



Predicting the outcome of intracytoplasmic sperm injection in women using machine learning combining Doppler and clinical data : SPIRL

Zeinab Abbas

► To cite this version:

Zeinab Abbas. Predicting the outcome of intracytoplasmic sperm injection in women using machine learning combining Doppler and clinical data : SPIRL. Acoustics [physics.class-ph]. Le Mans Université, 2023. English. NNT : 2023LEMA1029 . tel-04412683

HAL Id: tel-04412683

<https://theses.hal.science/tel-04412683>

Submitted on 23 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



THESE DE DOCTORAT

DE
LE MANS UNIVERSITE

SOUZ LE SCEAU DE
LA COMUE ANGERS – LE MANS

ECOLE DOCTORALE N° 602
Sciences de l'Ingénierie et des Systèmes
Spécialité : «Acoustique »

Par
« Zeinab ABBAS»

« Predicting the outcome of intracytoplasmic sperm injection in women using machine learning combining Doppler and clinical data»

« SPIRL »

Thèse présentée et soutenue virtuelle et à « ESEO, Angers », le « Decembre 18, 2023 ».
Numéro National de Thèse (NNT) : 2023LEMA1029

Rapporteurs avant soutenance :

Régine Le Bouquin, Pr. Université Rennes
Amine Nait Ali, Pr. Université Paris Est-Créteil

Composition du Jury :

Attention, en cas d'absence d'un des membres du Jury le jour de la soutenance, la composition du jury doit être revue pour s'assurer qu'elle est conforme et devra être répercutee sur la couverture de thèse

Président : Prénom Nom
Examinateurs : Denis KOUAME Pr, Université Paul Sabatier, Toulouse
Sofiane BOUDAOUD Pr, Université de technologie de Compiègne.
Amine NAIT ALI Pr, Université Paris Est-Créteil
Régine LE BOUQUIN Pr, Université Rennes

Dir. de thèse : Jean-Marc GIRAUT Pr, ESEO, Le Mans Université
Co-dir. de thèse : Jamal CHARARA Pr, Université Libanaise de Beyrouth
Co-encadrant : Sébastien MENIGOT MCF ESEO, Le Mans Université

Invité(s)

Chadi FAKIH ALHADI CENTER, Beirut, Lebanon

ACKNOWLEDGMENTS

I'd like to express my heartfelt gratitude to the people and institutions that follow for their important aid and support in completing this research project

Above all, I would like to convey my sincerest gratitude to my supervisor, Dr. GIRAUT Jean-Marc, for his unwavering guidance, support, and motivation throughout the entire research endeavor. His valuable perspectives and feedback have been pivotal in shaping the trajectory and extent of this study.

I'd wish to appreciate my supervisor, Dr. CHARARA Jamal who provided guidance, support, and encouragement during the entire research process.

Additionally, I extend my gratitude to the members of my research committee, Dr. MENIGOT Sébastien for their valuable input and constructive critiques. Their feedback has been instrumental in refining and enhancing the quality of this research.

Furthermore, I am grateful to Dr. IBRAHIM Zein for sharing their pearls of wisdom throughout this research.

I would like to extend our gratitude to Dr. AUBIN Sébastien for their valuable feedback on a previous iteration of the work.

I'd also want to thank Dr. FAKIH Chadi for providing assistance with data collection.

I would like to express my gratitude to Aurélie APIOU, Paul CRITON, Chloé DUFER, Shad KIJEK, and Nathan LE BOISSELIER for their help in the segmentation part.

I would like to sincerely thank and express my deep appreciation to my beloved mother, father, sisters and brother. Their boundless love, and encouragement, the guidance, sacrifices, and unwavering belief in my abilities that they have shown have profoundly shaped my personal and academic growth. I am eternally grateful for their presence in my life and for the countless opportunities they have provided me. Words cannot adequately express my gratitude for everything they have done for me.

I would like to express my heartfelt appreciation and acknowledgment to my beloved husband and my girls "MELINE" and "ELEINE" for their unwavering support, love, and understanding throughout my journey of pursuing a PhD. Their constant encouragement, sacrifices, and belief in my abilities have been instrumental in my success. I am truly grateful for their presence in my life and for being my pillars of strength.

Finally, I want to convey my appreciation to Le Mans University and ESEO for their support in providing the necessary resources and facilities for conducting this research.

Thank you for all the time you have spent in implementing this project.

TABLE OF CONTENTS

THESE DE DOCTORAT	i
DE LE MANS UNIVERSITE.....	i
ECOLE DOCTORALE N° 602 <i>Sciences de l'Ingénierie et des Systèmes</i> Spécialité : «Acoustique »	i
ACKNOWLEDGMENTS	i
TABLE OF CONTENTS	ii
LIST OF FIGURES	iv
LIST OF TABLES	vi
LIST OF ABBREVIATIONS.....	vii
CONTEXT	1
CHAPTER 1 GENERAL INTRODUCTION	3
1.1 BACKGROUND	3
1.2 PROBLEM STATEMENT	5
1.3 STATE OF THE ART OF PREDICTION SYSTEMS FOR THE SUCCESS RATE OF IVF/ICSI	7
1.4 OBJECTIVE OF THESIS.....	8
1.5 HYPOTHESIS OF THESIS.....	9
1.6 THESIS OUTLINE	10
CHAPTER 2 CLASSIFICATION BASED ON CLINICAL DATA (SPIRL VERSION I)	11
2.1 INTRODUCTION	11
2.2 MATERIALS AND METHODS OF SPIRL VERSION I	11
2.2.1 Flowchart	11
2.2.2 Data Collection.....	12
2.2.3 Inclusion and Exclusion Criteria	13
2.2.4 Data preprocessing	15
2.2.5 Feature selection techniques:	17
2.2.6 Why ANOVA from filter method in SPIRL V1?.....	18
2.2.7 Model selection.....	18
2.2.8 Hyper-parameter tuning techniques:	21
2.3 RESULTS	22
2.4 DISCUSSION.....	22
CHAPTER 3 CLASSIFICATION BASED ON CLINICAL AND POWER DOPPLER DATA (SPIRL VERSION II).....	24
3.1 INTRODUCTION	24
3.2 MATERIALS AND METHODS	24
3.2.1 Flowchart	24
3.2.2 Data Collection.....	27
3.2.3 Data Pre-Processing	28
3.2.4 Feature Selection Technique	28
3.2.5 Feature Extraction Techniques	30
3.2.6 Model Selection	33
3.2.7 Hyper-parameter tuning techniques	35
3.3 RESULTS	35
3.3.1 Description of the innovated results for the final classification for 94 patients:	35
3.3.2 Description of the innovated results for the final classification for 572 patients:	37
3.4 DISCUSSION.....	43
CHAPTER 4 CLASSIFICATION BASED ON CLINICAL DATA AND DOPPLER WITH AUTOMATIC AND SEMI-AUTOMATIC ULTRASOUND FEATURE EXTRACTION (SPIRL VERSION III)	47
4.1 INTRODUCTION	47
4.2 STATE OF THE ART	49
4.3 MATERIAL AND METHODS.....	51
4.3.1 Data Set collection	51
4.3.2 Preprocessing techniques	52
4.3.3 Measurement of all features for the segmentation part.	54
4.3.4 Segmentation	56

