (GLRL)* method.

2.4 Image Texture for Classification

Medical applications often require the automatic extraction of features for image classification tasks, such as to distinguish between normal and abnormal tissues schemes for detection of tumors. When analyzing images, radiologists usually not only observe brightness variations but the patterns present in the images. Features derived from texture analysis can provide useful information not only for locating an organ (i.e segmentation) in an image but to evaluate tissue state for classification.

Texture analysis has been used in the diagnosis tissues like white matter damage in [16] to analyze the neonatal brain by Transcranial ultrasound (TUS). To distinguish between injuries plaques of different composition (calcified,fibrous and necrotic) in intravascular ultrasound images (IVUS), several texture analysis techniques were tested in [19]. Attributes such as first order statistics, Haralick method, Law's Energy method, Gray Level Difference Matrix method and Texture Spectrum were tested on 27 coronary plaques using discriminant analysis.

The average gray level, the standard deviation and the width of the histogram of fetal lung sonograms were used in [15] to predict fetal maturity. The liver was taken as a reference due to its stability. They concluded that the fetal lung is mature when pulmonary echogenicity (gray level intensity) is greater than that of the liver.

For the analysis of breast US images, the Self-Invariant Feature Transform (SIFT) has been proposed in the work of [13], authors argued that SIFT descriptors provide invariability to scale, rotation and minor affine transformations along with robustness to illumination changes which are currently issues in US imaging. In the same vein, fractal analysis [3] and [3], Wavelets transform derived parameters; variance contrast, autocorrelation contrast and distribution distortion of wavelet coefficients help to differentiate the benign and malignant breast tumors in sonograms. In figure 2.7 we can observe the textural difference in normal and diseased tissues.

Multi scale approaches for lung tissue analysis bases on texture have been published, using the Riesz wavelet transform [5], Wavelets frames [4] for high resolution digital computed tomography (HRDCT) images.

A statistical generalized texture analysis technique to characterize and recognize the most important diagnostically typical vascular patterns relating to cervical lesions in colposcopy images is developed in [10].

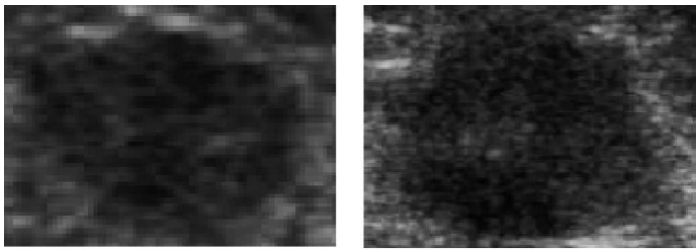


Figure 2.7: Breast Tumor images, left benign and right malignant.From reference [2]

Table 2.3: Summary of texture based US image processing approaches.

Application	Modality	Algorithm / Features
Atherosclerotic plaque	IVUS	Haralick's method
Cervical evaluation (cancer lesions)	Colposcopy images	Vascular patterns.
Fetal gray matter diagnosis	TUS	Texture modelled as a MRF
Cervical ripening evaluation	TA,TVU	Gray level first order statistics, co-ocurrence matrix
Breast tumor classification	BUS	Wavelets coefficients derived: variance contrast, autocorrelation contrast and distribution distortion
	BUS	Self-Invariant Feature Transform (SIFT)
	BUS	Fractal Analysis
Lung tissue classification	HRDCT	Riesz wavelet transform
Fetal lung maturity	TA (Trans. abd).	Mean gray level, Histogram width

For the evaluation of cervical ripening, there are some works that explore textural features of ultrasound images [8, 11, 12, 20]. These works are described in greater detail in the following section.

To improve performance of the algorithms, hybrid approaches have also been proposed in the literature calculate, e.g. calculating co-ocurrence matrices or histogram of wavelets coefficients at different scales.

Texture Analysis on Ultrasound Images The dominating methods in US image processing are those related to statistical approaches: Co-ocurrence matrix parameters (GLCM), width of gray level histograms (GLH), run-length matrix (RLE) parameters are among the most used. For cervical evaluation only statistical methods have been used so far.

Investigating about the application of model based, multi-resolution or multi-scale approaches on similar contexts of US imaging can shed some light concerning the choice of a particular method of analysis.

We are referring to US modalities using similar frequencies and transducer type. A logical choice to include is prostate evaluation by trans-rectal ultrasound (TRUS). Prostate examination by TRUS share similar probes, frequencies and an alternative digital examination. Applications on segmentation and classification

will be compared to gain insight about possible alternatives to current texture based methods for cervical assessment.

2.5 Cervical Assessment by Texture Analysis

Literature review has shown that among the methods using image processing for cervical evaluation only 4 studies are supported on texture analysis. These works however are concerning preterm birth, no previous work related to labor induction was found so far.

1. In his work, Wischnik [20] proposes the use of a parameter called texture score (TS) derived by multiple regression of various attributes such as texture co-occurrence matrix, gray level first order statistics and gradients. In this study participated 112 patients with normal pregnancies (14-41 weeks) and 57 patients admitted because of cervical insufficiency (20-35 weeks), representative regions of the image were analyzed using statistical software to find the best discriminatory parameters (entropy, contrast and co-occurrence matrix correlation). The study claims that his method can replace digital evaluation of the cervix.

2. Jörn [11] conducted a study to assess: 1) Changes in the cervix from non-pregnant women, and early and late pregnancy and (2) the differences between the cervices of pregnancies complicated by preterm labor and normal term pregnancy. 4 patients nonpregnant and pregnant women with the same number of first and second trimester of pregnancy were examined, and 5 patients with complications of premature birth by transvaginal ultrasound.

   In the study of regions of interest (ROI) within the images of TVU mean value of gray levels histogram was calculated, contrast and homogeneity of the co-occurrence matrix. ROIs were placed in the inner and external os and the anterior and posterior lips. The brightness of the texture decreases in the area of the external os from the state of non-pregnancy to early and late pregnancy. In the area of the inner os the contrast increases and homogeneity decreases; in the external os area texture changes are the opposite. The textures in the area of the internal os of pregnancies complicated by preterm labor were dark, showed more contrast and less homogeneity compared to term pregnancies.

3. In order to evaluate the echogenicity of the area surrounding the cervical canal (or Glandular Area CGA), Furtado [8] used the gray level histogram of transvaginal ultrasound (TVU) images in pregnancies between 20 and 25 weeks of gestation. The purpose was to objectively determine the absence of glandular area that has high variability between observers. In this study 149 patients in the second trimester were involved, the middle portion of the cervix was selected as the ROI because it is less influenced by the cervical length or position during the ultrasound examination. The results indicated that between 20 and 25 weeks the region surrounding the cervical canal is predominantly hypo-echoic (low gray level intensity). The mean, minimum and maximum value of the histogram were obtained for the glandular area and for the surrounding area and cervical tissue (stroma). The CGA/stroma ratio according to the authors could be considered a better indicator than the CGA only because this relationship is normally distributed with a mean value with less dispersion and a narrower standard deviation.

4. The mean gray value (MGL) was measured in the midsection of anterior and posterior cervical walls in the study realized by [12]. The difference in MGL between anterior and posterior (AP difference) was related to the Bishop subscore for cervical consistency (0, 1, or 2). They found that a more echogenic anterior than posterior cervix indicates a hard cervix; the greater the difference in echogenicity between anterior and posterior walls the harder the cervix

Regarding the image processing of B-mode images of the cervix, studies have been limited mainly to statistical approaches. We think that by implementing multi-scale , multi-resolution schemes (using some kind of transform) of appropriate texture descriptors we can improve the classification performance of texture-based methods.

As mentioned in [18]: *"Texture description is highly scale dependent. To decrease scale sensitivity, a texture may be described in multiple resolutions and an appropriate scale may be chosen to achieve the maximum discrimination "*.

Transform based techniques are appealing for texture analysis for several reasons:

- They possess zooming capabilities to arbitrary scales in the analysis, thus allowing examination of textures at their appropriate scales. To successfully characterize textures, it is equally important to describe both their local

Table 2.4: Summary of texture based image processing methods applied to TVU images in cervical evaluation.

Year	Data	Features	Results
1999 [20]	112 normal pregnancy patients (14-41 weeks) scanned with TVU,TUS	125 parameters derived from several image features: gray level first order statistics, gradients and co-ocurrence matrix	Able to reproduce Digital examination
2008 [11]	4 non-pregnant patients of each first and third trimester and 5 with complication of preterm birth .	Gray level histogram mean value, homogeneity and contrast of co-ocurrence matrix	Dark an low contrast textures are developed during pregnancy and are also characteristic of preterm birth.
2010 [8]	149 patients on their second trimester of pregnancy were scanned with TVU	Mean value, standard deviation , minimum and maximum values of the Gray Level Histogram of the glandular area(CGA) and stroma.	The ratio CGA/stroma allows the objective of the presence or absence of CGA
2010 [12]	214 women with low-risk singleton pregnancy during 27-30(th) pregnancy week scanned by TVU	Mean Gray Value Histogram	A more echogenic anterior than posterior cervix indicates a hard cervix

(loosely referred to as micro-textures) and global (macro-textures) properties. Multi-scale methods mimic the human vision system (HVS) that analyze images at several levels of detail.

- They are computationally efficient, they can be calculated in a fraction of the time a model-based counterpart uses.
- They may provide rotation and translation invariant texture attributes.

2.6 Chapter Summary

Summarizing we can say that methods developed so far for cervical evaluation by texture analysis are based mainly on statistical local texture descriptors. Their use have been mostly as a Bishop sub- score category replacement (consistency), i.e, they replace digital examination of consistency by assessing firmness through image processing. They are also not use alone but combined with other ultrasound parameters such as cervical length. Implementing multi-resolution schemes for texture analysis of US cervical images or by combining statistical and multiresolution approaches may result in an increase in classification accuracy by analyzing texture at different scales not only locally.

References

[1] G Castellano et al. "Texture analysis of medical images." In: *Clinical radiology* 59.12 (2004), pp. 1061–9.

[2] Dar-Ren Chen et al. "Classification of breast ultrasound images using fractal feature." In: *Clinical imaging* 29.4 (2005), pp. 235–45.

[3] Dar-Ren Chen et al. "Diagnosis of breast tumors with sonographic texture analysis using wavelet transform and neural networks". In: *Ultrasound in Medicine & Biology* 28.10 (2002), pp. 1301–1310.

[4] Adrien Depeursinge et al. "Lung tissue classification using wavelet frames". In: *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE* 2007.FEBRUARY 2007 (2007), pp. 6259–6262.

[5] Adrien Depeursinge et al. "Multiscale lung texture signature learning using the Riesz transform." In: *Miccai* (2012), pp. 517–524.

[6] P Dhawan. *Medical image analysis*. John Wiley & Sons, 2011, pp. 353–368.

[7] Ahmed Elgammal, David Harwood, and Larry Davis. "Computer Vision ECCV 2000". In: *Computer Vision ECCV 2000* 1843.October (2000), pp. 751–767.

[8] Marcio R. Furtado et al. "Transvaginal grey scale histogram of the cervix at 20-25 weeks of pregnancy". In: *Australian and New Zealand Journal of Obstetrics and Gynaecology* 50.5 (2010), pp. 444–449.

[9] R M Haralick and I Dinstein. *Textural Features for Image Classification*. 1973.

[10] Qiang Ji, John Engel, and Eric Craine. "Texture analysis for classification of cervix lesions". In: *IEEE Transactions on Medical Imaging* 19.11 (2000), pp. 1144–1149.

[11] H Jörn. "Prediction of premature birth using texture analysis of the cervix". In: *Ultraschall in der Medizin-European Journal of Ultrasound* 29.S2 (2005), P0_16.

[12] Tomoyuki Kuwata et al. "A novel method for evaluating uterine cervical consistency using vaginal ultrasound gray-level histogram". In: *Journal of Perinatal Medicine* 38.5 (2010), pp. 491–494.

[13] Joan Massich et al. "SIFT texture description for understanding breast ultrasound images". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8539 LNCS (2014), pp. 681–688.

[14] World Health Organization. *Manual of Diagnostic Ultrasound Vol 2*. Vol. 2. World Health Organization, 2013.

[15] A Salamanca, A Carrillo, and M Palomino. "Prediction of fetal maturity through sonographic lung characterization". In: *Progr Obstet Ginecol* 45.3 (2014), pp. 96–103.

[16] M. Sanz-Cortés et al. “Fetal brain MRI texture analysis identifies different microstructural patterns in adequate and small for gestational age fetuses at term”. In: *Fetal Diagnosis and Therapy* 33.2 (2013), pp. 122–129.

[17] Amr A R Sharawi et al. “Ultrasound Velocity in cervix Uteri in Correlation with Structural Changes for Diagnosis Incompetence”. In: *Proceedings of the Annual Conference on Engineering in Medicine and Biology* 12.1 (1990), pp. 350–351.

[18] Milan Sonka, Vaclav Hlavac, and Roger Boyle. *Image processing, analysis, and machine vision*. Cengage Learning, 2008.

[19] D G Vince et al. “Comparison of texture analysis methods for the characterization of coronary plaques in intravascular ultrasound images”. In: *Computerized medical imaging and graphics* 24.4 (2000), pp. 221–229.

[20] A Wischnik, R Stöcklein, and T Werner. *Evaluating the pregnant cervix uteri by ultrasound with computer-assisted texture analysis*. 1999.

Chapter 3

Texture Analysis of B-Mode cervical images

Extracting useful information from Ultrasound (US) B-mode images is a challenging task. US images are low contrast, contain blurred edges and they are normally contaminated with speckle noise. Despite these drawbacks a lot of effort has been made on processing US images.

In this chapter we present some image processing methods applied to the transvaginal ultrasound images collected during the duration of this research work. We also present results obtained from classification experiments, where the objective is to study the feasibility of constructing texture-based reliable algorithms for predicting labor induction outcome. One important aspect of texture is scale. It is known that the human visual system processes images in a multi-scale way. There are many neurophysiological and psychophysical data indicating the multi-scale analysis by the human visual front-end system [20]. The visual cortex has separate cells that respond to different frequencies and orientations. Analyzing texture at several resolutions is required when dealing with non-stationary textures as those obtained in medical imaging. We resort to several multi-resolution texture analysis schemes to analyse the images in our database.

3.1 Image Database

The database used in this thesis consist of images from patients admitted for labor induction procedures at the Obstetrics and Gynecology Service, Biocruces Health research Institute (Bilbao,Spain) during a period of one year. The inclusion criteria were singleton pregnancies, and $\geqq$ 37 weeks of gestation. Pregnancies of fetus suffering from infections and abnormalities were not included.

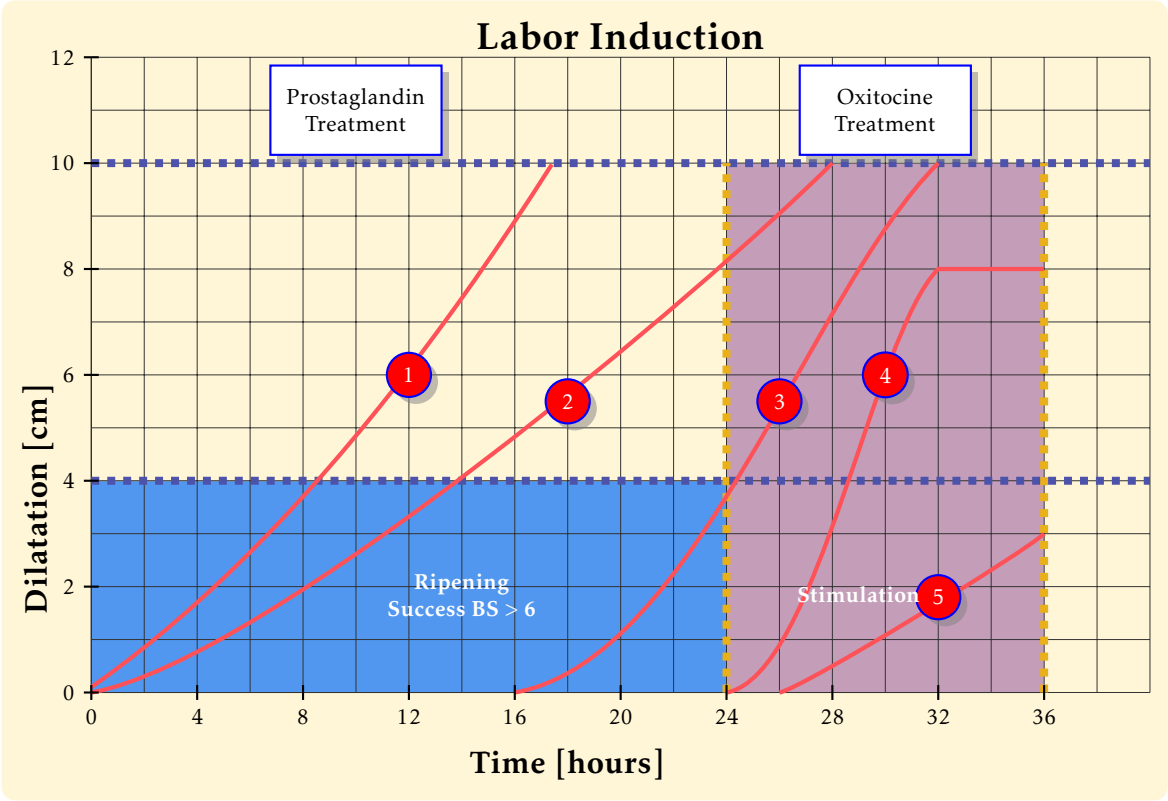


Figure 3.1: Labor induction stages. First 24 hours ripening stage and additional 12 hours stimulation stage. Curve legend: 1. Vaginal delivery and ripening success, 2. Vaginal delivery and ripening failure, 3. Vaginal delivery, stimulation success, 4. Cesarean section, stimulation failure, 5. Failure in both ripening and induction.

Annotations about weeks of pregnancy, labor induction indication and outcome were also attached to the collected images. Settings for the ultrasound scanner were defined in a protocol and practiced for all obstetricians participating in this study. Images were acquired during routine patient transvaginal scanning prior to labor induction. All images were acquired as DICOM files but pixel information was extracted as bitmaps for further processing. Both, images and annotation data were stored in a MYSQL database.

The labor induction process has been divided into two stages: A 24 hours *ripening stage* where prostaglandins are administered, followed by an additional 12 hours *stimulation stage* where the treatment is changed to oxytocin in case cervical ripening is not achieved. The whole procedure is illustrated in figure 3.1.

A successful ripening is thus defined when a vaginal delivery was obtained within 24 hours after induction is started. Labor induction failure is considered when after 36 hours the cervix still has a Bishop Score ≤ 6.

Table 3.1: Summary of cesarean section cases.

Cesarean section cause	Relation to Cerv. Ripening			
	Null	Low	medium	High
Breech presentation	✓			
Failure to descend	✓			
Failure to progress in labor	✓			
Fetal distress		✓		
Secondary arrest of dilatation			✓	
Induction failure				✓

Image Acquisition All images were acquired with a Voluson E8 Ultrasound scanner from General Electric. A total of 82 DICOM files were acquired, and from these 60 belong to patients with a vaginal delivery and 22 to cesarean section. These latter were further classified according to their relation to cervical ripening. Four categories were established for this purpose, namely: null, low, medium and high as summarized in table 3.1.

In the experiments we only took into account cesarean sections corresponding to categories medium and high. Also we excluded from analysis low quality images. The cases fulfilling those criteria were 54, divided as follows: 44 vaginal deliveries and 9 cesarean section.

Regions of Interest (ROI) All images in the database included up to eight ROIs. These ROIs were manually delineated by an expert obstetrician and defined several regions in the cervix lips as described in figure 3.2. The motivation for including all these ROIs was to study if there is an optimal lip region for texture analysis as it has been found in similar texture analysis.

Image resizing. Images in the database are rather big (975 x375 pixels) and some areas do not contain relevant information such as annotations and dark areas re-

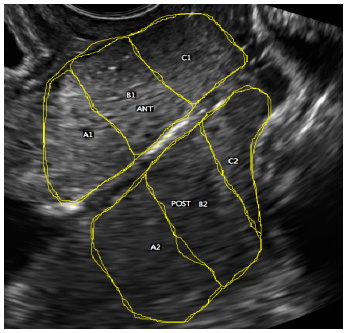


Figure 3.2: A sample TVU image showing the eight ROIS.

ROI	Region
ANT	Anterior lip
POST	Posterior lip
A1	Left anterior lip region
B1	Center anterior lip region
C1	Right anterior lip region
A2	Left posterior lip region
B2	Center posterior lip region
C2	Right posterior lip region

Table 3.2: ROI names and descriptions.

sulting from scan conversion and not corresponding to sector scanning. For this reason a squared area of 600 x 600 containing our regions of interest were cropped from every image and the ROI coordinates were also transformed for their use in the smaller images.

3.2 Texture operators.

To analyze the micro texture in the US images, we chose texture operators susceptible to be applied to region of interest of arbitrary shape, not necessarily squared, this include for example, histogram-based features.

Texture operators to be used in the experiments are the Local Binary Patterns (LBP), Gray Level co-occurrence matrix (GLCM), Gray Level Difference Matrix (GLDM) and First Order statistics (FOS).

3.3 Multiresolution methods.

For the multiscale and multiresolution analysis of the images in our database, some of the most popular transform for image texture processing were used. The purpose was to find out if the orientation, scale or frequency were important aspects during the classification of the images. The selected transforms were:

1. Wavelets.
2. Pyramidal directional filter banks, a.k.a Contourlets.
3. Conventional Gabor filter.
4. Circular Gabor filters.

3.3.1 Wavelets.

The wavelet transform decompose a signal by means of a series of elementary functions, created from dilations and translations of a basis function ψ known as mother wavelet. The basis functions of a discrete wavelet transform, $\psi_{j,k}(t)$, of time independent variable t, can be expressed as

$$\psi_{j,k}(t) = 2^{-\frac{j}{2}}\psi(2^{-j}t - k) \tag{3.1}$$

where j and k are integers that guide the dilations and translations of the func-

tion ψ to generate a family of wavelets, such as Haar and Daubechies. Wavelet transforms provide simultaneous time and frequency localization and thus are useful for analyzing time-variant, non-stationary signals.

Wavelets have been used extensively since its development. In image processing wavelets have become popular tools for denoising, compression and enhancement.

3.3.2 Pyramidal Directional Filter Banks

One way to assess the way orientation influences image classification is to study the image at several orientations, for this purpose the pyramidal filter banks is a good tool to be used.

Pyramidal Directional Filter Bank (Contourlets) is a 2D directional multiscale image decomposition developed to efficiently approximate images made of smooth regions separated by smooth boundaries. The Contourlet transform has a fast implementation based on a Laplacian Pyramid decomposition followed by directional filterbanks applied on each bandpass subband (see figure 3.3)

The Contourlet transform has properties of multiresolution, localization, directionality, almost critical sampling and anisotropy (important for finding dominant or preferred orientations) . Its basic functions are multiscale and multidimensional. The contours of original images, which are the dominant features in natural images, can be captured effectively with a few coefficients by using Contourlet transform.

Contourlet transform since its conception has found numerous applications in image processing: image retrieval [1], texture analysis of US images for thyroid nodules detection [12], brain image segmentation of MRI images [11], small bowel tumors detection in capsule endoscopy [2] to mention some. Is has been proposed also for tissue classification on cervical ultrasound images [17].

3.3.3 The Gabor Filter.

Gabor filters are constructed by combining oriented complex sinusoidal modulated by Gaussians functions. Gabor filtering has emerged as one of the leading approaches. The capability of texture discrimination of Gabor functions seems to be related both to their optimal joint resolution in space and frequency, and to their aptitude of modeling the response of cortical cells (simple cells) devoted to the processing of visual signals. The link between Gabor functions and the visual

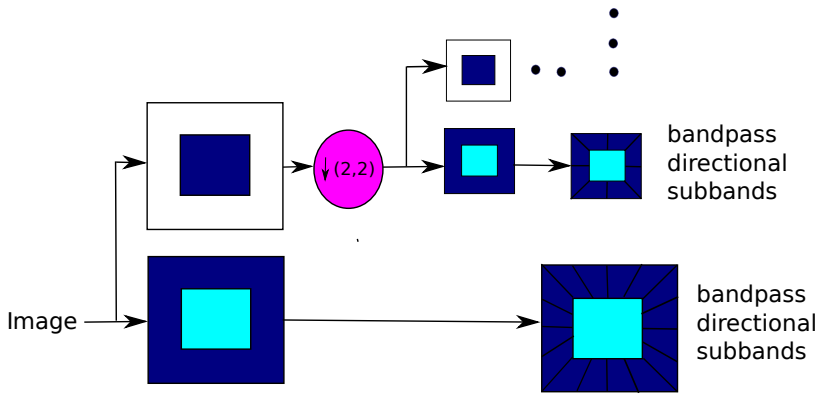


Figure 3.3: The contourlet filter bank: first, a multiscale decomposition into octave bands by the Laplacian pyramid is computed, and then a directional filter bank is applied to each bandpass channel.

system of mammals has been investigated and discussed by various authors.

Although Gabor filters are widely adopted, they suffer from certain limitations, mainly because they depend on various parameters that need to be set properly. This problem, sometimes referred to as *filter bank design*, involves the selection of a suitable number of filters at different orientations and frequencies.

These filters have been used extensively in image processing applications such as texture segmentation [4, 6, 23], image retrieval [24] and texture classification [8, 10]. It performs a localized and oriented frequency analysis of a two-dimensional signal. The formulation in the spatial domain is as follows:

$$g(x,y) = \frac{1}{2\pi\sigma_x\sigma_y}\exp\left[-\frac{1}{2}\left(\frac{\tilde{x}^2}{\sigma_x^2}+\frac{\tilde{y}^2}{\sigma_y^2}\right)\right]\exp\left(2\pi j W\tilde{x}\right) \tag{3.2}$$

$$\begin{cases}\tilde{x} &= x\cos\theta + y\sin\theta\\ \tilde{y} &= -x\sin\theta + y\cos\theta\end{cases} \tag{3.3}$$

Where σ_x and σ_y characterize the spatial extent and bandwidth of the filter, and w_X is the modulation frequency. θ ($\theta \in [0,\pi)$) specifies the orientation of the filter. W is is the radial frequency of the sinusoid. The Fourier transform of the Gabor function en equation 3.2 is given by:

$$G(u,v) = \exp\left[-\frac{\pi^2}{F^2}\left(\gamma^2(\tilde{u}-W)^2+\eta^2\tilde{v}^2\right)\right] \tag{3.4}$$

$$\begin{cases}\tilde{u} &= u\cos\theta + v\sin\theta\\ \tilde{v} &= -u\sin\theta + v\cos\theta\end{cases} \tag{3.5}$$

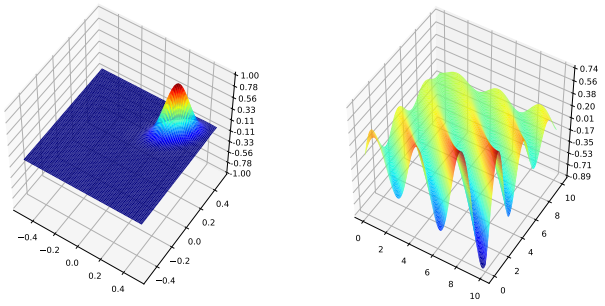


Figure 3.4: Gabor filter responses in frequency domain (left) and spatial domain (right). Parameters of the filter are as follows: $\theta = \frac{\pi}{4}$, $\sigma_x = 20$,$\sigma_y = 40$ and $F = 0.4$.

where $\gamma = 2\pi\sigma_x$, $\eta = 2\pi\sigma_y$. The Fourier representation in equation 3.4 specifies the amount by which the filter modifies each frequency component of the input image.

3.3.4 The Circular Gabor Filter

A modified version of Gabor filters termed circular Gabor filter [25] is used here in order to study TVU images at different frequency scales. In these filters the sinusoid varies along all orientations, leading to a circular symmetric response. These filters have been found particularly useful in rotation invariant texture analysis [5], [14]. The circular Gabor filter is defined as follows:

$$g(x,y) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2)}{2\sigma^2}} e^{2\pi i F\left(\sqrt{x^2+y^2}\right)} \tag{3.6}$$

$$F(u,v) = \frac{\sqrt{2\pi}}{2}\alpha e^{-\frac{\left(\sqrt{u^2+v^2}-F\right)^2}{2\alpha^2}} \tag{3.7}$$

In equation 3.7, we define, $\alpha = \frac{1}{2\pi\sigma}$. Equations 3.6 and 3.7describe the filter in the spatial and frequency domain respectively.

For the parameter selection, we make use of the following relationships (see figure 3.5)

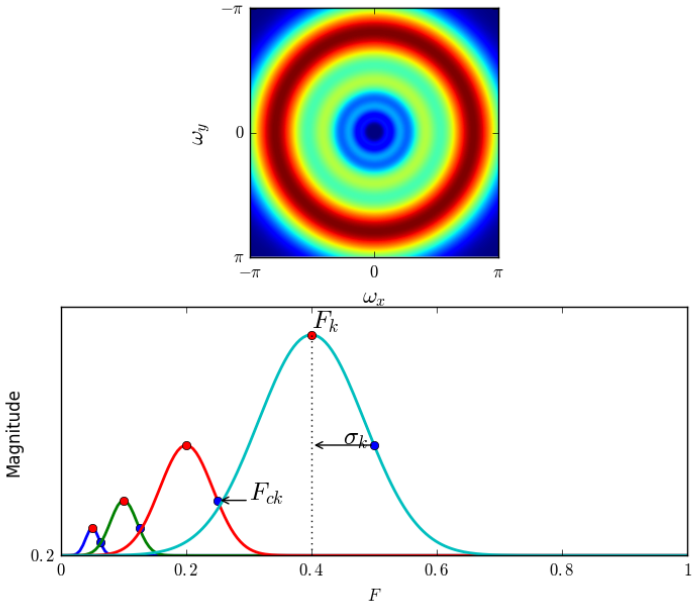


Figure 3.5: Frequency response of a circular Gabor filter bank using 4 scales and a section showing parameters used in the design process

$$F_k = f_o 2 * Bk \tag{3.8}$$

$$F_{ck} = \frac{1}{2} F_k \left(2^B + 1\right) \tag{3.9}$$

$$\sigma_k = \frac{\lambda}{\left(F_k - F_{c(k-1)}\right)} \tag{3.10}$$

where $\lambda = \frac{\sqrt{(2ln2)}}{2\pi}$, B is the bandwidth in octaves, F_k is the central frequency of the filter, $F_{ck}$is the frequency of half bandwidth and $f_0$is the lower limit of the frequency range under consideration. The frequency is usually normalized by the image size N giving a maximum frequency of 0.5.

3.4 Classifiers.

When choosing a classifier for a particular task one has to have in mind the type of data at hand. Aspects like the available amount of data for training, the dimensionality (number of features) of the sample data, or the nature of the labeled outputs (i.e numerical or categorical) are to be considered .

Usually a high dimensional feature space is generated when working with multiresolution techniques for texture analysis. This is due to the concatenation of the feature subsets obtained from different scales to be submitted to a classifier. The classifier suffers from the curse of dimensionality due to the high dimensional feature space, i.e. many data samples are required to train the classifier with a reasonable performance.

In medical applications, one usually face the problem of having only a limited amount do data. Classifiers performing well with high dimensional feature space and modest amount of data are desirable. So this is equivalent to say that a low variance estimator is needed in this case. The most frequently used classifiers in medicine are:

1. Support Vector Machine (SVM).
2. Multilayer Perceptron (MLP).
3. K-nearest neighbors (KNN).

K-Nearest neighbors. In pattern recognition, the k-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression. In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). KNN has some nice properties: it can be used for linear and nonlinear distributed data, it tends to perform well with a lot of data samples. Increasing the parameter k will decrease variance and increase bias. While decreasing k will increase variance and decrease bias.

Support Vector Machines. SVM can be used in linear or non-linear ways with the use of a Kernel, when you have a limited set of points in many dimensions SVM tends to be very good because it should be able to find the linear separation that should exist. SVM is good with outliers as it will only use the most relevant points to find a linear separation (support vectors). SVM needs to be tuned, the cost C and the use of a kernel and its parameters are critical hyper-parameters to the algorithm. Some common kernels are the linear kernel and the Gaussian kernel.

The multilayer Perceptron. The multilayer perceptron (MLP) is a type of feedforward network model that maps a set of inputs onto a set of outputs. Feedforward is a term used to describe networks where calculations are performed from input to output direction. A MLP is constituted by several layers of nodes in a directed graph, with each layer fully connected to the neighboring nodes in the next layer. Each connection can have an adjustable value or weight.

All nodes (except from the input layer) are activated by a nonlinear activation function (see figure 5.2) . The process to obtain appropriate values for the weights in each layer is termed training. MLP utilizes a supervised learning technique called *backpropagation* for training the network.

In back propagation an input training sample is propagated through the network. At the output the obtained values are compared to the target values by means of an error function (usually mean squared error or MSE). The network error is then minimized using a method called *stochastic gradient descent* although other methods are also available. The optimal value for each weight is that at which the error achieves a global minimum.

3.5 Estimation methods.

To evaluate the performance of a classifier algorithm on a data set with respect to a specific score (usually accuracy) is good idea to keep a fraction of the available data for testing purposes after the training step. Training and testing on the same data would make the classifier fail to predict anything useful on data not previously seen. This situation is called overfitting and prevent the classifier from generalizing properly on new data sets. To overcome this problem when a limited amount of samples is available we have at our disposal several techniques:

- Hold-out
- Cross-validation
- Random subsampling
- Bootstrapping.

Hold-out. In the holdout (train-test split) method, we randomly assign data points to two sets, usually called the training set and the test set, respectively. The size of each of the sets is arbitrary although typically the test set is smaller than the training set. We then train on the train set and test on the test set.

Cross-Validation. In cross-validation, we divide the available data into K equal sized parts. We leave one part k and fit the model to the remaining $k-1$ parts combined. Then we obtain predictions on the left-out part. This procedure is done for each part $k = 1, 2, ..K$ and then the results are combined. If $k = n$ where n is the number of data samples or examples then the *leave-one-out* cross-validation is obtained.

Bootstrapping. The bootstrap approach allow us to obtain distinct data sets by repeatedly sampling observations from the original data with replacement. Each of these bootstrap data sets is created by sampling with replacement , and is the same size as our original dataset. As a result some observations may appear more than once in a given bootstrap data set and some not at all.

Random subsampling. This method, also known as Monte Carlo cross-validation, randomly splits the dataset into training and validation data. For each such split, the model is fit to the training data, and predictive accuracy is assessed using the validation data. The results are then averaged over the splits. The advantage of this method (over k-fold cross validation) is that the proportion of the training/validation split is not dependent on the number of iterations (folds). The disadvantage of this method is that some observations may never be selected in the validation subsample, whereas others may be selected more than once. In other words, validation subsets may overlap. This method also exhibits Monte Carlo variation, meaning that the results will vary if the analysis is repeated with different random splits.

As the number of random splits approaches infinity, the result of repeated random sub-sampling validation tends towards that of leave-p-out cross-validation.

In our experiments we always relied mainly on cross-validation for assessing the performance of the different classifiers used in the classification experiments.

3.6 Multiresolution approaches.

In the following sections we are going to present experiments carried out using texture operators applied in the analysis of cervical images. The images are going to be studied from different aspects such as scale, frequency and orientation.

The common steps we follow in a typical multi-resolution scheme for texture analysis is as follows:

1. **Pre-processing**. Images need to be normalized in some sense for better comparison or classification. Denoising algorithms (despeckling) are applied in this stage.

2. **Decomposition** A decomposition is performed with some suitable transform (Wavelets, Fourier, Contourlets, Gabor).

3. **Feature Extraction** Texture features are calculated from the transformed images and combined. Feature selection can also be used if the dimensionality is very large.

4. **Classification** A classifier is trained to perform predictions on new data. More than one classifier can be tested and classification scores are compared.

3.6.1 Contourlet based image classification.

In this experiment we use the Contourlets transform to analyze the cervical images at several scales and orientations. A set of texture features is then calculated from the coefficient resulting from the image decomposition using this transform.

Preprocessing First, squared sections of size $512x512$ containing both ROIs were cropped from the original gray level images. This was done because all algorithms required squared shaped inputs with power of two sizes. Regions outside ROIs were assigned a gray value of 255 and not included in later calculations.

Images were also normalized to have zero mean and unit variance to mitigate the effect of gain and contrast variations. We tried to include all cervix lips region and not small patches to observe the documented variation in gray level during cervical ripening.