4.3.5	Segmentation using Chan Vese with Split-Bregman method and selection	60
4.3.6	Segmentation using the Chan-Vese model without any specific selection	62
4.3.7	Segmentation using Chan-Vese with Split-Bregman method and without selection. 66	
4.3.8	Comparison of segmentation methods	68
4.4	AUTOMATION	70
4.4.1	Automatic Algorithm	70
4.4.2	Exclusion criteria	72
4.4.3	Semi-automatic algorithm	75
4.5	FINAL RESULTS	75
4.5.1	Analysis of the Results	78
4.6	DISCUSSION FOR THE SEGMENTATION PART	81
4.7	CLASSIFICATION USING OUR SEGMENTED MEASUREMENTS	82
4.8	MATERIAL AND METHODS.....	82
4.8.1	Data collection and preparation:	82
4.8.2	Model Selection	83
4.9	RESULTS	83
4.10	DISCUSSION.....	85
CHAPTER 5 GENERAL DISCUSSION, FUTURE WORK and social impact		87
5.1	GENERAL DISCUSSION	87
5.2	FUTURE WORK	89
5.3	SOCIAL IMPACT.....	90
REFERENCES		91
APPENDIX A.....		103
A.1	the state of the art ON IVF	103
A.2	Studies on the classification and prediction the outcome of IVF and ICSI using machine learning techniques.....	104
A.3	Ultrasound measurements	108
A.4	Segmentation using Learning based approaches:	112
APPENDIX B GLOSSARY		116

LIST OF FIGURES

Figure 1–1: Average percentages of IVF success rate at Al Hadi Lab.....	4
Figure 1–2: Average percentages of ART cycles resulting in live births [3].....	5
Figure 1–3: The procedure of oocyte retrieval [4].....	6
Figure 2–1: A functional diagram for our SPIRL Version 1.....	11
Figure 2–2: Workflow of the SPIRL V1.....	14
Figure 2–3: Principle of the k-fold algorithm [25].....	16
Figure 2–4: A comparative principle among feature selection techniques	17
Figure 2–5: Choice the feature selection method based on the type of the data [27].	19
Figure 2–6: Machine learning algorithms [29].....	21
Figure 2–7: Accuracy and AUC for different machine learning models.....	22
Figure 3–1: A general view of SPIRL version II.....	24
Figure 3–2: The workflow of the SPIRL V2.....	26
Figure 3–3: Data decomposition employed in our study.....	29
Figure 3–4: The feature selection techniques employed in our study, showcasing the parameters that were subjected to tuning.....	33
Figure 3–5 : Selection of number of components in PCA.....	34
Figure 3–6: Algorithm comparison between Boosting and Bagging classifier.....	34
Figure 3–7 : (a) The ROC curve of all models, (b): The confusion matrix of the BC model (94 patients).	37
Figure 3–8: Comparison of the accuracy between Doppler, clinical and the combination of Doppler and clinical (94 patients).	38
Figure 3–9: Comparison of the AUC between Doppler, clinical and the combination of Doppler and Clinical (94 patients).....	39
Figure 3–10: (a) The ROC curve of the 10 classifiers, (b) The confusion matrix for the highest model (572 patients).....	40
Figure 3–11:The Boxplot of Model performance.	41
Figure 3–12: Feature ranking of the RF Importance feature selection technique.	42
Figure 3–13: Accuracy of the classifiers on three files for 572 patients. (Doppler, clinical and all parameters)	42
Figure 3–14: AUC of the classifiers on three files for 572 patients. (Doppler, clinical and all parameters).....	43
Figure 4–1: A Flowchart for the segmentation of the endometrium.....	48
Figure 4–2: (a) Selection of the important part of the image. (b) Manually mask of the endometrium using GIMP software.	52
Figure 4–3: Measurement of endometrial thickness.....	55
Figure 4–4: Measurement of endometrial surface area.....	56
Figure 4–5: Measurement of gray levels of the endometrium.....	56
Figure 4–6: Distribution of automatic BUS image segmentation approaches [78].	59
Figure 4–7: Demonstration of the principle of regularization.....	59
Figure 4–8: Selection of the area of interest.....	60
Figure 4–9: Successive treatments applied to the area of interest. The image on the left represents the original image, while the image on the right depicts the final preprocessed image.....	61
Figure 4–10: (a) Erosion applied to our detection., (b) Labeling of the detected shape, (c) Dilatation applied to our detection.	62
Figure 4–11: Segmentation using the Split-Bregman method with the endometrium selected in green and visualized alongside the manual mask represented in blue.....	62
Figure 4–12: First stage of image processing in the context of segmentation using the Chan-Vese model without prior selection. The left image is the initial image, the middle image is the cropped image and the image in the right is the final image after	63
Figure 4–13: Second stage of image processing in the context of segmentation using the Chan-Vese model without prior selection.	64
Figure 4–14: Selection of the shape after segmentation.....	65
Figure 4–15: Closing the contours of the shape.	66

Figure 4–16: Segmentation with the Chan-Vese model without endometrium selection, visualized in green, compared to the manual mask in blue.	66
Figure 4–17: Preprocessing of images before segmentation using the Split-Bregman method....	67
Figure 4–18: Shape selection and contour closure (identical to the Chan-Vese model).	67
Figure 4–19: Segmentation of the endometrium using the Split-Bregman method, visualized in green, overlaid with the manual mask in blue.	68
Figure 4–20: Results obtained for automatic segmentation.	69
Figure 4–21: Results obtained for semi-automatic segmentation.	69
Figure 4–22 : Ideal results obtained by summing the two methods.	70
Figure 4–23: Format of the saved results.	71
Figure 4–24: Workflow diagram of the automatic algorithm.	73
Figure 4–25: Activity diagram of the semi-automatic algorithm.	77
Figure 4–26: Results obtained for automatic segmentation.	79
Figure 4–27: Distribution of Jaccard indices for automatic segmentation.	79
Figure 4–28: Results obtained for semi-automatic segmentation.	80
Figure 4–29: Distribution of Jaccard indices for semi-automatic segmentation.	80
Figure 4–30: Final Results.	81
Figure 4–31: The Complete scheme of SPIRL Version III.	82
Figure 4–32: Histograms represent the feature Ranking after an Extra Tree Importance Feature selection technique for the two files.	84
Figure 4–33: The accuracy and AUC of semi-automated and SPIRL Version III.	84
Figure 5A-5–1 : Image of 3D ultrasound	108
Figure A-5–2: Image of pulsed and color Doppler	109
Figure A-5–3: Image shown the measurement of endometrial volume and the endometrial mean grey.....	109
Figure A-5–4: This image represents the measurement of endometrial power Doppler indexes.	110
Figure. A -5–5: Measurement of endometrial thickness.....	110
Figure A-5–6: A triple line endometrium represents in this image.....	111
Figure A -5–7: Image shown the measurement of myometrium volume and the myometrium mean grey.....	111
Figure A-5–8: Image represents the measurement of myometrium power Doppler indexes.....	112