In order to diminish the influence of speckle noise in our US images several despeckling filters were tested: Linear, wavelet based and Non linear. The best classification results were achieved by using the linear filter with a 5 x 5 pixels sliding window.

Feature Extraction Contourlet decomposition was performed only to two levels and 0, 2, and 4 orientations. We used the *"9-7"* pyramidal filter and the*"pkva"* directional filter as shown in figure 3.6 . A total of 24 detail subimages per image were obtained. For each detail coefficient matrix, first and second order statistics were calculated. Settings for GLCM were $d = 1$, $\theta =$ 0°, 45°, 90°and 135° and

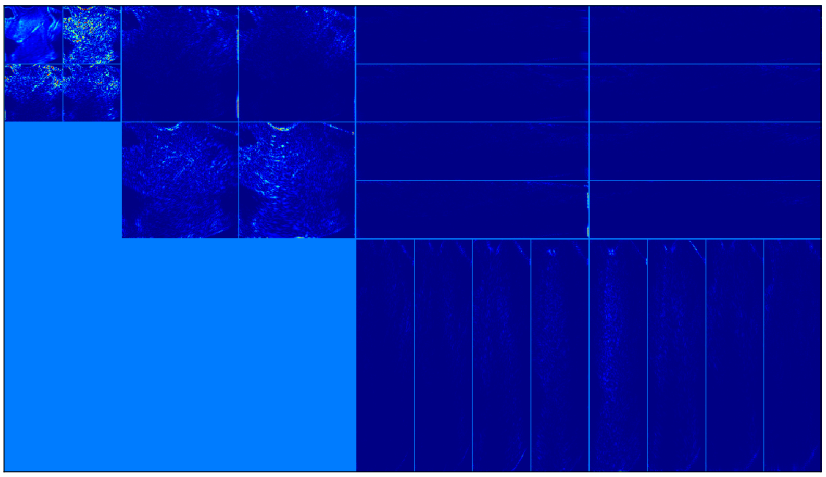


Figure 3.6: An example of contourlets decomposition of an ultrasound image up to three levels, using 0,2 and 4 orientations. For pyramidal decomposition the 'pkva' filter was used,and a '9/7' as a directional filter.

Table 3.3: Features with the highest discrimination power

Operator	Feature
FOS	Mean, median, standard dev.
GLCM	Energy,contrast, sum variance, sum entropy,correlation,
GLDM	Constrast,energy,entropy,mean

orientations . Matrices resulting for these orientations were combined and mean values and ranges of the measures were used.

Feature selection We used 4 first order statistic features, 14 features from GLCM along with their ranges, 5 features from GLDM. The application of these first and second order statistical methods to the coefficient of both transforms produced a large amount of features. This fact can be disadvantageous due to the high demand in computing power to carry out calculations and can also be detrimental to the overall algorithm performance due to presence of highly correlated features. In order to reduce the amount of features used for classification, a feature selection scheme was implemented. We used two methods: the Sequential Floating Forward Search algorithm (SFFSA), and the Sequential Backwards Search Algorithm (SBSA). In this analysis we found the best combination of features for performing classification. A total of 50 features were selected. In our analysis we found that the second order statistics features proved to be more helpful in the classification task, followed by FOS features. The most useful features for classification are summarized in table 3.3.

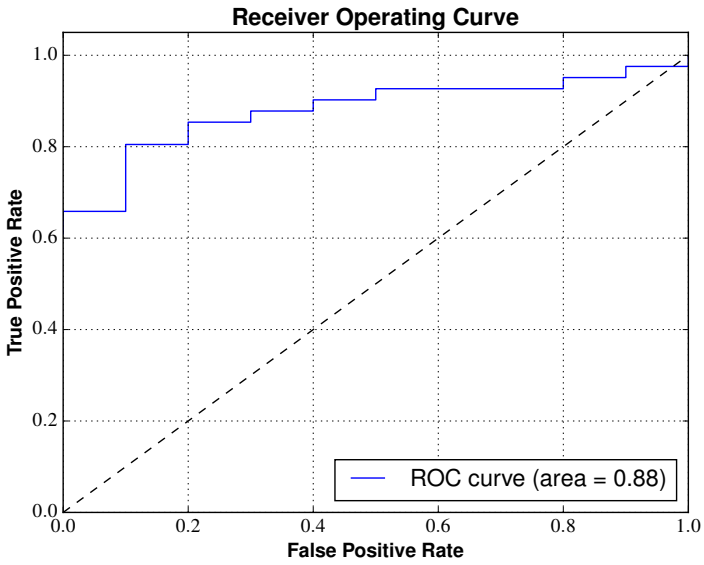


Figure 3.7: ROC curve obatined .

Classification For the classification stage we used a multi-layer perceptron with a hidden layer of 50 nodes. Crossvalidation using stratified k-fold (k=8) was performed on the data. With these settings an accuracy of about 82 % was obtained. ROC curves along with AUC values are shown in figure 3.7.

3.6.2 Multiscale Local Binary Patterns

The Local Binary Pattern texture descriptor is used here to analyze our images at different scales [16]. This is done by decomposing the images using the Wavelets transform and then using the LBP descriptor on every resultant (approximation) image.

Pre processing As a first step, a normalization of the images was performed by the 3-sigma method in which all pixels in an image are restricted to be within the interval $\mu \pm \sigma$ where μ is the mean gray value and σ is the standard deviation. Then we created binary masks from the provided ROI coordinates, a total of two ROIS were processed during the mask creation. (see fig, 3.8). These mask were also transformed by wavelets in such a way that they can be applied in every scale.

To allow for multi-resolution, the image is first decomposed in a pyramidal way using the wavelets transform. The Daubachies IV wavelets was utilized during decomposition. Only the approximation image was processed in this experiment. The LBP operator is applied to the different scale version of the image, by using

Table 3.4: Area under the curve (AUC) parameters obtained using the different mappings for the LBP.

Mapping	$P = 4$	$P = 6$	$P = 8$
riu2	0.83333	0.87037	**0.90741**
u2	0.72222	0.72222	0.7037
ri	0.83333	0.7963	0.66667

the calculated mask the points inside each ROI are retrieved. LBP histograms are then calculated for each ROI and from each resolution. Finally the histograms are all concatenated to form a combined histogram that represent the image under analysis.

Choosing the right mapping. In our implementation for the multi-resolution LBP, we tested neighborhood size of $P = 4$, 6, 8 or 16. We let $R = 1$ for all calculations since the resolution is changed by down-sampling operations carried out by the wavelets transform. We tried three different mappings (see section 2.3.1.2) for the LBP, uniform ('u2'), rotation invariant ('ri') and uniform-rotation invariant ('riu2') in a image subset to determine the most appropriate mapping for analysis. In table 3.4 a summary of the AUC values obtained using the different mapping, as shown, the best classification performance was achieved by using the 'riu2' mapping. For the case $P = 16$ the performance decreases again, thus $P = 8$ is considered optimum.

Classification The support vector machine classifier was used for classification. A Gaussian kernel was set for the SVM with a cost parameter of 0.8. It was found in the experiments that the performance of the classifier improves if the histograms were normalized to be zero mean and unit variance.

Two types of classifications were performed:

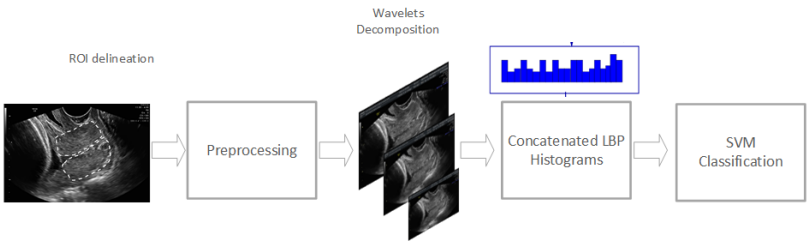


Figure 3.8: The different steps in the multiresolution methodology for LBP analysis of TVU images.

- One classification using four models, one for each cervical lip (L1 anterior , L2 posterior) of two classes ('Vaginal','Cesarean') .i.e VL1, VL2,CL1,CL2
- Classification of the cervix image using two models, one model for each class.

The first classification scheme was intended to study if the anterior and posterior lips show meaningful differences for each class because according to the previous works there are differences in the echogenicity of both lips. For error calculation we performed k-fold cross-validation on our dataset, with k = 8 folds.

A total of 54 files were used during classification experiments. The data set was divided into two groups one containing the patients with delivery within 24 hours (ripening success) and 36 hours, i.e those with stimulation stage and including only cases from high and medium category. Results for the first group (43 images) are summarized as follows:

1. For the individual lip models, we obtained 88.9%, 88,9%, 66.7% and 0% of correct identification for VL1,VL2,CL1,CL2. This might suggest that the anterior lip possesses more discriminant texture attributes.
2. For the global model, an accuracy of 80% is achieved (AUC 77%)

Results show that apparently the anterior lip of the cervix experience more echogenicity changes through the ripening process. This is probably due to the fact that the anterior lip is normally the first in the path of the ultrasound beam. This causes the posterior lips to receive less ultrasound power when a hard cervix is analyzed.

The percentage of good classification for the first group was 82%. When including the cases of the second category the performance decreases to almost 77%.

3.6.3 Multiscale Center-Symmetric Local Binary Pattern using Gabor filterbanks.

Another way of studying the texture of our images at different frequency scales and orientation is by means of a Gabor filter bank. These filter banks have been previously combined with some variants of LBP [3, 13, 26] for texture analysis. Here we are going to use the center symmetric local binary pattern (CS-LBP) due to its small feature size.

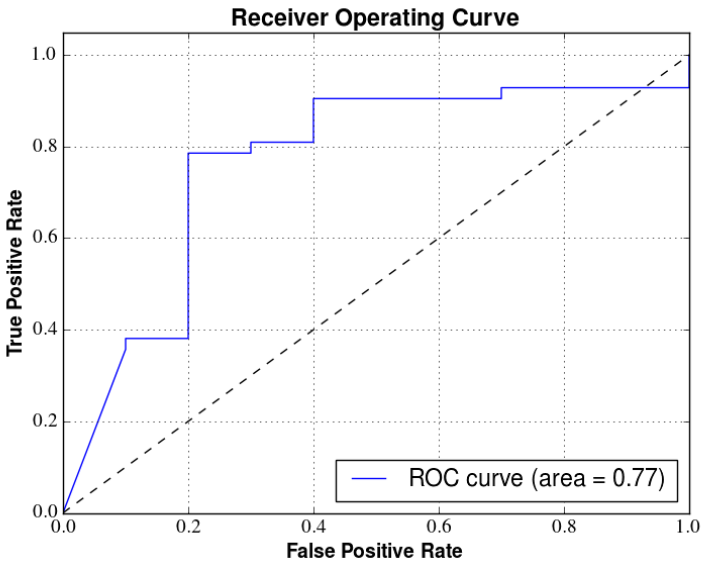


Figure 3.9: ROC curve for the Multiscale LBP method and using all patients from both groups.

The Center Symmetric Local Binary Pattern. The center symmetric local binary pattern is another modification of the original LBP descriptor [9]. The original implementation of LBP produced very long histograms and its feature is not robust on flat images. In CS-LBP instead of comparing the gray level value of each pixel with the center pixel, the center symmetric pairs of pixels are compared, see figure 3.10. CS-LBP is closely related to gradient operator. It considers the grey level differences between pairs of opposite pixels in a neighborhood. So CS-LBP take advantage of both LBP and gradient based features. It also captures the edges and the salient textures and it is less affected by noise.

To increase the operator's robustness in flat areas, the differences are thresholded at a typically non-zero threshold T. The histogram of CS-LBP values for an image I is stored as its feature. Three parameters have to be set during CS-LBP analysis: radius R, number of neighboring pixels N, and threshold on the gray level difference T.

Gabor filter bank. The design of a filter bank consist of a proper set of values for the filter parameters. Choosing optimum parameters for the filter bank is not a trivial task. By optimum, it is meant the ones providing the highest texture discriminating features. Due the many variables involved in the selection the search space is usually big. For this reason a complete search it is not advisable and

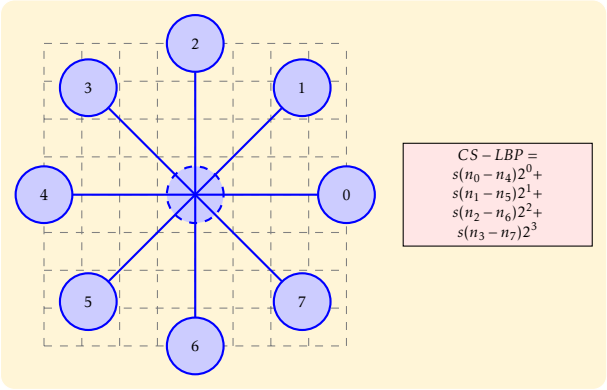


Figure 3.10: Example of CS-LBP feature calculation for a neighborhood $P = 8$ and radius $R = 1$. The n_i denotes the ith neighbor and $s()$ is a thresholding operation.

some heuristics are necessary. Some good methods to cope with this problem are based on genetic algorithms [13] or simulated annealing (ISA) [21]. In this work the artificial bee colony algorithm is used to find optimal or nearly optimal set of parameters.

The design of the filter bank involves the use of the following expressions to calculate the main parameters :F, η, γ, B_f, B_t.

$$\gamma = \frac{1}{\pi}\left(\frac{2^{B_f}+1}{2^{B_f}-1}\right)\sqrt{B_f \ln 2} \tag{3.11}$$

$$\eta = \frac{1}{\pi}\left(\frac{\sqrt{B_f \ln 2}}{\tan\left(\frac{B_t}{2}\right)}\right) \tag{3.12}$$

$$F_i = 2^{iB_f} F_0 \tag{3.13}$$

where B_f is the filter bandwidth in octaves, $B_t = \frac{2\pi}{N}$ is the angular spacing between different filters (N is the number of different orientations) and F_0 is the minimum normalized frequency to be considered during analysis, the maximum frequency is usually set as 0.4 or 0.5. Parameters η and γ are the same in equation 3.4. All of these parameters are chosen in such a way that the filter bank cover all the frequency domain. An example of this frequency partitioning performed by a filter bank can be observed in figure 3.11.

Artificial Bee Colony Algorithm (ABC). The artificial bee colony algorithm is an example of the swarm intelligence algorithm class. It was proposed by Pham [19] and try to imitate the food foraging behavior of swarms of honey bees. According to their authors ABC is applicable to both combinatorial and functional

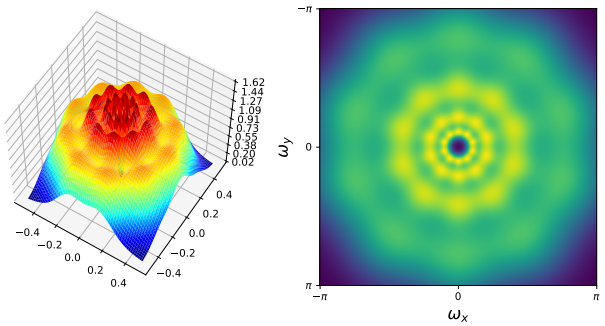


Figure 3.11: An example of a Gabor filter bank frequency partitioning. Parameters for the filters are: $N = 12$, $F_0 = 0.05$,$F_{max} = 0.5$ and $B_f = 1$

optimization problems. In ABC there exist three types of bees: employed bees, onlookers and scouts. ABC process requires cycle of four phases: initialization phase, employed bees phase, onlooker bees phase and scout bee phase.

Employed bees search food in the vicinity of the food source stored in their memory. They share food information with onlooker bees which tend to select food sources from those found by the employed bees. The source with the highest fitness (quality) is assigned a higher probability to be selected by the onlooker bees than others with lower quality. Scout bees are translated from a few employed bees, which abandoned their food sources after a predefined number of attempts and search now ones.

In the ABC algorithm, the first half of the swarm consists of employed bees, and the second half constitutes the onlooker bees. The number of employed bees or the onlooker bees is equal to the number of solutions in the swarm

During initialization, a randomly distributed initial population of SN solutions (food sources) is generated (equation 3.14), where SN denotes the swarm size. Each solution $x_i (i = 1, 2, ..., SN)$ is a D-dimensional vector, where D is the number of variables in the optimization problem and x_i represents the$^{\text{th}}$ food source in the population.

$$x_i^j = x_{min}^j + rand(0,1)(x_{max}^j - x_{min}^j), \forall_j = 1, 2.., D \tag{3.14}$$

$$v_{i,j} = x_{i,j} + \phi_{i,j}(x_{i,j} - x_{k,j}) \tag{3.15}$$

$$p_i = \frac{fit_i}{\sum_{i=1}^{SN} fit_i} \tag{3.16}$$

$$x_i^j = x_{min}^j + rand[0,1](x_{max}^j - x_{min}^j), \forall_j = 1, 2.., D \tag{3.17}$$

Each employed bee generates a new candidate solution in the neighborhood of its present position (equation 3.15) where $x_{k,j}$ is a randomly selected candidate ($i \neq k$) solution k, is a random dimension index selected from the set, and $\phi_{i,j}$ is a random number within [-1,1].

After the new candidate solution $v_{i,j}$ is generated its fitness is checked, if its value is higher than of its parent x_i^j then update it with $v_{i,j}$ otherwise keep the current value.

Once all employed bees complete the search process, they share the information of their food sources with the onlooker bees. An onlooker bee evaluates the nectar information taken from all employed bees and chooses a food source with a probability related to its nectar amount. This probabilistic selection is really a roulette wheel mechanism (equation 3.16), where fit_i is the fitness value of the i$^{\text{th}}$ solution in the swarm,we use$fit = e^{\frac{-c_i}{\bar{c}}}$, with c_i as the current cost and $\bar{c}$ is the average cost among all solutions.

As seen, the better the solution i, the higher the probability of the i$^{\text{th}}$ food source selected. If a position cannot be improved over a predefined number (called limit) of cycles, then the food source is abandoned. Assume that the abandoned source is x_i, then the scout bee discovers a new food source using equation 3.17 where and x_{min}^j and x_{max}^j are lower and upper boundaries of the j$^{\text{th}}$ dimension, respectively.