LIST OF TABLES

Table 2-1: The Clinical parameters with their definition.....	13
Table 2-2: A thorough classification of feature selection approaches, outlining the notable benefits and drawbacks associated with each methodology.....	19
Table 3-1: Doppler parameters with their definitions.....	28
Table 3-2: A comprehensive categorization of wrapper feature selection methods, presenting a detailed overview of the significant advantages and disadvantages associated with each methodology.....	30
Table 3-3: A comprehensive categorization of wrapper feature selection methods, presenting a detailed overview of the significant advantages and disadvantages associated with each methodology.....	31
Table 3-4: Techniques of hyperparameter tuning search with respect to the number of iterations, the number of iterations in order to obtain the optimal hyperparameters, the score and the time elapsed in second.....	36
Table 3-5: The precision, recall, F1 score and the accuracy for all models for 94 patients.....	36
Table 3-6: Representation of the precision, recall, F1 score and the accuracy for all models for 572 patients.....	39
Table 4-1: The most popular filters to eliminate noise from ultrasound images	53
Table 4-2: The values of PSNR, MSE and SNR for each filter applied on our data.....	54
Table 4-3: The segmentation techniques with their characteristics, advantages and disadvantages.....	58
Table 4-4: Exclusion criteria chosen in the segmentation.....	74
Table 4-5: The final results of the semi-automatic and automatic segmentation.....	78
Table 4-6: The selected features with their number.....	83
Table 5-1: Some studies for the segmentation of an organ from ultrasound images.....	103

LIST OF ABBREVIATIONS

AdaBoost:	Adaptive Boosting Classifier
AI:	Artificial Intelligence
ANOVA:	Analysis of Variance
ANN:	Artificial Neural Network
AUC:	Area Under the Curve
BG:	Bagging Classifier
CV:	Cross-Validation
CVM:	Chan-Vese Model
DT:	Decision Tree
EFS:	Exhaustive Feature Selection
GB:	Gradient Boosting Classifier
ICSI:	Intra-Cytoplasmic Sperm Injection
IVF:	In-Vitro Fertilization
KNN:	K-Nearest Neighbor
LR:	Logistic regression
LS:	Level-Set
NB:	Naïve Bayes Classifier
RF:	Random Forest
RFE:	Recursive Feature Elimination
RFECV:	Recursive Feature Elimination Cross-Validation
ROC:	Receiver Operating Characteristic
SPIRL:	System for Predicting the success Rate of IVF using machine Learning combining clinical and Doppler data.
SVM:	Support Vector Machine
US:	Ultrasound machine
VOCAL:	Virtual Organ Computer Aided Analysis
XGB:	Extreme Gradient Boosting classifier

CONTEXT

Located in the heart of Beirut, Al HADI IVF Center stands as the leading fertility center in Lebanon since its establishment in 2005. With a commitment to providing the most advanced techniques in the region, the center offers state-of-the-art services to its patients. Dr. Chadi Fakih, the founder of Al HADI IVF Center, brings a wealth of experience and exceptional education to the field, having successfully contributed to the conception of thousands of babies through the utilization of the latest assisted reproductive techniques.

Medicine remains up-to-date with the latest advancements in assisted reproductive technology, ensuring that patients receive the most current and effective treatments available. Moreover, the clinicians emphasize honesty, professionalism, and unwavering support, offering hope to couples seeking to fulfill their dream of having a child.

The motivation behind exploring the factors that influence IVF outcomes arises from the strong desire to enhance the success rates and overall effectiveness of the procedure. Given the complex and multifaceted nature of IVF, understanding and identifying these factors hold immense significance for both clinicians and researchers. This recognition has led to the idea of conducting systematic studies, thoroughly analyzing various parameters such as patient characteristics, hormone levels, embryo quality, and other relevant factors. Through these endeavors, the aim is to gain a deeper understanding of these parameters and develop strategies to optimize the success rates of IVF treatments.

During his experience in the field of IVF, Dr. Chadi has observed a relationship between endometrium and myometrium 3D Power Doppler parameters and the outcomes of IVF treatments. The success of IVF treatment relies on the important roles played by the endometrium. The endometrium, which is the inner lining of the uterus, undergoes cyclic changes throughout the menstrual cycle. It thickens and prepares itself for the potential implantation of a fertilized embryo. In the context of IVF treatment, the receptivity and quality of the endometrium are critical factors in achieving successful implantation. Before the embryo transfer, the endometrium is carefully assessed for its thickness, texture, and hormonal responsiveness to ensure it is optimal for implantation. On the other hand, the myometrium refers to the middle layer of the uterine wall, composed of smooth muscle tissue. Its primary function is to contract during labor to facilitate childbirth. While there is limited research on the correlation between endometrial blood flow and the outcomes of IVF, it remains uncertain whether the same correlation exists for myometrium blood flow. However, in the laboratory of ALHADI, accurate information about endometrial blood flow and myometrium blood flow and its correlation with IVF outcomes is considered vital. Changes in treatment protocols, such as the addition of medications, may be influenced by this information. Ultimately, the decision-making process involving the couple is crucial in determining the appropriate course of action.

Choosing to conduct a thesis on predicting the outcome of IVF with the ALHADI center presents a valuable research opportunity with clinical relevance. It allows for the exploration of innovative methodologies, collaboration with experts, and a chance to enhance IVF outcomes and patient well-being. Access to extensive clinical data and IVF outcomes facilitates thorough and dependable analysis. The research findings hold the potential to greatly influence couples undergoing IVF treatments, enabling them to make informed decisions, effectively manage expectations, and alleviate associated stress.

CHAPTER 1

GENERAL INTRODUCTION

1.1 BACKGROUND

Infertility is a medical condition characterized by the inability to achieve a clinical pregnancy after 12 months or more of regular unprotected sexual intercourse. The reported prevalence of infertility varies significantly due to different research methodologies. However, global estimates indicate that approximately 48.5 to 72.4 million couples worldwide experience infertility. Among couples of reproductive ages, the prevalence of infertility ranges from 12.6% to 17.5% on a global scale, with higher rates observed in regions such as the Americas, Western Pacific, Africa, and Europe. These regional disparities in prevalence reflect variations in sexual and reproductive health, access to and quality of healthcare, and societal and cultural factors, which are further influenced by environmental conditions. Infertility is not exclusively attributed to women; it can affect both men and women. Approximately one-third of infertility cases are caused by issues specific to women, while another one-third is attributed to male factors. The remaining cases involve a combination of male and female factors or have unidentified causes [1].