Since the ABC algorithm was developed for optimization of continuous functions it has been modified to work with discrete type values. In particular equations 3.14, 3.15, 3.17 have the form:

$$k^j = randint(0, N_j) \tag{3.18}$$

$$x_i^j = x[k^j] \tag{3.19}$$

$$v_{i,j} = x_{i,j} + round(a_j \phi_{i,j}(x_{i,j} - x_{k,j})) \tag{3.20}$$

Where N_j is the number of different discrete values of each variable. a_j controls de displacement of the variable. If the displacement goes beyond maximum values in each variable range then the maximum value is used.

Optimum values for our filterbank . The modified ABC algorithm was used to find optimal or nearly optimal set of parameters for our Gabor filter bank. The search space is described in table 3.5.

The cost function utilized was the distance between two model (vectors) representing each category (cesarean, vaginal). These vectors are actually the average of histograms calculated using the concatenated CS-LBP texture descriptor histograms at each scale and orientation and for each ROI (ANT,POST). The distance metric used in the calculation was *histogram intersection*. Vector a_j was set to $a_j = [1, 0.5, 0.0, 0.5, 0.2]$. Using this methodology, the optimum values found are summarized in table 3.6.

Pre-processing. The images were normalized to have zero mean and unit variance. This is needed since we don't want the DC component to be present when performing convolution. The images were of size $512x512$ and included two ROIs corresponding the anterior (ANT) and posterior (POST) regions.

Image decomposition. The filter bank with the optimal parameters was used for decomposing the images into $k*N$ different components where $k = \frac{1}{B_f}\log_2\left(\frac{F_{max}}{F_0}\right)$. Every component is then processed using the CS-LBP.

A binary mask is applied to each processed component to select only the points inside the ROIs. With the obtained pixels a histogram of 2^{np} bins is created. $np = \frac{P}{2}$ is the number of pixel pairs used.

Classification. Because obtained vectors were rather high-dimensional a feature selection step was introduced before feeding data into classifiers. Only the best 200 features were selected and sent to a multilayer perceptron (MLP) with a hidden layer of 100 nodes. The area under the curve (AUC) obtained was 0.79.

3.6.4 Multi-Frequency Resolution GLCM-LBPV using circular Gabor filters.

Local Binary Pattern despite being a powerful texture descriptor discard the spatial relationship between LBP codes when generating the histograms. In this sec-

Table 3.5: Parameter search space.

Parameter	Values
N	4 , 6, 8, 10, 12,14
B_f	0.5, 0.75, 1, 1.25, 1.5
F_0	0.05, 0.10, 0.15, 0.2, 0.25
F_{max}	0.35, 0.40, 0.45, 0.5
T	0.1, 0.2, 0.3, 0.5

Table 3.6: Optimal values found using ABC.

N.Orient	B_f	F_0	F_{max}	Threshold
14.	0.75	0.05	0.5	0.1

tion we analyze our dataset at different frequency scales obtained by passing the image through a circular Gabor filter bank. Then the LBP descriptor and the co-ocurrence matrix of the generated LBP images at each scale is calculated and used as features for classification [22].

Co-ocurrence Local Binary Pattern A method that uses that information by creating a Gray Level Co-occurrence Matrix from LBP images is described in [15]. In this experiment we investigate if the inclusion of the spatial relationship between GLCM entries improve the classification rate of a LBP-based algorithm.

We do not use the auto correlation methodology proposed in the reference to create the GLCM but calculate the codes in a pointwise manner using a sliding windows of 3x3 pixels. This can be done because we only process points inside predefined ROIs. Only four neighbors are considered in the calculation (north, east, west and south) and a distance $\delta = 1$. The GLCM is subsequently unfolded into a histogram

Pre-processing Normalization of the images was performed again by the 3σ method in which all pixels in an image are restricted to be within the interval $\mu \pm 3\sigma$, where μ is the mean gray value and $\sigma$is the standard deviation.

Image decomposition. After pre-processing, all images were filtered using a circular Gabor filter bank. For the filter bank we set $f_0 = 0.15$ and $f_{max} = 0.5$, the bandwidth B was set to one octave. This gives 3 different scales or frequency bands for analysis as shown in table 3.7.

For each scale we calculate the Local Binary Pattern histogram and Co-occurrence Local Binary Pattern histogram of the selected ROIs (we considered a total of 8 ROIs used in pairs). A neighborhood size of $P = 8$ and $R = 2$, were utilized as well as a uniform-rotation invariant mapping which produces LBP histograms with

Table 3.7: Frequency scales used for filtering.

	F	F_c	σ
1	0.15	0.112	5
2	0.3	0.225	2.5
3	0.6	0.450	1.25

length 10.

The histogram corresponding to each image (with 2 ROIs each) are concatenated to form a single histogram that represent the image. In the case of GLCM-LBP we concatenate the LBP histogram and the GLCM-LBP histogram together. Features were further processes by normalizing each input vector individually to have unit norm.

Previous to the classification stage we perform a feature selection step where we retain only the 60/200 (LBP/GLCM+LBP) most informative features.

Classification: For classification we used two classifiers: a K-nearest neighbor classifier with k = 2 and a Neural Network with 20 nodes hidden layer. For these classifiers we use the implementation in the scikit-learn Python package [18] , as we used Python to program all the functions used for analysis.

For the K-nearest network classifier we tried several distance metrics such as Euclidean, CityBlock, Canberra and Minkowski. Best classification results were achieved using the Canberra distance.

For the error calculation, we performed cross validation using a stratified K-fold (K=8) scheme due to the unbalanced classes. ROC curves for the best region of interest (ANT-POST) are shown in figure 3.12.

Among the different ROIs used in this study the one corresponding to the whole cervical lips provides the best classification result followed by the center ROI (B1, B2). The BPNN for this ROI selection provides a AUC score of 0.83.

Our results show that it is possible to differentiate by means of image processing techniques a ripe cervix and therefore the type of outcome from labor induction, with an accuracy of about 92% (BPNN). This accuracy was obtained when using the whole dataset. The spatial information of the LBP codes obtained from the GLCM of the LBP image improve the classification rate as it was observed from the ROC curves for both classifiers.

3.6.5 Including Contrast information using LBPV Analysis.

In LBP analysis an image is considered to be comprised of two aspects : the pattern and the constrast. Until now we have used pattern information without paying attention on contrast variations in the image. The LBP operator was designed to be invariant to intensity, so it discards any contrast information. As we saw in chapter 2 there is a complementary contrast measure in LBP analysis, the $VAR_{P,R}$ operator (equation 2.15).

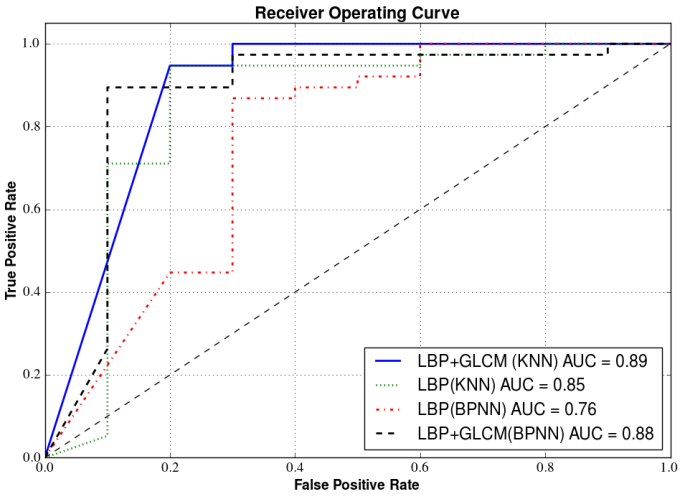


Figure 3.12: ROC curves and area under the curve (AUC) for the LBP and GLCM-LBP schemes using both classifiers.

It is known that contrast variations exist in the texture of cervical lips region in response to ripening. In this section we analyze the images including a contrast measure.

3.6.5.1 The Joint LBP/VAR distribution.

One way to include contrast information for classification, is just to concatenate the VAR histogram to the LBP histogram to form a long vector. However a better approach consist in calculating the joint LBP/VAR histogram which constitutes an approximation to the joint probability density function (PDF) of both LBP and VAR. To construct a LBP/VAR histogram at several scales we first decompose the image by means of the wavelets transform. The LBP and VAR operators are then applied to the different scaled versions of the image and then 2D histograms are calculated at each resolution for every defined ROI. The bins for the VAR histograms are calculated by taking evenly distributed sections in the cumulative distribution of gray levels of all images in the data set. Only those pixels contained within each ROI were included in the histogram calculations. Histograms corresponding to different ROIs were then concatenated to form a combined histogram representing the image, a sample is shown in fig 3.13.

By using these models a classification scheme can be designed including some sort of distance to compare a sample image to both models. Membership to one of the classes can be assigned by selecting the shorter distance in a nearest neighbor

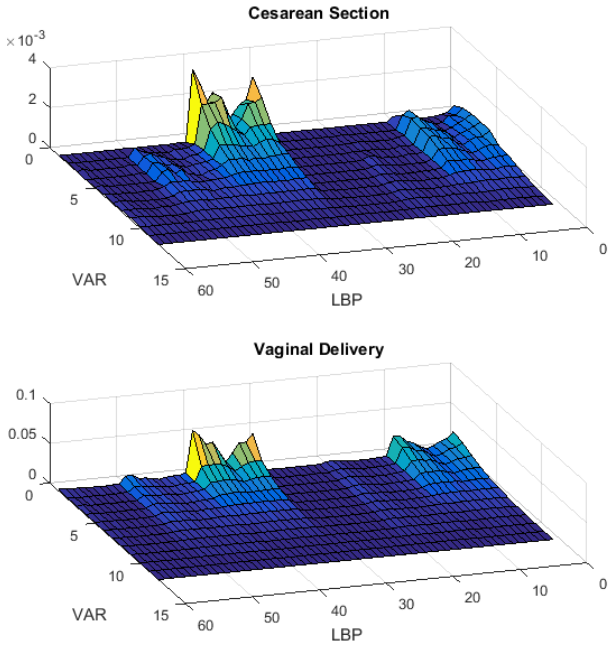


Figure 3.13: Sample of 2D Histogram Models for the ANT/POST ROIS of both lips. The histogram were obtained using the *riu2* mapping and 12 bins for the VAR histogram at 3 decomposition levels.

fashion. A drawback of this method is that you have to train the algorithm with all the images in order to calculate the bins in the VAR histogram.

A simpler solution is presented in [7] where an operator termed LBPV is described. The LBPV includes contrast information using a weighting scheme during the LBP histogram calculation, instead of simply counting the appearances of a LBP code, the LBPV weigths the contribution of a determined code, by the value of the VAR operator.

Let I and image of size MxN and a histogram of LPV values having K bins, then for each k in K we calculate :

$$LBPV_{P,R} = \sum_{i=1}^{N}\sum_{j=1}^{M} w\left(LBP^{riu2}_{P,R}(i,j),k\right), \quad k \in [0,K] \tag{3.21}$$

$$w\left(LBP^{riu2}_{P,R}(i,j)\right) = \begin{cases} VAR_{P,R} & LBP^{riu2}_{P,R}(i,j) = k \\ 0 & \text{Otherwise} \end{cases} \tag{3.22}$$

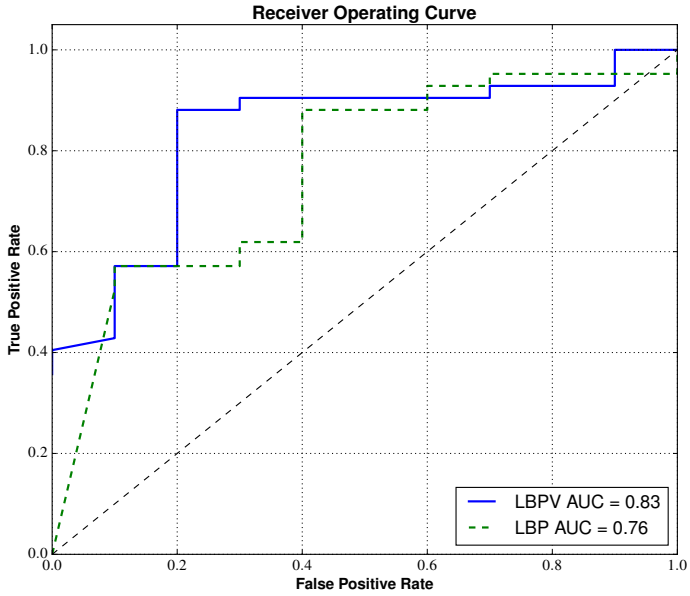


Figure 3.14: ROC curves comparing classification using LBP alone and after including contrast information (LBPV). A good improvement (about 9 %) is obtained.

3.6.5.2 Classification using LPBV.

In order to verify the contribution of the contrast information for discrimination, we performed an experiment using the same image set as before. The preprocessing steps are as in section 3.6.2. Wavelets Daubachies type IV were used for decompostion of the images at three levels, a neighborhood of $P = 8$ and a radius $R = 1$ for LBP.

For classification we used again a multilayer perceptron with 25 hidden nodes, and a stratified k-fold crossvalidation with $k = 8$. Using these settings an accuracy of 84.6% was obtained and an AUC of 83% (see figure 3.14) what is a good improvement over the simple LBP.

3.7 Chapter Summary

In this chapter we tested several multi scale / multi resolution schemes to investigate the utility of such tools in the problem of classifying cervical status and predicting the outcome of a labor induction process. Several aspects were considered: scale, frequency and orientation. The results from our experiments suggest that there is no a strong preferred orientation in cervical texture and that frequency scale is a more important parameter as demonstrated using the circular gabor fil-

ters. Additionally, including contrast information can increase the performance of the proposed algorithms, nevertheless it is advisable to include some form of gray level normalization prior to LBPV calculations.

References

[1] B Balasuganya. "Contourlet Based Feature Extraction Method for Classification of Breast Cancer using Thermogram Images". In: *International Journal of Scientific & Engineering Research* 5.4 (2014), pp. 285–287.

[2] Daniel Barbosa, Dalila Roupar, and C. S. Lima. "Multiscale texture descriptors for automatic small bowel tumors detection in capsule endoscopy". In: *Discrete Wavelet Transforms - Biomedical Applications* 9 (2011), pp. 155–176.

[3] Chen Chen et al. "Gabor-Filtering-Based Completed Local Binary Patterns for Land-Use Scene Classification". In: *Proceedings - 2015 IEEE International Conference on Multimedia Big Data, BigMM 2015* (2015), pp. 324–329.

[4] Chung Ming Chen, H. H S Lu, and Ko Chung Han. "A textural approach based on gabor functions for texture edge detection in ultrasound images". In: *Ultrasound in Medicine and Biology* 27.4 (2001), pp. 515–534.

[5] J Doublet, O Lepetit, and M Revenu. "Contactless palmprint authentication using circular Gabor filter and approximated string matching". In: *Signal and Image Processing (SIP 2007)*. 2007.

[6] I. Fogel and D. Sagi. "Gabor filters as texture discriminator". In: *Biological Cybernetics* 61.2 (1989), pp. 103–113.

[7] Zhenhua Guo, Lei Zhang, and David Zhang. "Rotation invariant texture classification using LBP variance (LBPV) with global matching". In: *Pattern Recognition* 43.3 (2010), pp. 706–719.

[8] G.M. Haley and B.S. Manjunath. "Rotation-invariant texture classification using modified Gabor filters". In: *Proceedings., International Conference on Image Processing* 1 (1995), pp. 262–265.

[9] M Heikkilä, M Pietikäinen, and C Schmid. "Description of interest regions with center-symmetric local binary patterns". In: *Computer Vision, Graphics and Image Processing.* 2 (2006), pp. 58–69.

[10] Mahamadou Idrissa and Marc Acheroy. "Texture classification using Gabor filters". In: *Pattern Recognition Letters* 23.9 (2002), pp. 1095–1102.

[11] Arshad Javed, Wang Yin Chai, and Narayanan Kulathuramaiyer. "Contourlet transform based enhanced brain MR image segmentation". In: *Proceedings of the IIEEJ Image Electronics and Visual Computing Workshop 2012*. 2012.

[12] Stamos Katsigiannis, Eystratios G. Keramidas, and Dimitris Maroulis. "A Contourlet Transform Feature Extraction Scheme for Ultrasound Thyroid

Texture Classification". In: *Engineering Intelligent Systems, Special issue: Artificial Intelligence Applications and Innovations* 18.3 (2010).

[13] Ma Li and R. C. Staunton. "Optimum Gabor filter design and local binary patterns for texture segmentation". In: *Pattern Recognition Letters* 29.5 (2008), pp. 664–672.

[14] Li Ma, Yunhong Wang, and Tieniu Tan. "Iris Recognition Using Circular Symmetric Filters". In: *16th International Conference on Pattern Recognition (ICPR'02)* (2002), pp. 414–417.

[15] Ryusuke Nosaka, Yasuhiro Ohkawa, and Kazuhiro Fukui. "Feature extraction based on co-occurrence of adjacent local binary patterns". In: *Advances in Image and Video Technology* (2012), pp. 82–91.

[16] Pablo J Vásquez Obando et al. "Multi-resolution Local Binary Pattern for Assessing Cervical Ripening". In: *2015 International Conference on Frontiers of Signal Processing (ICFSP 2015)*. 2015.

[17] Pablo Vásquez Obando et al. "Multiscale Texture Analysis of Cervical Tissue for Labor Induction". In: *2015 IEEE International Workshop of Electronics, Control, Measurement, Signals and their application to Mechatronics*. 2015.

[18] F Pedregosa and G Varoquaux. "Scikit-learn: Machine learning in Python". In: *The Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[19] D.T. Pham et al. "The bees algorithmA novel tool for complex optimisation". In: *Proceedings of the 2nd International Virtual Conference on Intelligent Production Machines and Systems* (2006), pp. 454–459.

[20] Bart M. Haar Romeny. *Front-End Vision and Multi-Scale Image Analysis: Multi-scale Computer Vision Theory and Applications, Written in Mathematica*. 2003, p. 470.

[21] Du-Ming Tsai, Song-Kuaw Wu, and Mu-Chen Chen. "Optimal Gabor filter design for texture segmentation using stochastic optimization". In: *Image and Vision Computing* 19.5 (2001), pp. 299–316.

[22] Pablo Vásquez O et al. "Labor Induction failure prediction based on B-Mode Ultrasound Image Processing using Local Binary Patterns". In: *2016 International Conference on Optoelectronics and Image Proccessing*. 2016.

[23] T.P. Weldon and W.E. Higgins. "Design of multiple Gabor filters for texture segmentation". In: *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings* 4.4 (1996), pp. 2243–2246.