However, the global assisted reproductive technology (ART) market has experienced significant growth in terms of the number of clinics and procedures performed. The assisted reproductive services market was valued at approximately USD 26 billion in 2019 and is projected to reach USD 45 billion by 2025. Between 1997 and 2016, ART treatments saw a remarkable increase, with more than a fivefold increase in Europe, a 4.6-fold increase in North America, and a threefold increase in Australia and New Zealand [2].

The market solutions for enhancing the success rates of IVF encompass a range of strategies and advancements within the field of assisted reproductive technology (ART). ALHADI Laboratory, prioritizes the improvement of success rates.

As shown below, the success rate of IVF cycles in Al HADI IVF Center is around 65% and increases to 85% in case of cumulative IVF cycles (two to three times repeat of the IVF cycle) also, this percentage increase in case of Preimplantation Genetic Screening (PGS), a screen embryo for chromosomal abnormalities to 70% since studying the genetic information prior to implantation increases the implantation. In addition, the success rate of cumulative PGS (two to three times repeat) increases to 90%.

At this level, the main objective is to improve the success rate from 65% to 80% using a procedure or AI, we could then imagine approaching 100% for several cumulative IVF cycles, leading to a revolution in the field of IVF.

Assisted reproductive technology (ART) is widely recognized as a crucial biomedical intervention for addressing infertility in couples or individuals worldwide. However, significant disparities exist in the availability, quality, and delivery of infertility care services. Despite ART being available for over four decades, it remains inaccessible or unavailable to the majority of individuals in resource-poor settings. In addition to being costly, ART procedures are often time-consuming, physically and emotionally demanding, and offer no guarantees of success (see Figure 1-1) [1].

Success rates of assisted reproductive techniques (ART) can vary and are influenced by several factors such as the age of the partners, the cause of infertility, the expertise and experience of the clinic, different assisted reproductive procedures, the use of fresh or frozen eggs, and the use of fresh or frozen embryos.

IVF and ICSI are two assisted reproductive methods employed to address infertility. In IVF, the process entails combining eggs and sperm in a laboratory dish, followed by the transfer of resulting embryos into the uterus for potential pregnancy. In contrast, ICSI is a modified version of IVF that involves the direct injection of a single sperm into an egg to assist with fertilization. ICSI has the potential to increase the likelihood of achieving a successful pregnancy. Due to this advantage, ALHADI Laboratory has incorporated ICSI as one of its preferred methods for assisting couples in their journey towards fertility.

The success rate of IVF hinges primarily on the occurrence of a positive pregnancy test or the detection of a gestational sac within 21 days following the transfer of the embryo into the uterus. This represents the concluding stage of the IVF process.

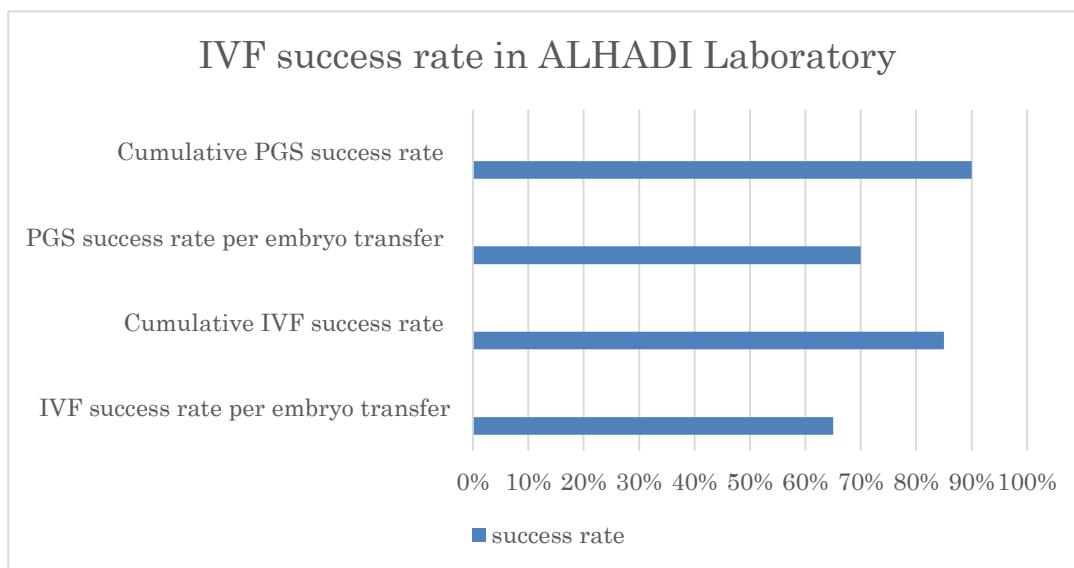


Figure 1–1: Average percentages of IVF success rate at Al Hadi Lab.

The Centers for Disease Control and Prevention (CDC) in the United States collect data on success rates of assisted reproductive techniques from certain fertility clinics [3]. A graph illustrating the average percentages of assisted reproductive cycles resulting in live births based on the mother's age is shown Figure 1-2. This study shows that the success rate decreases with the age of the woman.

In many countries, such as Lebanon, infertility is often overlooked due to competing health priorities, relatively high fertility rates, and large family sizes, which can mask the prevalence of infertility within populations and discourage public funding for infertility treatment. As a result, publicly funded infertility treatments are either limited or nonexistent, and they are generally not covered by health insurance schemes, despite the heavy financial burden on patients.

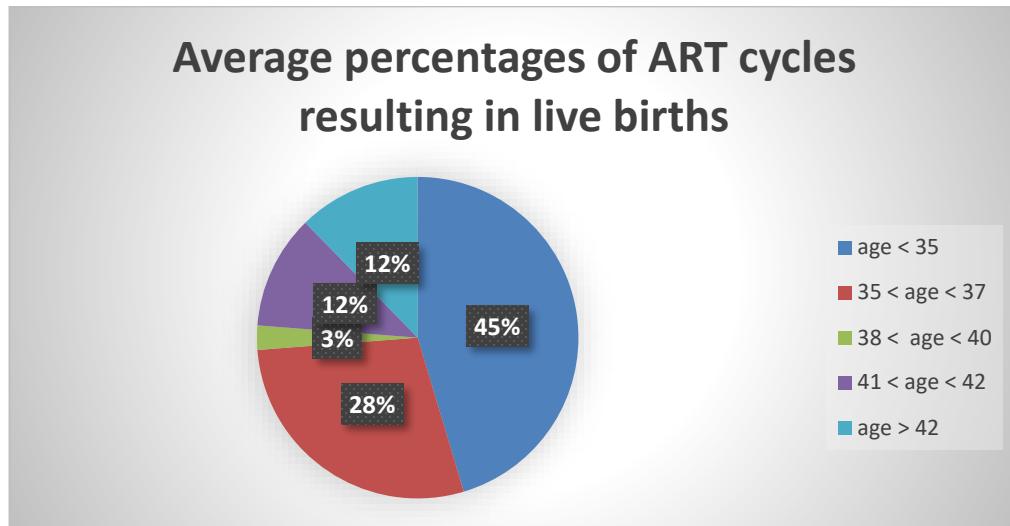


Figure 1–2: Average percentages of ART cycles resulting in live births [3].

The lack of government capacity or commitment in addressing infertility means that many couples are forced to pay for their treatment out of pocket, making cost a significant barrier to accessing such treatment. However, the deep desire to have a child often compels couples to make substantial financial sacrifices and endure severe financial hardships in order to access infertility care. Although assisted reproduction can be costly and time-consuming, it has enabled many couples to conceive a child when they would have otherwise faced challenges.

The success rate of this procedure can vary from country to country. Therefore, it is crucial to consider high success rates when choosing a facility to achieve your goal of having a baby.

1.2 PROBLEM STATEMENT

When a couple decides to undergo an IVF procedure, the physician begins by providing information about possible reasons for infertility and the selection criteria for IVF. A series of consultations then follow. The fertility specialist offers an initial consultation, during which they review medical history, conduct a physical examination, and gather relevant information.

To ensure a favorable outcome, numerous tests, which can vary depending on individual circumstances, are typically necessary to assess the couple's fertility status and determine the most appropriate therapeutic approach. These various pre-IVF protocol tests are performed to evaluate fertility factors, some specifically designed for women and others for men. These tests include ovarian reserve and tubal patency testing, screening for infectious and genetic diseases, chlamydia screening, rubella immunity testing, and sperm quality evaluation, to name a few.

Additionally, an ultrasound examination is performed to clearly examine the endometrium and myometrium, extracting qualitative and quantitative measurements through gray-scale image levels. Pulsed and color Doppler allow measurements of uterine and endometrial/myometrium perfusion through

receptivity factors such as the vascularization index (VI), flow index (FI), and vascularization flow index (VFI).

Additional examinations may be recommended based on the individual's medical history and risk factors. Finally, couples may be required to attend counseling sessions to address the emotional, psychological, and ethical aspects associated with IVF treatment.

After this series of consultations, if the couple agrees to follow the doctor's advice (complete smoking cessation, optimizing body mass index, maintaining a healthy lifestyle), the doctor can provide them with a prediction of the treatment success rate and advice. This prediction, based solely on the doctor's experience and available information, even if imperfect, represents an important milestone as it can be the starting point of the IVF process.

The challenge we aim to address in this thesis is the establishment of an "automatic" prediction of IVF success rate. This new measure will assist and support the doctor in their own prediction of IVF success rate. The doctor can then advise the couple on the continuation of the protocol.

If the couple decides to proceed with the IVF protocol, the prospective parents must sign a consent form, acknowledging that they understand and accept all the steps and risks associated with IVF. As you can see in Figure 1-3, the IVF/ICSI cycle continues with stimulation treatment, ovarian activation with gonadotropins, ultrasounds to monitor follicle development (following serum hormone injections), final oocyte maturation and Human Chorionic Gonadotropin test (HCG) administration, oocyte retrieval through transvaginal aspiration, artificial insemination, embryo transfer, and finally, pregnancy testing.

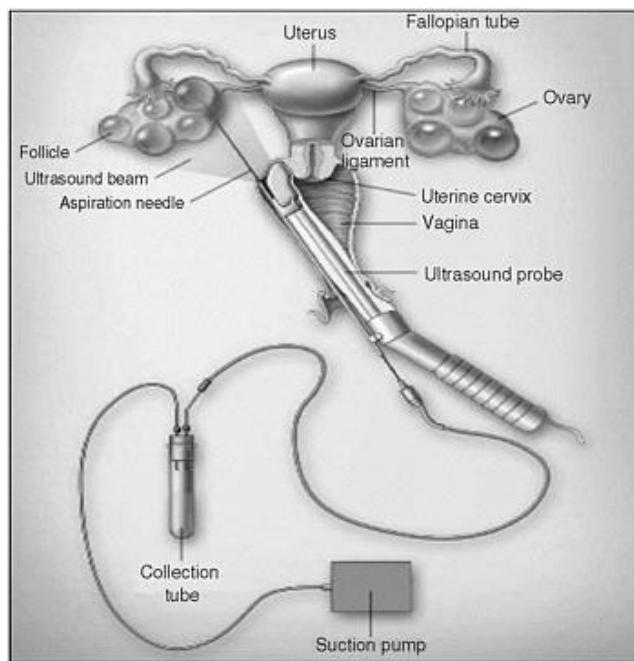


Figure 1–3: The procedure of oocyte retrieval [4].

1.3 STATE OF THE ART OF PREDICTION SYSTEMS FOR THE SUCCESS RATE OF IVF/ICSI

One out of seven couples is affected by infertility. Therefore, assisted reproductive technology (ART), particularly in-vitro fertilization (IVF), is regarded as a common method for encouraging the birth of children among infertile people [4] [5]. IVF is a sophisticated assisted reproductive therapy that has advanced since the first pregnancy achieved with this technique in 1979. The rates of pregnancy after IVF treatment range from 30 to 70%, depending on the female's age and other variables. The causes of successful pregnancies have been extensively investigated [6]. Age, body mass index (BMI), uterine and ovarian variables, the kind of stimulation protocol utilized, the stimulation dose, and the use of fresh or frozen protocols are all factors that can affect the outcome of ARTs [7] [8]. To help practitioners make informed decisions, machine learning algorithms have been suggested to automate the intricate modeling of the IVF procedure for result prediction and patient counseling. These models are capable of exploring and analyzing the population pattern in a sizable dataset and can aid in the extraction of the dataset's latent knowledge.

A variety of machine learning techniques are widely used for prediction [9], including Artificial Neural Networks (ANN), Support Vector Machines (SVM), Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), k-Nearest Neighbors (KNN), Multilayer Perceptron (MLP), Radial Basis Function (RBF), Recursive Partitioning (RPART), Adaptive Boosting (AB), Light Gradient Boosting Machine (LGBM), Extreme Gradient Boosting (XGBoost), Logistic Regression (LR), Linear Discriminant Analysis (LDA), Hill Climbing Wrapper Algorithm (HCWA), Classification and Regression Trees (CART), and Conditional Inference Tree (CIT). Table A-1 in the appendix A summarizes the studies that predict the outcome of IVF using machine learning techniques. The table shows for each study the sample size, the machine learning technique used, the features selection technique, the selected features, and the performance. Among these studies, Kaufmann et al. [10] used ANN to identify the most influential factors. They then used those as input features to train a neural network. Uyar et al. [11] looked into a variety of classifiers to predict the outcome of embryo implantation in IVF. They used the Receiver Operating Characteristics (ROC) curve to analyze the performance of the classifiers. Better results were obtained using the NB and RBF classifiers. In a further investigation conducted in 2015, these authors reduced the feature set by weighing each feature's relative relevance, and this resulted in a marginally improved performance by NB [12]. Moreover, Guh et al. [4] proposed a hybrid method that integrates both genetic algorithm (GA) and DT techniques to study 67 IVF attributes. The GA algorithm was used to identify the best set of features and the learning parameters of the DT model simultaneously. This integration of GA and DT reduced the number of significant features to twenty-eight. In addition, Durairaj & Kumar [13] used 14 male and female variables to predict the success rate of IVF treatment. Hafiz et al. [14] investigated five techniques to predict the success of IVF. They reported that the RF and RPART outperformed other methods. In 2019, Qui et al. [15] applied different variables to compare predictive performances based on the AUC. XGBoost and RF

algorithms achieved higher AUCs. Amini et al. [5] predicted the outcomes of Intracytoplasmic Sperm injection (ICSI) using male and female clinical attributes. In 2020, Hassan et al. [16] proposed a Hill climbing wrapper algorithm to select the best subset of clinical features. SVM attained the highest accuracy. Additionally, Liu et al. [17] used variables such as clinical female factors to predict ICSI success rates. Barreto et al. [18] used five techniques with male and female clinical factors and showed that RF achieved the best performance while Wen et al. [19] applied six classifiers and showed that XGBoost attained the best performance.