[24] Dengsheng Zhang et al. "Content-based Image Retrieval Using Gabor Texture Features". In: *IEEE Transactions PAMI* 3656 LNCS (2000), pp. 13–15.

[25] J Zhang, T Tan, and L Ma. "Invariant texture segmentation via circular Gabor filters". In: *Pattern Recognition, 2002. Proceedings. 16th International Conference on* 2.2 (2002), pp. 901–904.

[26] Lin Zhang, Zhiqiang Zhou, and Hongyu Li. “Binary Gabor pattern: An efficient and robust descriptor for texture classification”. In: *Proceedings - International Conference on Image Processing, ICIP*. 2012, pp. 81–84.

Chapter 4

Effects of illumination variations and noise on cervical US image classification.

In this chapter we study cervical ultrasound images from different aspects that are relevant for the successful application of image processing algorithms. Particularly we devote this chapter to describe what it has been found related to image gray level statistics, noise and their relationship with cervical tissue transformation during pregnancy. We also tested several image normalization schemes in order to find out if normalization plays a role in the accuracy obtained from LBP-based methods.

4.1 Echogenicity changes of the cervix during pregnancy.

As it has been mentioned in previous chapters, the chemical and structural changes the cervix experiments during pregnancy have been found to give raise to changes in gray level parameters of the US images. In studies [2, 3] analyzing cervix for preterm birth prediction the observed changes on echogenicity from the state of non-pregnancy to early and late pregnancy were:

1. The brightness of the texture decreases in the area of the external os (external opening of the cervix).
2. In the area of the inner os (internal opening) the contrast increases and homogeneity decreases.
3. In the area of external os the contrast decreases and homogeneity increases.

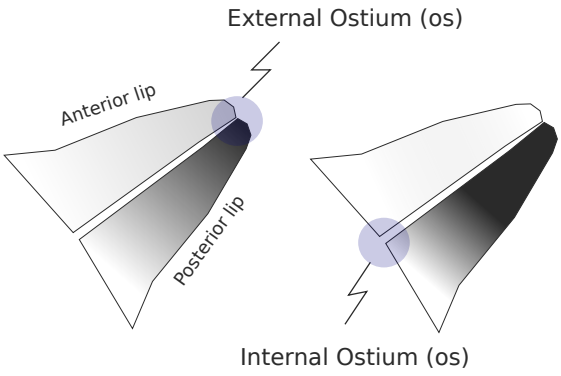


Figure 4.1: Echogenicity changes in the cervical tissue during pregnancy are depicted. Left, a pregnant cervix, to the right a non pregnant cervix. Observe the difference in gray levels between both lips. The difference in mean gray level is a measure of tissue consistency.

4. The greater the difference in echogenicity between anterior and posterior lips the harder the cervical tissue consistency.

In figure 4.1 a picture summarizing the above mentioned changes is presented. Here we are interested in knowing if these conclusions are also valid in the case of labor induction procedures. For this purpose we analyzed the 82 images in our database. Because we divided each lip into three ROIs (A,B and C) which contains the internal os (A), the external os (C) and the middle region (B) it is possible to assess if this changes occur by looking at the mean gray level of each region as it is done in next section.

4.2 Gray level statistics.

We calculated some statistical parameters of the gray levels corresponding to each defined ROI to determine how these change in each condition (cesarean vs vaginal delivery). We tested three parameters: mean, median and standard deviation from the whole set of gray level histograms. After analyzing the results, it was confirmed that mean gray level of anterior lip is higher that the posterior and that this difference is more accentuated in the case of cesarean section as shown in figure 4.2, they exhibit however high variability and it is frequent to encounter images for which the opposite is true.

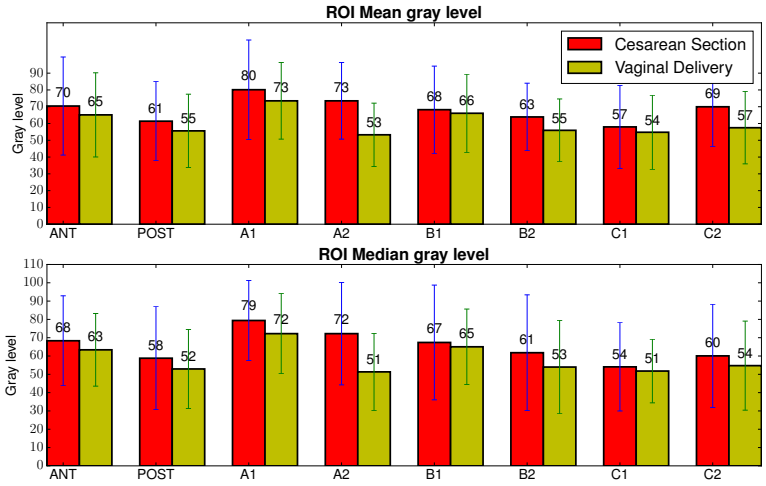


Figure 4.2: Gray level statistics of the image dataset. The eight ROIS and the two classes are shown.

It was also noticed that for both categories (vaginal vs cesarean) it is true that the external os region (C) has always a smaller gray level mean value compared to internal os (A).

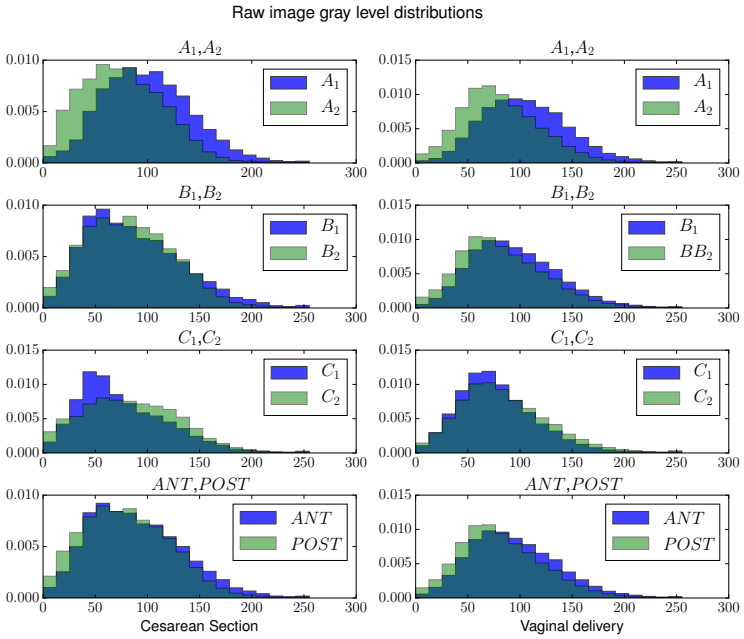


Figure 4.3: Distribution of gray levels by ROIs and category. The x axis corresponds to the gray level intensity (0-255) and the y axis to the probability of a particular gray level.

From all the tested ROIs A_1, A_2 (those corresponding to the internal os region) were found the ones that show more echogenicity differences (se also figure 4.3).

4.3 Image Normalization

Although care has been taken during image acquisition for using the same ultrasound scanner settings, this is not always possible. As it can be observed from the previous section there changes in the gray levels of the images in the data set. One way to reduce the impact of different settings, or acquisition conditions is by means of image normalization (standardization) which aims to reduce these differences and make the image gray level lay on the same range to ease comparison and classification. Experiments were carried out with the following normalization schemes:

1. Histogram equalization.
2. Histogram normalization.
3. Contrast strecth normalization.
4. Unitary variance and zero mean.
5. Linear scaling using two reference values.
6. Linear scaling using two reference values and predefined output range.

The first four normalization methods are very popular and we will not describe them in detail, however the last two were specifically designed for our data set and we will present them thoroughly.

Histogram of an image represents the relative frequency of occurrence of the various gray levels in the image. The histogram gives primarily the global description of the image.

4.3.1 Histogram equalization.

Histogram equalization is a technique in which the gray scale of the image is adjusted so that the gray level histogram of the input image is mapped onto a uniform (flat) histogram, in which the percentage of pixels of every gray level is the same. Commonly the transformation is performed using the cumulative distribution function of the gray levels of the input image as follows:

$$h_n(i) = round\left(\frac{C(i) - C_{min}}{1 - C_{min}}(L-1)\right) \tag{4.1}$$

Here, C_{min} is the minimum value of the cumulative distribution function $C(i)$ in the image, i the bin index, L is the total number of gray level values (256) and $h_n(i)$ is the equalized histogram.

4.3.2 Histogram normalization.

The original image histogram is stretched, and shifted in order to cover all the gray scale levels in the image as follows:

$$I_n(i,j) = \frac{I_{max} - I_{min}}{h_{max} - h_{min}} * (I(i,j) - h_{min}) + I_{min} \tag{4.2}$$

If the original histogram of the initial image starts at h_{min} and extends up to h_{max} brightness levels, then we can scale up the image so that the pixels in the new image, $I_n(i,j)$ lie between a minimum level and a maximum level (0,$L-1$).

4.3.3 Contrast stretch normalization.

Contrast stretching is used to increase the pixel value range by rescaling the pixel values in the input image.

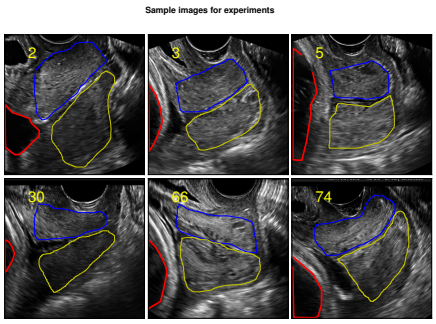
$$I_t(i,j) = I_{max}\frac{I(i,j) - I_{low}}{I_{high} - I_{low}} + I_{low} \tag{4.3}$$

The values for I_{low} and I_{high} are obtained by first searching for the minimum and maximum gray level values in the image and then calculating the average inside a small windows of $9x9$ pixels centered on these values.

4.3.4 Unitary variance.

This normalization perform subtraction of the image mean gray value $\bar{I}$ and division by the image standard deviation σ_i to make it zero mean and unitary variance.

$$I_n(i,j) = \frac{I(i,j) - \bar{I}}{\sigma_I} \tag{4.4}$$



N	Code	Labor outcome
1	0002	Vaginal delivery
2	0003	Vaginal delivery
3	0005	Vaginal delivery
4	0030	Cesarean section
5	0066	Cesarean section
6	0074	Cesarean section

Figure 4.4: Samples images showing the reference region (fetal head region) used during image normalization. The other region used for reference is the anterior lip.

4.3.5 Linear scaling using two reference values.

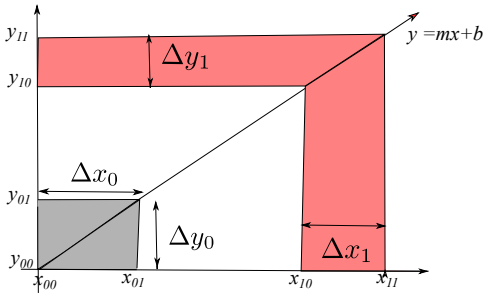
The image gray values are forced to fit within a pair of predefined values. Two regions that have potential to be used as reference gray levels are the fetal head and the anterior lip. The first has always low gray values due to the amniotic liquid present in the uterus. The anterior lips on the other hand is the most echogenic of the cervix. The median gray level value of the fetal head region and that of the anterior lip were used in our experiments as reference gray level values. The scaling is performed as follows:

$$I_n(i,j) = I_{max}\frac{I(i,j) - I_{low}}{I_{high} - I_{low}} + I_{low} \tag{4.5}$$

In last equation I_{low} is the median gray level calculated from the fetal head region and I_{high} the corresponding value of the anterior lip, $I(i,j)$ the input image and $I_n(i,j)$ the normalized output image.

4.3.6 Linear scaling using two reference values and predefined output range

A method for normalization that has been used for ultrasound images consist of using two reference values [11],[5] . These reference values are used to linearly transform the image in such a way that statistics of the ROIs after processing satisfy some criteria. Usually they consist of mean or median gray values of selected regions of interest in the image. These parameters are modified to lie around some reference values. As in the last section the fetal head region and the anterior lip region were used as references.



$$\Delta y_i = y_{i1} - y_{i0}$$
$$\bar{y_1} = \frac{y_{11} - y_{10}}{2}$$
$$\bar{y_0} = \frac{y_{01} - y_{00}}{2}$$
$$x_{10} \leq x_1 \leqq x_{11} \mapsto y_{10} \leqq y_1 \leq y_{11}$$
$$x_{00} \leq x_0 \leqq x_{01} \mapsto y_{00} \leqq y_0 \leq y_{01}$$

Figure 4.5: Linear transform and ranges proposed for fetal head region and anterior cervical lip. Here x_{1i} represents the range of anterior lip (R_1) values and x_{0i} the corresponding values for the fetal head region. y_{0i} and y_{1i} are the obtained output values.

The general idea with standardization is to reduce the variability of the gray value statistics. For example, as it can be seen from figure 4.2 the median value of R_1 fluctuates between 40 and 90. A sample set of 6 images representative of both classes were chosen to carry out experiments, see figure 4.4. For the selected sample images, an interval of 70-80 was set for the anterior lip (R_1) and 0-10 for the fetal head. Most of the time this transformation is carried out manually, however it is desirable that this transformation can be done automatically, when a large amount of images is to be processed. A linear transformation of the form $y = mx+b$, should provide such mapping as shown in figure 4.5.

In order to find appropriate values for m and b a grid search was performed in the variable space. For the optimization a cost function based upon a quadratic error is considered as in equation 4.6

$$\begin{aligned} e_1 &= (y_0 - \bar{y_0})^2 \\ e_2 &= (y_1 - \bar{y_1})^2 \\ e_{tot} &= e_1 + e_2 \end{aligned} \tag{4.6}$$

where y_0 and y_1 represent the output of the linear transformation for the fetal head and anterior lip regions respectively. On our tests we set $y_{00} = 0$, $y_{01} = 10$, $y_{10} = 70$ and $y_{11} = 80$. This gives $\bar{y_0} = 5$ and $\bar{y_1} = 75$ for the averages. The results after applying the linear transformation to each of six sample images is summarized in table 4.1.

Table 4.1: Results from the linear transformation applied to the sample set.The subscript n stands for normalized.

H	R_1	H_n	R_{1n}
5.000	72.000	5.000	75.000
22.000	75.000	17.000	71.000
5.000	62.000	5.000	75.000
17.000	62.000	16.000	72.000
16.000	108.000	8.000	85.000

4.4 Noise in TVU images.

US images are usually contaminated with speckle noise. Speckle noise has a random and deterministic nature as it is formed from backscattered echoes of randomly or coherently distributed scatterers in the tissue. It can be classified into four types depending on the scatterer density, the way they are organized within the resolution cell, the existence of deterministic elements influenced by the relative size of scatterers when compared to the wavelength of the ultrasound signal, result in four different types of speckle [6]:

1. *Fully developed*: large number of scatterers and non-existence of deterministic components, modeled by Rayleigh distribution;

2. *Fully resolved:* large number of scatterers and presence of deterministic components (for instance, specular reflection), modeled by Rician distribution;

3. *Partially developed*: small number of scatterers and non-existence of deterministic components, modeled by K distribution.

4. *Partially resolved*: small number of scatterers and presence of deterministic component, modeled by K-Homodyne distribution.

Although it is known that the cervix collagen fibrils are aligned in layers and that during ripening this fibrils become misaligned, a recent study on echo signals from ex-vivo cervix tissue [9] reported that this alignment could be possibly too weak to produce coherent component is the backscattered image. The mean length of collagen fibrils in the rat cervix is reported to be 2268 ± 77 nanometers [1] which is much smaller than the typical wavelength of TVU probes (0.3 mm using a 6.5 Mhz transducer). They also found that much of the cervical tissue can be classified as having sparse scattering sources yet specular reflections are expected.

In view of these facts a Rician distribution would better describe the cervical tissue. There are however more general and simpler distribution that can be used for modeling speckle, for instance the Nakagami distribution.

4.5 Speckle Noise Reducing Algorithms.

Speckle reduction in ultrasound imaging is desired mainly for two reasons: An improvement in image quality for visualization purposes to help human interpretation and as preprocessing step before segmentation or registration. It has been argued that speckle may contain diagnostic information and should be retained [10]. On the other hand there is always a detail lose when despeckling is performed. To investigate if image classification can benefit from speckle reducing techniques we tried some of the most popular and effective noise reducing algorithms. These are:

1. Anisotropic diffusion (AD).
2. Speckle Reducing Anisotropic Diffusion (SRAD)
3. Wavelets Bayesian shrinkage.
4. Linear filtering.

4.5.1 Anisotropic diffusion

The use of partial derivative equations in the context of noise removal started with the work of Perona and Malik [7]. The authors designed a filter based on the diffusion equation that apply smoothing depending on the image edges and their directions. Anisotropic diffusion is an efficient nonlinear technique for simultaneously performing contrast enhancement and noise reduction. It smooths homogeneous image regions, but retains image edges. Anisotropic diffusion is defined by the following equation:

$$\begin{aligned} \frac{\partial I}{\partial t} &= \operatorname{div}\left(c(\|\nabla\|).\nabla I\right) \\ I_{(t=0)} &= I_0 \end{aligned} \tag{4.7}$$

where I is an image, ∇ denotes the gradient, *div* is the divergence operator and $c(\|\nabla\|)$ is the diffusion coefficient that depends on the gradient magnitude $c(x,y,t) = g(I(x,y,t))$. This coefficient controls the rate of diffusion and is designed to preserve edges in the image.

A discrete form of the former equation is,

$$I_s^{t+\Delta t} = I_s^t + \frac{\Delta t}{|\bar{\eta_s}|} \sum_{p \in \bar{\eta_s}} c\left(\nabla I_{s,p}^t\right) \nabla I_{s,p}^t \tag{4.8}$$

where I_s^t is the sampled image, s denotes the current pixel position and Δt is the time step size, $\bar{\eta_s}$ represent the spatial neighborhood. $|\bar{\eta_s}|$ is the number of pixels in the window.

The authors proposed two implementations of the $g(.)$ function:

$$g(\nabla I) = e^{-(\|\nabla I\|/K)^2} \tag{4.9}$$

$$g(\nabla I) = \frac{1}{1 + \left(\frac{\|\nabla I\|}{K}\right)^2} \tag{4.10}$$

The parameter K is a positive gradient threshold parameter, known as diffusion or flow constant. In our implementation 8 neighbors were used and the discrete derivatives were calculated by means of convolution kernels.