As early as 2012, ultrasound assessment has been used and found to be useful in predicting outcomes in assisted reproduction patients [20] [21]. Indeed, Zhang et al. [22] demonstrated the benefit of adding new echo-Doppler parameters to clinical parameters. As a matter of fact, ultrasound examination of endometrial receptivity can detect a variety of combinations of imaging parameters. The prediction models used a variety of ultrasound features, including endometrial volume and endometrial vascularization index. This study was the only one that used subendometrial blood flow as measures to demonstrate the correlation with IVF. Nevertheless, myometrium receptivity has not been addressed in [22]. Based on Dr. Chadi's observations, this study will serve as our starting point in establishing our own prediction system. **This is the first ingredient that will serve as the foundation for our proposed solution in this thesis.** However, Zhang et al. [22] have developed simple nomograms combining clinical and Doppler parameters that predict early clinical outcomes in embryo transfer cycles using the parameters of age, birth BMI, type and number of embryos transferred, endometrial thickness, flow index (FI), resistivity index (RI), pulsatility index (PI) and the number of endometrial and subendometrial blood flows. In the training process, the AUC showed a value of 0.698, and similar results were shown in the subsequent validation cohort. Furthermore, none of the studies described above used machine learning approaches with endometrial and myometrium Doppler measurements. **Therefore, machine learning approaches with endometrial and myometrium Doppler measurements were the second ingredient that served as the foundation for our proposed solution in this Thesis.**

1.4 OBJECTIVE OF THESIS

The challenge we are tackling is to provide doctors with a "pre-IVF protocol" score for the success rate of IVF based on diverse data. To achieve this, we will develop a tool which we will refer to as SPIRL (System for Predicting the success Rate of IVF using machine Learning combining clinical and Doppler data). The SPIRL tool aims to revolutionize the prediction of IVF/ICSI success rates by combining machine learning techniques with clinical and Doppler data. The goal is to provide an automated tool capable of calculating the success rate of ICSI procedures based on various data.

The primary purpose of the SPIRL tool is to assist doctors in their decision-making process rather than replacing their expertise. It is a valuable aid that relies on advanced algorithms to analyze and interpret data and ultimately provide a prediction of the success rate. The current prediction of ICSI success rates depends

on the specific types of data used and the methods employed. With the SPIRL tool, the aim is to create a fully automated version in the long run that minimizes manual interventions as much as possible. The aim is to extract important measurements and features from Echo-Doppler images as automatically as possible. This involves automatically extracting Echo-Doppler images and their corresponding measurements and features. Three different approaches are considered for extracting biomarkers from Echo-Doppler images for analysis: manual, semi-automatic, and fully automatic segmentation independent of the operator. These approaches aim to minimize dependence on human operators and reduce potential variations or biases in the biomarker extraction process.

By developing an automated tool like SPIRL, healthcare professionals can benefit from faster and more accurate forecasts of ICSI success rates. This information can help guide treatment decisions and improve the overall efficiency of the ICSI process.

The proposed strategy for SPIRL includes the following steps:

- State of the art
- Collection of clinical parameters and power Doppler parameters.
- Creation of three separate databases: one containing Doppler parameters, another containing clinical parameters, and a third combining both clinical and power Doppler parameters.
- Use of different machine learning models to predict the outcome of ICSI for each database.
- Comparison of the obtained results for each dataset to evaluate the impact of power Doppler parameters on the success rate.

1.5 HYPOTHESIS OF THESIS

The combination of established clinical parameters and new indicators aims to improve the success rate of ICSI. This approach involves the use of advanced techniques such as power Doppler ultrasound, which allows measurement of the volume and vascularization indices of the endometrium and myometrium, as well as the mean grayscale values of these uterine layers.

The vascularization indices of the endometrium and myometrium are quantitative measures that assess the density, intensity of blood vessels and perfusion of blood flow in these uterine layers. These indices provide valuable insights into the overall vascularization of the uterus, which can influence the receptivity of the endometrium and myometrium and the success of embryo implantation.

The mean grayscale values represent the intensity of pixels in the ultrasound image of the endometrium and myometrium. These values reflect tissue characteristics and can indicate the health and quality of these uterine layers. Deviations from normal mean grayscale values may indicate underlying issues that could impact the success of ICSI.

By combining these advanced measures with established clinical parameters, healthcare providers can obtain a more comprehensive and accurate assessment of the uterine environment. This information allows for personalized treatment planning, facilitating the identification of potential issues and the implementation of appropriate interventions to optimize the success rate of ICSI.

By using various machine learning techniques and different feature selection methods, it is possible to identify key factors that influence the outcome of intracytoplasmic sperm injection (ICSI). In particular, the importance of Doppler parameters can be demonstrated, enabling the prediction of the success rate of ICSI outcomes.

Virtual Organ Computer-aided Analysis (VOCAL software) measures the Power Doppler parameters. The drawback of using the VOCAL software is that it can be time-consuming. The manual interaction and adjustments required during the segmentation process can prolong the analysis time, which may not be ideal in time-sensitive clinical settings. So, it is important to explore automated segmentation methods that can achieve comparable or even superior accuracy while minimizing the time required for the process. Making a fully automated prediction of the outcome of ICSI is a challenge.

1.6 THESIS OUTLINE

The thesis is structured into five chapters, each serving a specific purpose. Chapter one begins by introducing the study. It presents background information, the objectives, and defines the scope of the work. Chapter two provides a preliminary work regarding the use of machine learning in predicting the success of IVF/ICSI procedure using clinical parameters. Chapter three presents a thorough analysis of the success of IVF/ICSI procedure using both clinical and Doppler parameters. Chapter four presents a detailed description and analysis of the segmentation techniques employed in the research work with the application and the results. And then explores the classification of the segmentation results. Chapter five concludes the thesis by summarizing the overall discoveries and contributions of the research work. Additionally, it highlights the areas for future research and potential avenues for further exploration.

CHAPTER 2

CLASSIFICATION BASED ON CLINICAL DATA (SPIRL VERSION I)

2.1 INTRODUCTION

Our initial approach serves as a first draft for the SPIRL (System for Predicting the success Rate of IVF using machine Learning combining clinical and Doppler data) we propose, which aims to use machine learning to improve the estimation of the optimal endometrial implantation period.

The functional diagram of our prediction system consists of several functions as illustrated in Figure 2-1. It integrates influential clinical attributes of IVF. Then, a preprocessing technique is a must for the management of the data. Through feature selection techniques, we identify the most significant factors that impact the outcome of IVF/ICSI. Our goal is to assist doctors in making informed decisions regarding the most appropriate treatment for each couple, aiming to achieve a 100% success rate in conceiving a baby.



Figure 2-1: A functional diagram for our SPIRL Version 1.

2.2 MATERIALS AND METHODS OF SPIRL VERSION I

2.2.1 Flowchart

The study's workflow is depicted in Figure 2-2. After gathering clinical parameters, the raw data underwent preprocessing using data classification tools before being inputted into the machine learning application. Missing values were handled, and data normalization was performed using Min-Max Scaling, a function from the Scikit-learn library. Categorical data, such as embryo quality (grade 0 for good quality and grades 1-3 for bad quality), was encoded into numerical representations using One Hot Encoding, also available in Scikit-Learn [23]. The data was then split into training and testing sets with an 80/20 ratio using the train-test-

split function from Scikit-Learn. The training set, consisting of 80% of the data, was used for hyperparameter tuning, feature selection, and feature extraction. Stratified k-fold cross-validation was employed during training to ensure effective results. The remaining 20% of data was reserved for evaluating the final model's performance after hyperparameter tuning and feature engineering, allowing unbiased scoring of prediction results.

Feature selection was employed to reduce the number of features in the dataset, enhancing the classifier's performance by selecting important features that maximize separation between classes. ANOVA was employed as the feature selection technique for each classifier in our initial version of SPIRL. The selection of this particular technique was based on its widespread use in the current state of the art for our specific data type. To determine the optimal values of hyperparameters, three different optimization algorithms (Grid Search, Bayesian Search, and Random Search) were compared for each classifier. These techniques differ in their search methods for finding the best hyperparameters.

Machine learning algorithms, Artificial Neural Networks (ANN), Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), were used to automate the modeling of IVF procedures for result prediction and patient counseling. Tree-based classifiers, in particular, were chosen for their effectiveness, low cost, and interpretability compared to more complex algorithms. To predict pregnancy outcomes after ICSI, the performance of prediction models built using five classifiers (RF, LR, SVM, NB, KNN) was compared. The evaluation of each model's performance involved using different metrics, such as accuracy and AUC-ROC, to select the best model for predicting the outcome of ICSI.

The calculations in this study were conducted using the Python programming language [23], utilizing various libraries suitable for machine learning tasks, including sklearn.metrics, model_selection, ensemble, Decomposition, Discriminant Analysis, and feature selection. Additionally, commonly used libraries such as NumPy, Matplotlib, and Skopt were employed. Classification accuracy is the most commonly used metric for evaluating classification models. It is preferred due to its simplicity in calculation and interpretation, providing a single number summarizing the model's performance. A higher AUC (Area Under the Curve) indicates better prediction of 0 (non-pregnant) and 1 (pregnant) class.

2.2.2 Data Collection

Between January 2018 and May 2018, data was collected from all patients undergoing ICSI treatments at ALHADI IVF Center. The study involved the inclusion of all eligible patients who met the criteria for participation. Every patient approached willingly provided their consent and actively engaged in this study. The data exhibits a balanced distribution, making it well-suited for utilization in machine learning algorithms. For this retrospective study, a total of 94 patients who underwent ICSI were included. The inclusion criteria were women under the age of 45 with at least one embryo of good morphology. Clinical parameters including age, embryo and oocyte characteristics such as the number of embryo transfers, retrieved oocytes, oocytes fertilized, 2PN (pronuclei) stage, utilization rate, retrieved embryos, embryo quality, and the day of transfer were extracted from the "TrakMD platform"

application, which was filled out by the embryologist at ALHADI IVF Center. The selection of these clinical parameters is founded upon an extensive review of numerous articles in the literature. These carefully chosen clinical parameters have demonstrated their significance in predicting the outcome of IVF and play a crucial role in determining its success. Additionally, the extensive availability of data on the platform and the low occurrence of missing values contribute to the selection of these clinical parameters. The abundance of data ensures a robust analysis, while the minimal missing values enhance the reliability and accuracy of the predictions in IVF outcomes. Table 2-1 presents the clinical parameters used in our study along with their respective definitions.

Table 2-1: The Clinical parameters with their definition.

Clinical Parameters	Definition
Age (Numerical)	The age of the patient
Number of the Embryo Transferred (ET) (Numerical)	The number of embryos transferred into the uterus per ICSI cycle.
Transfer day (Numerical)	The day of transfer.
Oocytes Retrieved (Numerical)	The number of Oocytes retrieved during ICSI treatment.
Oocytes Put in Fertilized (Numerical)	The number of Oocytes put in fertilization during ICSI treatment.
Embryos sampled (Numerical)	The number of embryos chosen in the treatment
Utilization rate (Numerical)	The rate of live baby born per oocyte
2 PN (Numerical)	Normal fertilization of an oocyte is defined by observing two distinct pronuclei (2PN) and two polar bodies after insemination
The quality of the Embryo (Categorical)	The quality of transferred embryo determined by the embryologist based on morphology.

2.2.3 Inclusion and Exclusion Criteria

All medical procedures were conducted according to standard protocols, without any additional or unusual diagnostic or treatment procedures. Therefore, the recorded analysis in this case is retrospective research that does not have any direct impact on patient care. The inclusion and exclusion criteria play a crucial role in the study, as they allow for the selection of a specific study population. The exclusion of certain criteria serves multiple purposes. Firstly, by ensuring a homogeneous study population, we can derive more precise conclusions regarding the factors that influence IVF outcomes. This increases the likelihood of identifying the specific factors that impact the success of IVF. Additionally, the exclusion of certain criteria helps eliminate confounding effects that may arise due to unique characteristics of individuals.