4.5.2 Speckle reducing anisotropic diffussion

In the SRAD filter the gradient-based edge detector is replaced by the so-called instantaneous coefficient of variation [12]. For a four pixel neighborhood, the update equation and the coefficcient of variation are:

$$I_{i,j}^{t+\Delta t} = I_{i,j}^t + \frac{\Delta t}{|\bar{\eta_s}|} div\left[c\left(C_{i,j}^t\right) \nabla I_{i,j}^t\right] \tag{4.11}$$

$$C_{i,j}^2 = \frac{\frac{1}{2}|\nabla I_{i,j}|^2 - \frac{1}{16}\left(\nabla^2 I_{i,j}\right)^2}{\left[I_{i,j} + \frac{1}{4}\nabla^2 I_{i,j}\right]^2} \tag{4.12}$$

4.5.3 Wavelets Bayesian denoising.

Wavelets have been used extensively in denoising algorithms for images. The denoising algorithms usually perform thresholding of the wavelet coefficients, which have been affected by noise. During thresholding only large coefficients are retained and setting the remaining to zero. The way a threshold is chosen gives raise to various solutions from fixed to adaptive. In [8] a method using the stationary wavelets transform is presented. In this method the decomposition is performed to L levels and with each detail image $\mathbf{w} = \{w_1, w_2..w_L\}$ a binary mask is associated

$\mathbf{x} = \{x1_1, ..x_L\}$. In the masks $x_l = 0$ if w_l represents mainly noise and $x_l = 1$ if it contains useful information.

To estimate the mask values, a threshold value is evaluated.

$$\hat{x_{i,j}} = \begin{cases} 0 & |W_{i,j}||\hat{y_{i,j+1}}| \leq \hat{\sigma_{i,j}}^2 \\ 1 & |W_{i,j}||\hat{y_{i,j+1}}| \geq \hat{\sigma_{i,j}}^2 \end{cases} \tag{4.13}$$

where $\hat{\sigma_{i,j}}$ is the estimate of the standard deviation of noise at the resolution scale 2^j. It is calculated as the median absolute deviation (MAD) of the coefficients at a given detail image $\sigma_{i,j}^2 = MAD_j/0.6745$.

For denoising , wavelets shrinkage is applied to obtain an estimated of the true coefficients $\hat{y_l} = q_l w_l$ where $0 \leq q_l \leq 1$ is the shrinkage factor defined as

$$q_l = \frac{\xi_l \eta_l}{1 + \xi_l \eta_l} \tag{4.14}$$

In the last expression ξ_l is the likelihood ratio at the current position l and η_l is related to the spatial neighborhood $\partial(l)$.

$$\xi_l = \frac{p(m_l \mid 1)}{p(m_l \mid 0)}, \quad \eta_l = exp\left(\gamma \sum_{k \in \partial(l)} (2\hat{x_k} - 1)\right) \tag{4.15}$$

$$\tag{4.16}$$

The parameter γ controls the importance attributed to the local spatial neighborhood. These calculations are to be performed in a coarse to fine direction along the scales.

4.5.4 Linear Filtering.

This filter use local statistics (variance and mean) calculated in a neighborhood centered on the pixel of interest [4].

$$f_{i,j} = \bar{g} + k_{i,j}\left(g_{i,j} - \bar{g}\right) \tag{4.17}$$

where $f_{i,j}$ is the estimated noise-free pixel value, $g_{i,j}$ is the noisy pixels inside the sliding window, $\bar{g}$ is the local average of an $N_1 x N_2$ region around $g_{i,j}$, k is a weighting factor , with $k \in [0, 1]$ and i, j are the pixel coordinates. The k factor depends upon local statistics and can be calculated as,

Table 4.2: Summary of the different parameters used for the despeckling algorithms.

Filter	Parameters
AD	Δt=1, iterations = 5,K=10
SRAD	λ=2, iterations =5
WS	l = 3 , wavelet type = Daubachies type 4, γ=0.2

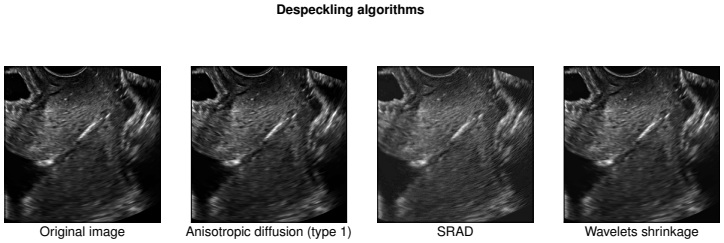


Figure 4.6: Comparison of resulting images after despeckling using three diffrent algorithms.

$$k_{i,j} = \frac{\sigma^2}{\left(\bar{g}^2\sigma_n^2 + \sigma^2\right)} \tag{4.18}$$

The parameters σ^2 and σ_n^2 correspond to the variance of the sliding window and the variance of the noise in the whole image respectively. This variance can be approximated by the following expression:

$$\sigma_n^2 = \sum_{i=1}^{p} \frac{\sigma_p^2}{\bar{g_p}} \tag{4.19}$$

where σ_p and g_p are the variance and the mean of the noise in the selected windows, respectively, and p is the index covering all windows in the whole image. If the value of $k_{i,j}$ is 1 (in edge areas), this will result to an unchanged pixel, whereas a value of 0 (in uniform areas) replaces the actual pixel by the local average $\bar{g}$ over a small region of interest.

A sample image (the one with code 002) was filtered using the algorithms presented in this section and the resulting images are shown in figure 4.6 The parameters used during the experiments are summarized in table 4.2.

4.6 Effect of filtering and Normalization on Classification.

In order to verify if filtering and /or standardization affects the classification results obtained using the methods described in the previous chapter, a dataset of 54 images was used. Two models representing both categories ,i.e. Vaginal Delivery or Cesarean Section, were constructed from the histograms obtained after processing the images using the multiscale circular Gabor filter Local Binary Pattern method. The distance between these two model is a way to measure the influence of the normalization methods on the classification accuracy; the larger the class separation, the better classification obtained.

Model creation. The models were built as the average of all histogram vectors belonging to each class. The LBPV method was chosen among the different method used so far as it is the one which is more sensitive to gray level intensity variations and can therefore show performance changes due to normalization.

After model creation several histogram distance metrics were used to assess a potential class separation. In our experiments we used Euclidean, Battachardya, Chisquare and Histogram Intersection distance metrics. The results from the analysis of the image dataset is summarized in table 4.3. According to these results the best normalization schemes are the histogram normalization and the normalization with two reference points, the remaining schemes did not improve the class separation significantly. Filtering does not seem to have too much impact in the obtained distances, however the anisotropic diffusion (AD) and the Speckle reducing anisotropic diffusion (SRAD) provided slightly better results than the no filtering case.

4.7 Chapter summary.

In this chapter we analyzed the statistics of gray levels in our images. It was found that the observations reported on cervical echogenicity changes regarding preterm birth (section 4.1), also hold for the labor induction case. We also tested several ways of image normalization and denoising. After evaluating these methods on our image data set, one is able to see that they do affect class separation (and the accuracy of classification) using LBP methods to some extent (see table 4.3). However this is not an exhaustive investigation using all the methods currently avail-

Table 4.3: Results obtained using the different filters presented so far, the metrics with the highest scores are shown along with the case of no normalization (None). For this data the ANT and POST regions were analyzed.

	Distance metric.			
Normalization	Bhat	Chisq	Euclid	Hintersec
	SRAD			
HN	**-3.35**	**3.39**	**4.76**	**22.64**
HE	**-3.67**	**1.04**	**3.67**	**35.17**
None	-3.02	5.37	4.95	14.22
	AD			
HN	**-15.03**	**257860.23**	**417236.55**	**2783020.85**
N2ref	**-14.78**	**456507.08**	**490724.66**	**1968917.08**
None	-14.32	353849.55	367419.80	1207274.15
	Linear			
HN	**-14.24**	**60714.55**	**192465.20**	**1331104.92**
N2ref	**-13.99**	**130087.42**	**248178.28**	**951440.93**
None	-13.54	108818.09	190765.50	582573.18
	No filtering			
HN	**-15.03**	**257860.23**	**417236.55**	**2783020.85**
N2ref	**-14.78**	**456507.08**	**490724.66**	**1968917.08**
None	-14.32	353849.55	367419.80	1207274.15

able, so it is possible that exist normalization and denoising algorithm allowing an improved classification performance.

References

[1] H Feltovich et al. "Effects of selective and nonselective PGE2 receptor agonists on cervical tensile strength and collagen organization and microstructure in the pregnant rat at term". In: *American Journal of Obstetrics and Gynecology* 192.3 (2005), pp. 753–760.

[2] H Jörn. "Prediction of premature birth using texture analysis of the cervix". In: *Ultraschall in der Medizin-European Journal of Ultrasound* 29.S2 (2005), P0_16.

[3] Tomoyuki Kuwata et al. "A novel method for evaluating uterine cervical consistency using vaginal ultrasound gray-level histogram". In: *Journal of Perinatal Medicine* 38.5 (2010), pp. 491–494.

[4] Jong See Lee. "Digital image enhancement and noise filtering by use of local statistics." In: *IEEE transactions on pattern analysis and machine intelligence* 2.2 (1980), pp. 165–168.

[5] Christos P. Loizou et al. "Quality evaluation of ultrasound imaging in the carotid artery based on normalization and speckle reduction filtering". In: *Medical and Biological Engineering and Computing* 44.5 (2006), pp. 414–426.

[6] J a Noble. "Ultrasound image segmentation and tissue characterization." In: *Proceedings of the Institution of Mechanical Engineers. Part H, Journal of engineering in medicine* 224 (2010), pp. 307–316.

[7] P Perona and J Malik. "Scale-space and edge detection using anisotropic diffusion". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12.7 (1990), pp. 629–639.

[8] Aleksandra Pizurica et al. "Despeckling Sar Images Using Wavelets and a New Class of Adaptive Shrinkage Estimators". In: *Image Processing, 2001. Proceedings. 2001 International Conference on*. Vol. 2. 2001, pp. 233–236.

[9] I Rosado-Mendez et al. "Analysis of coherent and diffuse scattering using a reference phantom". In: *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* PP.99 (2016), p. 1.

[10] Thomas L Szabo. *Diagnostic ultrasound imaging: inside out*. Academic Press, 2004.

[11] T.Elatrozy et al. "The effect of B-mode ultrasonic image standardisation on the ehodensity of symptomatic and asymptomatic carotid bifurcation plaques". In: *International Angiology* 17.3 (1998), p. 179.

[12] Yongjian Yu, Scott T Acton, and Senior Member. "Speckle Reducing Anisotropic Diffusion". In: 11.11 (2002), pp. 1260–1270.

Chapter 5

Deep Learning for US image classification

5.1 Introduction

Deep learning (DL) is a branch of machine learning based on a set of algorithms that attempt to model high level abstractions in data by using hierarchical structures. These algorithms try to imitate the functioning and structure of the mammal brain.

DL is nowadays an emergent technology which has received a lot of attention in the research community. Applications of Deep Learning based methods are numerous, from computer vision to natural language processing, automatic speech recognition and semantic learning.

In this chapter experiments using *Deep Convolutional Neural Networks* are carried out for the problem of image classification where these networks excel. It is usually a hard task to find the appropriate features for image classification; in previous chapters we have tested texture operators such as LPV or GLCM used in a multi-scale fashion and tried to find the best combination of features to better classify the images. In contrast these networks are good at "learning" automatically good features for discrimination. In the following sections we test some DL architectures on our image database and compared them with the results obtained so far. Though these networks do not perform texture based classification only, they process the images in a hierarchical way just as multi-resolution schemes do.

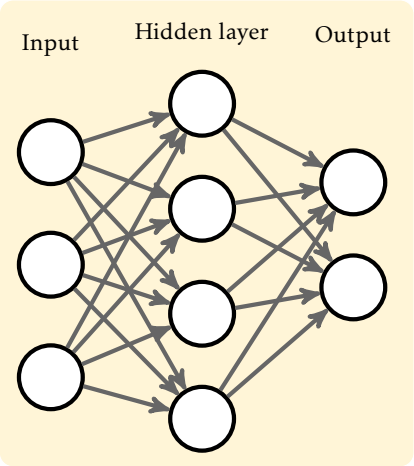


Figure 5.1: A multilayer perceptron with three layers. Each circle represents a neuron and the arrows represent connections between neurons.

5.2 Artificial Neural Networks.

Implementation of DL algorithms usually involves the use of artificial neural networks (ANN). One of the most employed ANNs is the multilayer perceptron.

The original linear perceptron is a type of ANN first conceived in the 1950s by Rosenblatt which used a linear activation function. MLPs are modifications of the standard linear perceptron and can distinguish data that are not linearly separable unlike the original linear perceptron.

In MLPs there exist three types of layer: the input layer where data is fed into, an output layer, and one or more hidden layers where intermediate calculations are performed (see figure 5.1) .

Activation functions: The firing of a neuron (transition of output values) is driven by an activation function. Activation functions try to mimic the firing of action potentials of neurons in biological systems. Common activation functions are the logistic, the hyperbolic tangent and the rectifier function as shown in figure 5.2.

Deep vs Shallow architectures. Not so long ago, MLPs with more than two hidden layers were rarely used. One of the reasons for this to happen was the difficulty of training such networks. A problem known as the *vanishing gradient* associated with the backpropagation method, caused the error in classification increase when adding more layers. Artificial neural networks with less than two hidden layers are termed *shallow* as opposed to multi-layered architectures which are now known as *deep* nets.

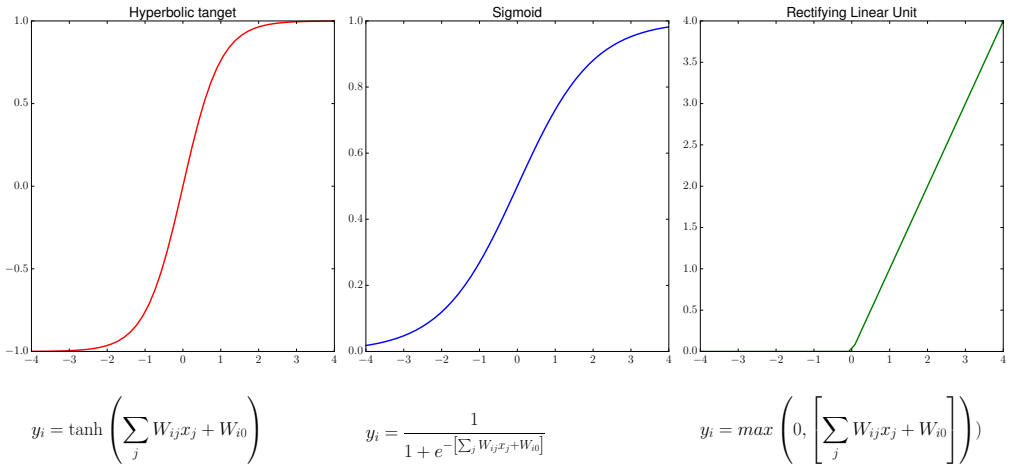


Figure 5.2: Some common activation functions. Here y_i is the output of the i_{th} node, W_i are the weights associated to the input synapses and W_{i0} is a bias term.

Motivations for choosing deep architectures [27]:

- Brains have a deep architecture.
- Humans organize their ideas hierarchically, through composition of simpler ideas.
- Insufficiently deep architectures can be exponentially inefficient.
- Distributed (possibly sparse) representations are necessary to achieve non-local generalization.
- Intermediate representations allow sharing statistical strength.

In deep nets, the several layers are aimed to learn levels of data representation and abstraction allowing the net to represent function of increasing complexity.

Before 2006 training deep architectures was unsuccessful except for convolutional neural nets. In 2006 some key research works like [11] on Deep Belief Networks demonstrated how multilayered feed forward neural networks could effectivley be pre-trained using a one-layer at a time strategy treating each layer as Restricted Boltzmann Machine (RBM). Other seminal paper were written by [4]. This paper explores and compares RBMs and auto-encoders (neural network that predicts its input, through a bottleneck internal layer of representation). In and

[16] the author uses sparse auto-encoder (which is similar to sparse coding) in the context of a convolutional architecture.

The following key principles are found in all three papers [4, 11, 16]:

- Unsupervised learning of representations is used to (pre-)train each layer.
- Unsupervised training of one layer at a time, on top of the previously trained ones. The representation learned at each level is the input for the next layer.
- Use supervised training to fine-tune all the layers (in addition to one or more additional layers that are dedicated to producing predictions)

5.3 Deep Learning Architectures

There are huge number of deep architectures variants. Most of them are branched from some original parent architectures. It is not always possible to compare the performance of multiple architectures all together, because they are not all evaluated on the same data sets. Deep learning is a fast-growing field, and new architectures, variants, or algorithms appear every few weeks.

1. Auto encoders
2. Restricted Boltzmann Machines
3. Convolutional Neural Networkks
4. Recurrent Neural Networks
5. Deep Sparse coding.

5.3.1 Autoencoders

Autoencoders are typically feedforward networks trained to copy their input to their outputs. They are commonly used to learn efficient encodings for data. The hidden layer of an autoencoder has less neurons than the input and output layers. In the hidden layer the network is forced to find the most important features of the input data achieving in this way a compact representation.

As unsupervised learning algorithms they are used to pre-train (from labeled or unlabeled data) features. These features can then be used as initialization for a supervised Multi-Layer Perceptron (MLP)

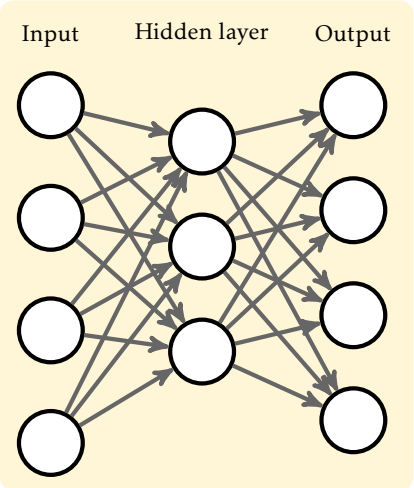


Figure 5.3: An autoencoder with three layers. The hidden layer has less neurons forcing the system to find a compressed representation of the input.

5.3.2 Restricted Boltzmann machines.

Restricted Boltzmann machines (RBM) are generative stochastic neural networks that can learn a probability distribution over their set of inputs. RBMs are composed of hidden and visible layers. The connections between the layers are undirected (i.e, the values can be propagated in both directions) and fully connected (each unit from a given layer is connected to eah other in the next)

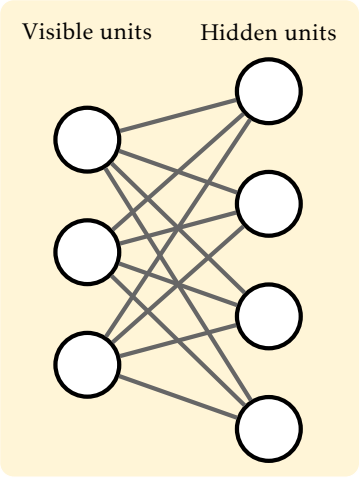


Figure 5.4: A restricted Boltzmann machine. The absence of arrow heads means that data calculation are performed in both directions.

5.3.3 Recurrent Neural Nets.

Recurrent neural networks are networks where neurons have feedback connections. They are appropriate for learning sequential tasks not learnable for traditional machine learning methods.

5.3.4 Deep Sparse coding.

Deep Sparse-Coding Networks (DepSCNets) are not ANN based nets, these networks are based on sparse coding techniques. The Convolution layer of ConvNets is reformulated to encode local patches using sparse coding. A DeepSCnet consists of four types of layers: Sparse-coding layers, pooling layers, normalization layers and map reduction layers [28].

5.3.5 Convolutional Neural Nets.

Convolutional Neural Nets (ConvNets) are special types of feed forward network and has become a popular choice for image recognition and other two-dimensional data. Building blocks for Convnets construction are convolutional layers, fully connected layers and pooling layers. These networks are easier to train than other regular deep feed forward nets and have fewer parameters to estimate.

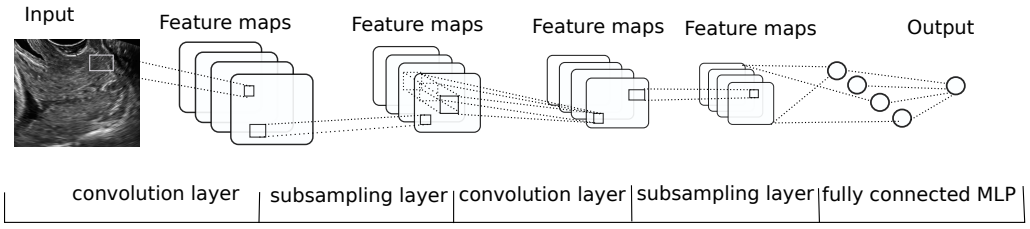


Figure 5.5: Typical Convolutional Neural Net architecture.

Convolution layers. The Convolution layer is the core building block of a Convolutional Network. They do the most heavy calculations in the net. In these layers a number of spatial filters are convolved with the image. They have usually small size or footprint, 3x3,7x7 or 11x11 sizes are common. The result of one filter applied across the image is called feature map (FM) and the number feature maps is equal to the number of filters. The spatial extent of the filter is a hyper-parameter called the *receptive field*. By using convolutional layers the amount of parameters to be learned or tuned is reduced.

Pooling (subsampling) layers. These layers reduce the size of the input. For example, if the input consists of a 32x32 image and the layer has a subsampling region of 2x2, the output value would be a 16x16 image, which means that 4 pixels (each 2x2 square) of the input image are combined into a single output pixel. There are multiple ways to subsample, but the most popular are max pooling, average pooling, and stochastic pooling.

Fully connected layers. The last pooling (or convolutional) layer is usually connected to one or more *fully connected layers*, the last of which represents the target data.

Training is performed using modified back-propagation that takes the pooling layers into account and updates the convolutional filter weights based on all values to which that filter is applied.

All of these architectures posses their own area of application where they are more appropriate, for instance to study systems changing over time a recurrent neural net would be the most useful. For the image processing case convolutional neural nets seem to be the right choice. These nets have been proven to be excellent for image classification and object recognition tasks.

5.4 Medical Applications

Currently there is a number of applications of DL on medicine. Most of them are related to computer vision and interpretation tasks. Medical diagnosis based on radiology images are among the most common. Some examples are Magnetic Resonance imaging (MRI) of brain white matter [29], cardiac MRI [1], breast micro calcification detection in mammography images [3], [14], invasive ductal carcinoma detection in whole side histopathology images [9], or pulmonary nodule detection [24].

There is also applications on organ segmentation, for instance [29] on fetal brain segmentation on computer tomography (CT) images, [18] automatically retrieve missing or noisy cardiac acquisition plane information from magnetic resonance imaging (MRI) and predict the five most common cardiac views. An application using multiscale deep networks for direct estimation of cardiac ventricular volumes is presented in [30].

Big companies like Microsoft are developing automated tools like Inner Eye for medical image diagnosis based on deep learning.

Applications on ultrasound imaging. Although most of the applications of DL on medical imaging are devoted to Computer Tomography or Magnetic Resonance imaging, there exist however several application on ultrasound images or videos.

A new method for automatic detection and classification of suspected breast cancer lesions using ultrasound images is proposed in [15] where authors use a combination of convolutional neural network and Fuzzy support vector machines. Automatic blood vessel detection is performed in [21] using a convolutional neural net.

Classification of liver disease based on contrast - enhanced ultrasound (CEUS) videod is proposed in [26]. The authors use a Deep Belief Network for classification and sparse non-negative matrix factorization as preprocessing step. Fetal Ultrasound plane detection is automatized by using a Recurrent Neural Net [6]. A more or less comprehensive survey of medical applications of DL an be found in [12].

5.5 Mitigating two common problems in Deep Learning models.

The appeal for deep learning based method is currently driven for the following reasons [17]:

- DL networks learn both features and classifier in the same step. The need for hand-crafted features is avoided.
- Outperforms other systems in multiple domains. This inludes speech, language, vision and gaming by a considerable amount.
- Architectures that can be adapted to new problems relatively easy.

However this technology usually requires large amounts of data for training a network, if the aim is to outperform traditional approaches. Nets can be extremely computationally expensive to train. Depending on the size of the model can take weeks for training.

Also, since there may be thousands or even millions of parameters to tune, we can easily be affected by overfitting. To fullfil these two needs, dimensionality reduction and overfitting prevention, several methods have been developed.

5.5.1 Reducing overfitting.

Overfitting occurs when a model is tighly attached to a particular dataset, i.e the classification error is very small in that dataset but is much bigger in any other. Overfitting reduces the generalzition of a classification system and is therefore an undesirable phenomenon.

Data augmentation. The easiest and most common method to reduce overfitting on image data is to artificially enlarge the dataset using label-preserving transformations. Some of the most used are:

1. Image translations.
2. Horizontal reflections.
3. Changing RGB intensities.
4. Adding noise.
5. Image rotations.

Drop out. Drop out consist of randomly setting to zero the output of each neuron with a predefined probability (usually 0.5). By doing in this way we prevent these neurons to contribute to the forward pass and cancel their influence in back propagation. So every time an input is presented, the neural network samples a different architecture.

L1/L2 Regularization. Regularization works by penalizing large neuron weights which helps in generalization. The objective function to be optimized is changed to $E(x)+\lambda L_p(W)$ where L_p is a regularization function and W are the neuron weights.

1. L_1 regularization: $\sum_{i=1}^{m}|w_i|$ also known as Lasso.
2. L_2 regularization: $\sum_{i=1}^{m}|w_i^2|$ also known as ridge, or weight decay. This is the most widely used.
3. L_{12} regularization: $\lambda_1 L_1 + \lambda_2 L_2$ also known as elastic net regularization.

Early stopping. This method tries to find the parameters that give the best validation error. In early-stopping the training is stopped before overfitting.

Batch normalization. During the training process the distribution of input of each layer changes as the parameters of previous layers change. This increase the training time and makes it difficult to train models with saturating non-linearities (as softmax layers) . In batch normalization inputs layers are normalized. The normalization is set as part of the model architecture and it is performed for each batch. Batch Normalization allows to use much higher learning rates and be less careful about initialization [13].

5.5.2 Dealing with dimensionality.

To reduce the amount of parameters to be learned by the network, clever techniques has been designed. As mentioned in section 3.4 a high dimensional feature space is generated when working with images using multi-resolution techniques.In image processing the dimensionality of the feature space has been always an issue.

Pooling A pooling layer is always placed after a convolution layer to perform a sub-sampling action on the feature maps. A spatial window of a predefined size without overlapping is applied to the feature map and a pooling function (e.g. maximum selection) is calculated inside each window (max-pooling). Other types of pooling like mean or median pooling are also used.

5.6 Experiments with Convolutional Neural Networks on Transvaginal Ultrasound Images.

5.6.1 A small ConvNet model architecture.

The code used for testing ConvNets was written in Python using the keras library [7]. This library allows for easy development and testing of deep learning nets.

The first convolutional neural net used in our experiments, is based upon a model described in the Keras blog [1] and it resembles the models proposed by Yann LeCUn in the 1990s. The model consist of several layers:

[1] https://blog.keras.io

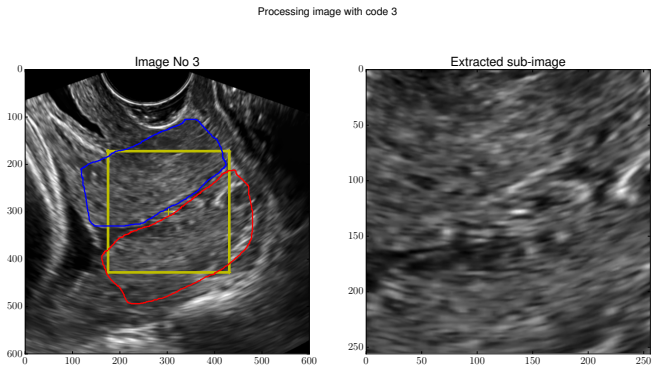


Figure 5.6: Extracting a squared region (in yellow) from an image. ANT and POST region are shown in blue and red respectively. To the right, the extracted image.

1. Convolutional layer with 32 filters of 3 by 3 pixels window. The activation used in this layer was the rectifying linear unit (ReLu).
2. Max Pooling layer with a 2 by 2 pixels window.
3. Convolutional layer with 32 filters of 3 by 3 pixels window with 'ReLu' activation.
4. Max Pooling layer with a 2 by 2 pixels window.
5. Convolutional layer with 64 filters of 3 by 3 pixels window with 'ReLu' activation.
6. Max Pooling layer with a 2 by 2 pixels window.
7. Dense layer with 64 hidden nodes. Activation 'Relu'. Used Drop Out with 0.5 for regularization.
8. Output dense layer with two unit (binary output) with sigmoid activation.

Preprocessing. To limit the size of the image to be processed, a square portion of 256×256 pixels was cropped from every image in the data set. The extracted region was chosen to be centered on the section containing the ROIs ANT and POST, as shown in figure 5.6. Since the images were bitmaps our input images were 256x256x3.

A re-scaling of the gray values (0-255) was also performed. The intensities of the input images were set to be ranging from 0 to 1.

Data augmentation. Since our image database is very small, artificial augmentation of the available amount of images for training and testing our net was necessary. Several changes were made to the images in order to create new examples:

- Rotating the images by ± 20°
- Shifting horizontally the images by 20% of total width.
- Shifting vertically the images by 20% of total height.
- Flippingg the images horizontally.
- Flipping the images vertically

Model fitting. The network model was trained using the augmented data during 20 epochs. An epoch in the neural network terminology means one forward pass and one backward pass of all the training examples in a dataset and the batch size is eqyal to the number of training examples in one forward/backward pass. In each epoch 2,000 samples, in batches of 16 samples are generated from the training set. These images were used as inputs for the model during training. The model is then tested in a set of 800 samples in batches of 16 samples generated now from the validation set.

During the fitting process, the loss function was set to *categorical cross-entropy* and the model was fitted using the *stochastic gradient descent* (SGD) method with a learning rate of 0.01 and a momentum of 0.9. These parameters control the speed of weight updates in every iteration. The total of model parameters that were tuned was 3,714,593.

Results. The accuracy reached with this simple model was 93.5% on this dataset which is better than previously achieved. In figure 5.7 the learning curves during training and validation are shown and the feature maps corresponding to the first convolution layer can be observed in figure 5.8

5.6.2 Transfer learning: Using pre-trained models.

When working on Deep Learning one come across very often with the models used in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). The purpose of this annual contest is to develop models that can correctly classify an input image into one of the 1,000 existing categories in the database.

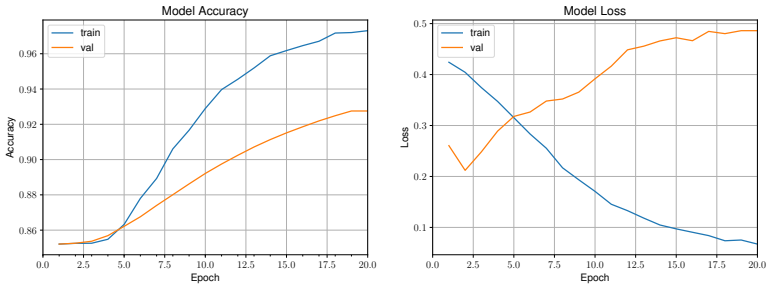


Figure 5.7: Learning curves for the model. To the right training and validation learning curves, to the right the loss function for both cases.

To train models 1.2 million images are available for training, another 50,000 for validation and 150,000 images for testing. The kind of objects in the database is diverse: cats, dogs, cars, vehicle types and many more.

The ILSVRC has become a vey important reference for computer vision classification algorithms,and since 2012 it has been dominated by convolutional neural nets and Deep Learning techniques.

Currently a medical image data base with annotations comparable in size to *ImageNet* is not available. However, several research works have shown that by applying some modifications to a previously trained model and fine-tuning it, a good classification can be achieved [2, 5, 19]. This is true even though the images used for training in the model were rather different from the type of images to be classified. This fact can be used as another way to overcome data scarcity; the weights learned by a network trained in a much larger data set such as ImageNet can be used in a new image classification problem.

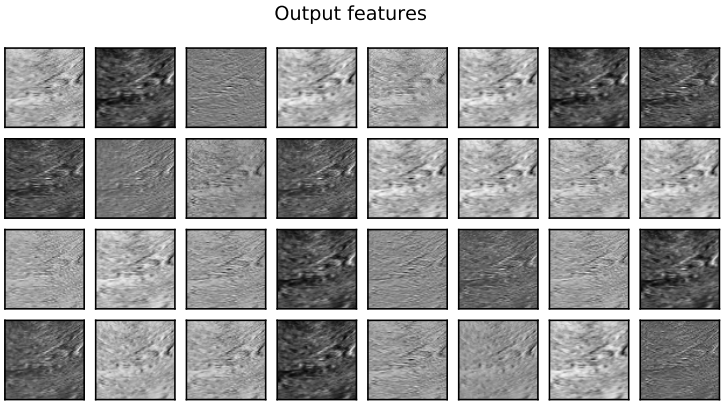


Figure 5.8: 32 feature maps corresponding to the first layer

Here we present some of the highest performing Convolutional Neural Networks on the ImageNet challenge over the past few years. These networks also demonstrate a strong ability to generalize to images outside the ImageNet dataset via transfer learning, such as feature extraction and fine-tuning.

1. **AlexNet.** AlexNet is a convnet designed by Alex Krizhevsky that participated in the ImageNet Large Scale Visual Recognition Challenge in 2012. It has only 8 layers, first 5 layers were of convolutional type and the remaining 3 were fully connected layers. On the test data, the model achieved top-5 error rate of 17.0% which was considerably better than the previous state-of-the-art at that time.

2. **VGG16/19**. VGG16 and VGG19 are 16/19-layers Convnets [20] used by the Visual Geometry Group (VGG) at Oxford University in the 2014 ILSVRC (ImageNet) competition. The VGG16/19 models achieved a 7.5% top 5 error rate on the validation set and 7.4/ 7.3 % on the test set of ILSVRC-2012.

3. **ResNet50.** Residual networks [10] are easier to optimize, and can gain accuracy from considerably increased depth. These networks were build up to 152 layers in depth (ResNet50 has only 50 layers) but still having lower complexity i.e. fewer parameters to be tuned, than previous models like VGG nets. An ensemble of these residual nets achieves 3.57% error on the ImageNet test set. This result won the first place on the ILSVRC 2015 classification task.

4. **Inception V3.** Inception-V3 achieved the second place in the 2015 ImageNet competition with a 5.6% top 5 error rate on the validation set. The model is characterized by the usage of the Inception Module [22], which is a concatenation of features maps generated by kernels of varying dimensions.

5. **Xception.** Xception is an extension of the Inception architecture which replaces the standard Inception modules with depthwise separable convolutions [8].

5.6.3 Transfer Learning with Inception V3 and ResNet50 on TVU images

.

Transfer learning is considered as the transfer of knowledge from one learned

task to a new task in machine learning [25]. Regarding Neural Networks this means the transferring of learned features from a trained network to be used in a new problem. Transfer learning usually results in faster training times than training a new convolutional neural network because you do not need to estimate all the parameters in the new network.

The transferring can be performed in the following ways:

1. Feature extraction We can use a pre-trained model as a feature extraction mechanism. What we can do is that we can remove the output layer(the one which gives the probabilities for being in each of the 1,000 classes) and then use the entire network as a fixed feature extractor for the new data set.

2. Use the Architecture of the pre-trained model. What we can do is that we use architecture of the model while we initialize all the weights randomly and train the model according to our dataset again.

3. Train some layers while freeze others. Another way to use a pre-trained model is to train is partially. What we can do is we keep the weights of initial layers of the model frozen while we retrain only the higher layers. We can try and test as to how many layers to be frozen and how many to be trained.

The choice of using one of the mentioned approaches is dictated by the type of our data, the amount of training examples and even computing resources at hand.

5.6.3.1 Fine Tuning AlexNet.

The ALexNet weights were loaded from [2]. For the fine tuning, we used the strategy delineated in [23] which essentially consist of performing the tuning in a layer-wise manner. We used the same data set as before, but the images were 227x227x3 pixels since this is the image size used when training this model. The weights for all layers are loaded except for the last one, this is our "base model". The fully connected layer (usually termed the "bottleneck") is going to be substituted by our own layer (the original one is for 1,000 classes and we use only two).

Once the new model is created, it is compiled. A very small learning rate is used (usually about 10^{-4}) using stochastic gradient descent (SGD) as optimizer. We used a learning rate of 0.001 and a momentum of 0.9.

The training process is started with the base model layer "frozen" which means

[2] http://files.heuritech.com/weights/alexnet_weights.h5

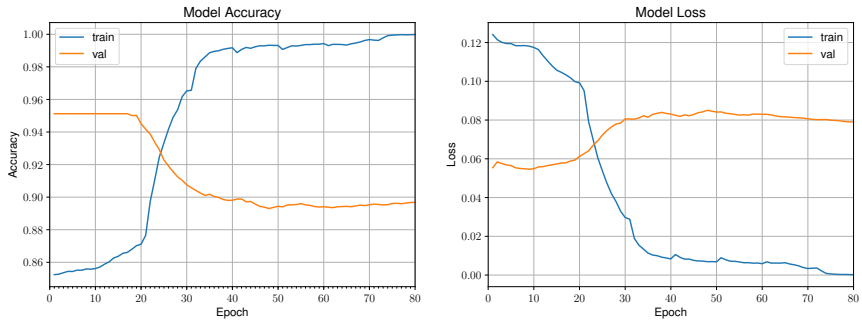


Figure 5.9: Learning curves (left) and loss function curves (right) for the AlexNet mode for bth training and validation.

that their weights are not updated and only the last added layer weights. Additional layers from the base model can be unfrozen (set as trainable) and the cycle is again started until an acceptable accuracy is achieved.

5.6.3.2 Inception V3

The goal of the inception module is to act as a multi-level feature extractor by computing 1x1, 3x3, and 5x5 convolutions within the same module of the network the output of these filters are then stacked along the channel dimension and before being fed into the next layer in the network (figure 5.10).

The original incarnation of this architecture was called GoogLeNet, but subsequent manifestations have simply been called Inception vN where N refers to

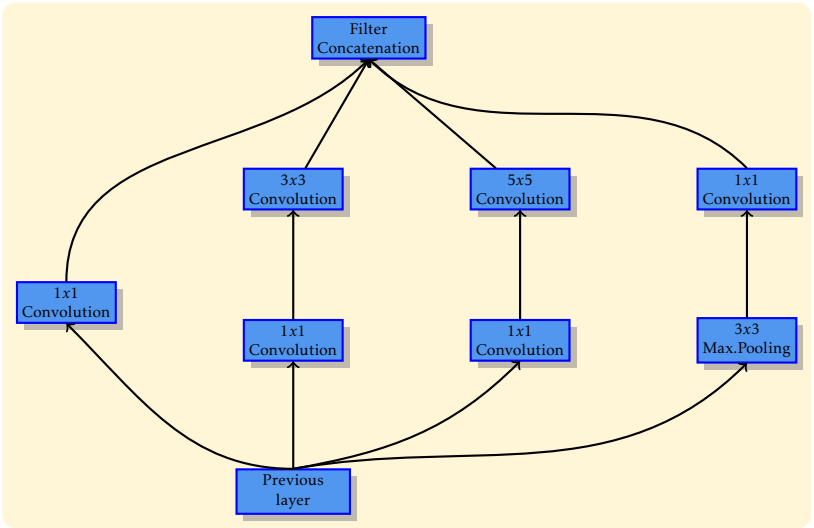


Figure 5.10: The original Inception module with dimensionality reduction used in GoogLeNet.

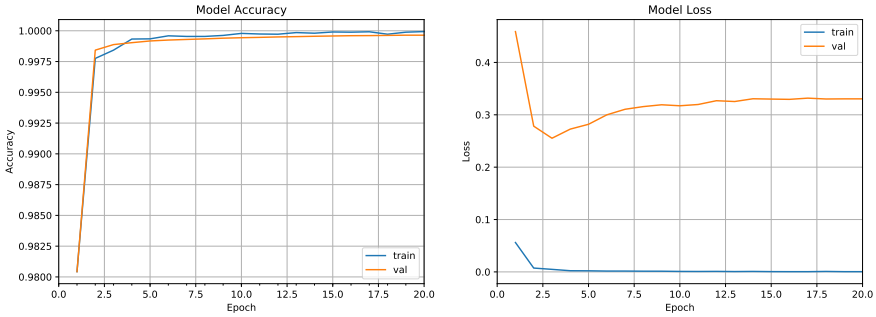


Figure 5.11: Learning curve for the Inception v3 model. Right: Accuracy for training and validation, Left: The loss function for both cases.

the version number put out by Google. For this model the input images should be $224x224$ pixels.The images are re-scaled to have values in the range (0,1) before being used as inputs to the convnet.

Model fine-tuning The model weights for the 27 layers in this convnet are publicly available at Github. [3]. The structure of the network is available in the keras library. The size of the images to be used for this model were 224x224x3 pixels. As pre-processing we normalize each image to have values in the range (0,1), but did not subtract the mean as in the original method. We proceeded as before replacing the last soft max layer for one with two classes. The model was compiled using SGD with a learning rate of 0.001 and a momentum of 0.9.

5.6.3.3 ResNet50.

Unlike traditional sequential network architectures such as AlexNet and VGG, ResNet is instead a form of exotic architecture that relies on micro-architecture modules (also called network-in-network architectures).

The term micro-architecture refers to the set of building blocks used to construct the network. A collection of micro-architecture building blocks (along with your standard Conv, Pool, etc. layers) leads to the macro-architecture (i.e, the end network itself).

First introduced in [10], the ResNet architecture has become a seminal work, demonstrating that extremely deep networks can be trained using standard SGD (and a reasonable initialization function) through the use of residual modules:

[3] https://github.com/fchollet/deep-learning-models/releases/download/v0.5/inception_v3_weights_tf_dim_ordering_tf_kernels.h5

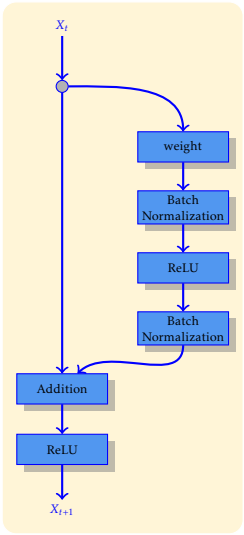


Figure 5.12: The residual module in ResNet as originally proposed by He et al. in 2015.

Model fine-tuning. The model weights are available from [4]. The process is similar to InceptionV3 tuning. Image for this model are 224x224x3 pixels. The preprocessing is also similar to the one applied before.

Again we truncate and replace the softmax layer for transfer learning. The fine-tuning process will take a while, depending on the available hardware resources. After it is done, we use the model to make predictions on the validation set and return the score. The accuracy obtained by the three networks is show in table 5.1.

5.7 Chapter summary.

The simple metric of accuracy was used as measure for easy comparison of the performance of the different models used in this chapter. Due to the lengthy simulations needed for the model to be trained, it was impractical to carry out

[4] https://github.com/fchollet/deep-learning-models/releases/download/v0.2/resnet50_weights_th_dim_ordering_th_kernels.h5

Table 5.1: Classification results obtained after fine-tuning networks. Accuracy is calculated as defined in section 1.5

Model	Accuracy
Model trained from scratch	
Custom net	0.93
Pre-trained models	
AlexNet	0.90
InceptionV3	0.99
ResNet50	0.96

cross validation which implies several runs of the training and testing processes. The accuracy values obtained in the experiments with convolutional neural nets has shown the potential of these networks for US image classification. Training the model from scratch although we obtained good accuracy is less effective than taking advantage of a pre-trained model. By using the weights of a pre-trained model we decrease the time used for the net to learn. This is so, because we keep the weights of the bottom layers(layers closer to the input layer) which contains very general information about the images to be analyzed and concentrate only on training the top layers (closer to the output) where more specific features are to be learned. For the transvaginal US images we tested, the best classification results in terms of accuracy were obtained using the inceptionv3 model followed by the ResNet50. These two nets outperform AlexNet presumably because they require less parameters to train apart of being deeper which enable them to describe more complex relationship.

References

[1] M. R. Avendi, Arash Kheradvar, and Hamid Jafarkhani. "A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI". In: *Medical Image Analysis* 30 (2016), pp. 108–119. arXiv: 1512.07951.

[2] Avi Ben-cohen B et al. "Fully Convolutional Network for Liver Segmentation and Lesions Detection". In: *International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis* 10008 (2016), pp. 77–85. arXiv: arXiv:1608.04117v1.

[3] Alan Joseph Bekker, Hayit Greenspan, and Jacob Goldberger. "A multi-view deep learning architecture for classification of breast microcalcifications". In: *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)* (2016), pp. 726–730.

[4] Yoshua Bengio et al. "Greedy Layer-Wise Training of Deep Networks". In: *Advances in Neural Information Processing Systems* 19.1 (2007), p. 153. arXiv: 0500581 [submit].

[5] Gustavo Carneiro, Jacinto Nascimento, and Andrew P. Bradley. "Unregistered multiview mammogram analysis with pre-trained deep learning models". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9351 (2015), pp. 652–660.

[6] Hao Chen et al. "Automatic fetal ultrasound standard plane detection using knowledge transferred recurrent neural networks". In: *International Confer-*

ence on Medical Image Computing and Computer-Assisted Intervention. Springer. 2015, pp. 507–514.

[7] François Chollet. *Keras*. \url{https://github.com/fchollet/keras}. 2015.

[8] François Chollet. "Xception: Deep Learning with Depthwise Separable Convolutions". In: (2016). arXiv: 1610.02357.

[9] Angel Cruz-Roa et al. "Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks". In: *Proc. SPIE* 9041.216 (2014), pp. 904103–904115.

[10] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *Arxiv.Org* 7.3 (2015), pp. 171–180. arXiv: 1512.03385.

[11] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. "A Fast Learning Algorithm for Deep Belief Nets". In: *Neural Computation* 18.7 (2006), pp. 1527–1554. arXiv: 1111.6189v1.

[12] David Hutchison. *Deep Learning and Data Labeling for Medical Applications*. Springer, 2016.

[13] Sergey Ioffe and Christian Szegedy. "Batch Normalization : Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *CoRR* abs/1502.0 (2015). arXiv: arXiv:1502.03167v3.

[14] AR Jamieson and K Drukker. "Breast image feature learning with adaptive deconvolutional networks". In: *SPIE Medical Imaging* (2012), pp. 831506–831506.

[15] B Karimi and A Krzyak. "A novel approach for automatic detection and classification of suspicious lesions in breast ultrasound images". In: *Journal of Artificial Intelligence and Soft Computing Research* 3.3 (2013), pp. 265–276.

[16] Yann LeCun et al. "Efficient Learning of Sparse Representations with an Energy-Based Model". In: *Nips* 1 (2006), pp. 1137–1144. arXiv: 1112.6209.

[17] Hui Ming Li. "Deep Learning for Image Recognition". In: *International Journal of Signal Processing, Image Processing and Pattern Recognition* 7.3 (2013), pp. 171–180. arXiv: 1512.03385.

[18] J Margeta, A Criminisi, and R Cabrera Lozoya. "Fine-tuned convolutional neural nets for cardiac MRI acquisition plane recognition". In: *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* (2015).

[19] HC Shin et al. "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning". In: *IEE Transactions on Medical imaging* 35.5 (2016), pp. 1285–1298.

[20] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *International Conference on Learning Representations (ICRL)* (2015), pp. 1–14. arXiv: 1409.1556.

[21] Erik Smistad and Lasse Løvstakken. "Vessel Detection in Ultrasound Images Using Deep Convolutional Neural Networks". In: *International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. Springer. 2016, pp. 30–38.

[22] Christian Szegedy et al. "Going Deeper with Convolutions". In: (2014). arXiv: 1409.4842.

[23] Nima Tajbakhsh et al. "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?" In: *IEEE Transactions on Medical Imaging* 35.5 (2016), pp. 1299–1312. arXiv: 1706.00712.

[24] Bram Van Ginneken et al. "Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans". In: *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)* (2015), pp. 286–289.

[25] Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. "A survey of transfer learning". In: *Journal of Big Data* 3.1 (2016), p. 9.

[26] Kaizhi Wu, Xi Chen, and Mingyue Ding. "Deep learning based classification of focal liver lesions with contrast-enhanced ultrasound". In: *Optik - International Journal for Light and Electron Optics* 125.15 (2014), pp. 4057–4063.

[27] Masood Zamani and Stefan C Kremer. *Handbook on Neural Information Processing*. Vol. 49. 2013, pp. 505–525.

[28] Shizhou Zhang et al. "Constructing Deep Sparse Coding Network for image classification". In: *Pattern Recognition* 64 (2017), pp. 130–140.

[29] Wenlu Zhang et al. "Deep convolutional neural networks for multi-modality isointense infant brain image segmentation". In: *NeuroImage* 108 (2015), pp. 214–224.

[30] Xiantong Zhen et al. "Multi-scale deep networks and regression forests for direct bi-ventricular volume estimation". In: *Medical Image Analysis* 30 (2016), pp. 120–129.

Chapter 6

Conclusions

In this chapter we summarize the findings made along this thesis work. Aspects that were not sufficiently addressed are commented in the limitations section and finally some possible directions that this research can have as continuation and extension are pointed out is the future work section.

After analyzing our images in the various ways outlined in previous chapters, we can summarize the main findings in this thesis:

1. The TVU images of the cervix show gray level variations as reported in the reviewed literature (section 4.1) related mainly to preterm birth.

2. The texture of the images do not have a strong orientation. In spite of the fact that , as mentioned in chapter one, it has been discovered that the cervix collagen behaves anisotropically especially in the non-pregnant state.

3. The spatial frequency seems to have a higher discriminative potential than scale alone. So , analyzing the images at different frequency scales proved to be more successful. Including additional information as spatial distribution of local binary patterns or gray level variance can further increase the obtained accuracy.

4. Image gray level normalization has more influence on the classification accuracy than denoising, at least for the texture attributes used on our experiments.

6.1 Addressing research question

Now, we return to some questions that needed to be addressed from chapter one :

Table 6.1: Successful induction predictive values reported in several works.

Method	Condition	Sensitivity	Specificity	PPV	NPV
Bishop score [1]	>5	0.66	0.49		
Cervical length	>26mm	0.62	0.61	–	
Bishop score [3]	>3	0.58	0.77	–	
Cervical length	>28mm	0.87	0.71	—	
Bishop score [4]	>4	0.87	0.45	–	
Cervical length[2]	< 20mm	0.64	0.70	0.57	0.76
Cervical Area	$< 100mm^2$	0.64	0.70	0.57	0.76
Bishop score	> 5	0.62	0.57	0.46	0.71
Mean elastographic index	< 100	0.76	0.56	0.51	0.79
Cervical hard area	$< 200mm^2$	0.86	0.60	0.51	0.87

1. Is it feasible to predict the outcome from labor induction procedures by means of texture analysis?.

2. If the answer to the first item is yes, how good is our predicting capability?.

3. Is the proposed algorithm a better alternative to digital examination using the Bishop score?

To answer the first question we can say that the algorithms tried so far have provided promising results that suggest that this is really the case. The accuracy scores obtained until now are between 77 and 99 percent of correct classification. It may be argued that since the distribution of the classes are not symmetric (we have an asymmetric distribution for cesarean section) accuracy is not a good parameter for reasonable comparisons.

To elucidate that problem we resort to the ROC curves obtained in the experiments previously performed. From the ROC curve sensitivity and specificity values which measure the system capability of correctly detecting positive and negative classes, can be easily visualized. To compare with Bishop score we included the sensitivity and specificity values for the Bishop score already presented in chapter one, and we reproduce here for commodity in table 6.1.

From the figure is apparent that Bishop score prediction power is smaller than those obtained by texture analysis.

6.2 Limitations:

Despite the good results obtained in this thesis work, we have to point out some aspect that deserve to be mentioned:

- Our data base was not big enough to provide statistically solid results. This is specially true for the Deep Learning case. What it is even worse, for the class to be detected (i.e. cesarean section) we have rather small amount of samples. This essentially a problem of the current probability of a cesarean section to occur during a labor induction which is close to 20%. We have been collecting these images for a period of about two years a relatively short time to acquire a significant number of samples.

- Images have been, during experiments, selected to have good visual quality and excluded too-bad formed images. This probably will not be possible in a practical application, so a more robust preprocessing stage have to be designed. We did not care too much about preprocessing apart from some simple methods.

- The design of a quality control scheme for the US images would be advisable since, for the time being, the selection of good images is a manual task.

- A related issue is that the acquisition of the images have been performed following carefully a clinical protocol and even so, some artifacts were present in the images, e.g. specular reflections, shadowing and some others. It is known that specular reflection are dependent on insonation angle, i.e. there is a degree of dependency on the operator.

- We experienced limitations in the test of deep learning algorithms imposed by hardware requirements. The GPU used in the experiments was a modest category one and we have little memory at our disposal which caused run out of memory errors.

6.3 Future work

As a future work we may propose among other things to carry out experiments with a larger set of images. The inclusion of more than one texture operator may improve what have been achieved so far. Acquisition of images of the cervix using second harmonic, would probably provide a better quality images which together with elastography would serve as a comparison or control information.

Finally, based on the results attained with the experiments using Convolutional Neural Nets, we believe that classification performance obtained being good, could be further improved. If more data is available it would be worthy to continue

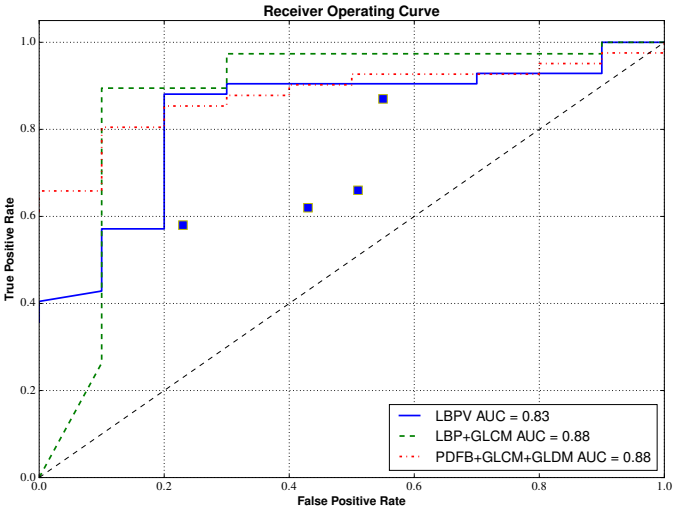


Figure 6.1: ROC curves for the algorithm tested in this thesis and some reported values of specificity and sensitivity for the Bishop score. Bishop score values are marked with squares.

trying the models used in this thesis and also to develop new ones.