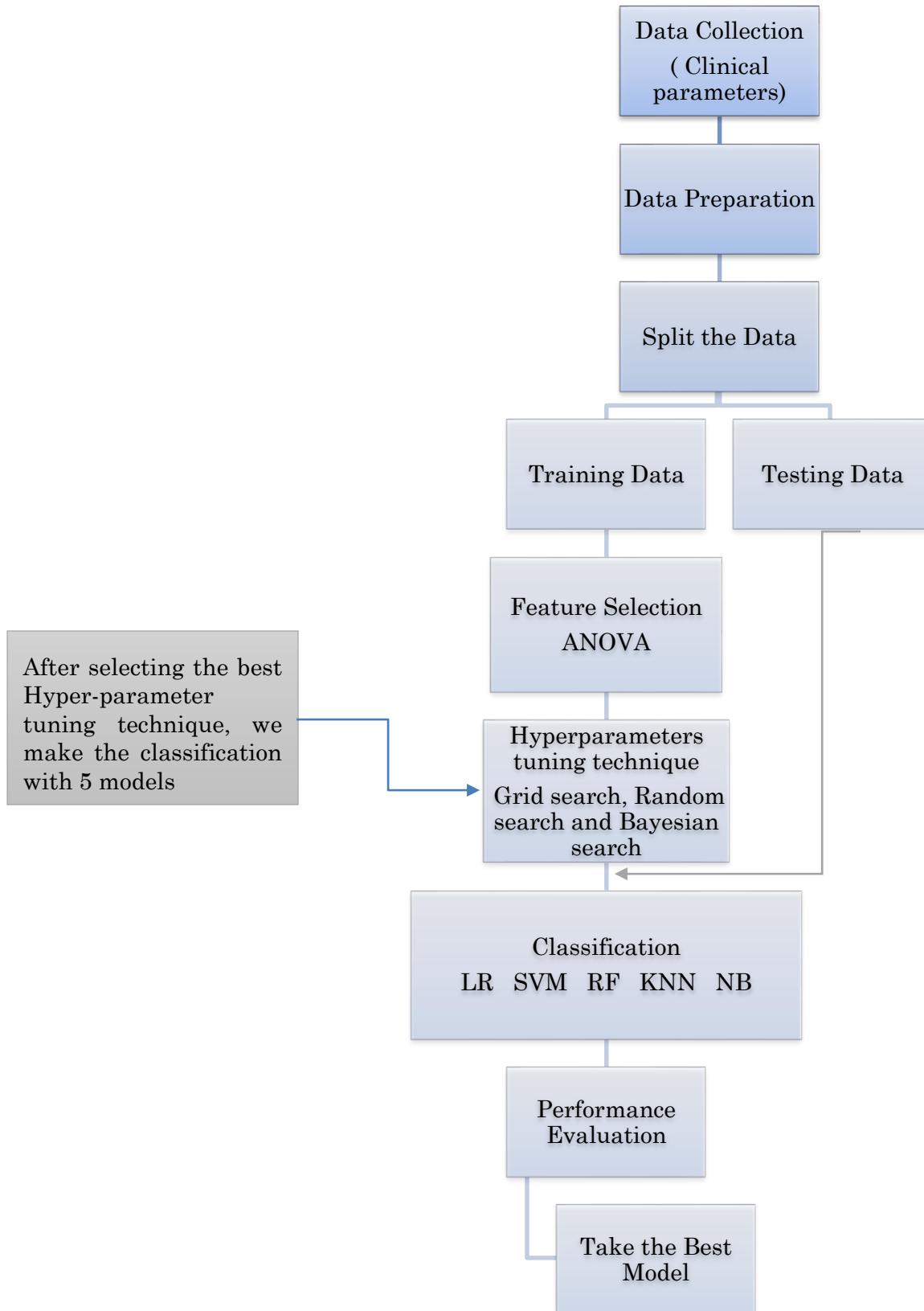


Figure 2–2: Workflow of the SPIRL V1.

The inclusion and exclusion criteria for this study were as follows:

- **Inclusion Criteria:**

Women who had one or 2 good quality embryos.

- **Exclusion Criteria:**

1. Women over age 45;
2. Poor responders;
3. Inappropriate endometrium for implantation;
4. Abnormal anatomy of uterine cavity;
5. Patients with fibroids and previous myomectomy.

For instance, individuals with uterine abnormalities and those who have undergone previous myomectomy may be at a higher risk of complications during IVF. By excluding such patients, we can focus on the specific factors affecting IVF outcomes without the potential confounding effects of these conditions. Similarly, poor responder patients may possess distinct characteristics that require different approaches. Including them in the study could complicate the analysis and potentially obscure the interpretation of results. Therefore, their exclusion allows for a more focused analysis of IVF outcomes.

2.2.4 Data preprocessing

During the data cleaning and preprocessing phase, various techniques are used to transform raw data into a suitable format that can be effectively utilized by machine learning algorithms. This involves performing typical steps to ensure data quality and prepare it for further analysis. Missing data can be problematic for machine learning models. Patients with more than one missing value, identified through a thorough search in TrakMD, will be excluded from the analysis. For patients with a single missing value, the missing value will be replaced using the Mean Imputation method for numerical features. However, in the case of missing values in the quality of the embryo feature, the patient will be removed from the dataset. This decision is based on the significance of this feature on the results and its categorical nature (0 or 1). All the factors were numerical, except for embryo quality, which was a categorical data type that needed to be encoded as numerical data. Normalizing the data is a process that aims to bring all features to a similar scale.

Cross validation (CV) is the ability to estimate a model's accuracy or quality using new, previously unseen data throughout the training process. This means that even during the learning process, it is possible to forecast how the AI would perform in practice. During this phase, the data set is divided into two sections: training data and test data. The training data is used to learn and update the model's weights during model training. The test data is then used to independently confirm the model's correctness and validate how good it already is. Depending on the outcome, either a new training phase or the program is terminated [24].

Among the available options, the two most suitable choices for our data are Stratified k-fold cross-validation and Repeated Stratified k-fold cross-validation. For each model, I perform both types of cross-validation and select the one that yields the highest performance.

a. K-fold cross-validation

With this technique, the entire dataset is partitioned into k equal-sized portions, and each partition is referred to as a fold. It is termed k-fold because it is divided into k pieces, where k can be any integer. The model is trained using the remaining $K-1$ folds and one-fold for validation. This technique is repeated k times until each fold is used once as a validation set and the remaining folds are used as a training set. Figure 2-3 illustrates the k-fold strategy. This validation strategy is not recommended for imbalanced datasets because the model will not be effectively trained due to the proper data ratio for each class. This validation strategy is used in the case of regression rather than classification [25].

b. Stratified k-fold cross-validation

K-fold validation cannot be used for imbalanced datasets since data is separated into k -folds with a uniform probability distribution. This is not the case with stratified k-fold, a more advanced version of the k-fold cross-validation technique. Despite the fact that it, too, divides the dataset into k equal folds, each fold has the same proportion of target variable instances as the full dataset. This enables it to work flawlessly with uneven datasets but not with time-series data. The ratio of instances of the target variable is preserved in all folds in the stratified k-fold cross-validation procedure, which is an advantage [25].

c. Repeated stratified cross validation

Model performance estimates based on k-fold cross-validation can be noisy. This means that a different k-fold split of the dataset can be implemented each time the technique is run, resulting in a different distribution of performance scores and a different mean estimate of model performance. The model employed and the dataset itself determine the degree of variation in projected performance from one run of k-fold cross-validation to the next [25].



Figure 2–3: Principle of the k-fold algorithm [25].

2.2.5 Feature selection techniques:

There are three main approaches to feature selection: filter techniques, wrapper approaches, and embedded techniques. Figure 2-4 below explains the principle of each technique to select the features.

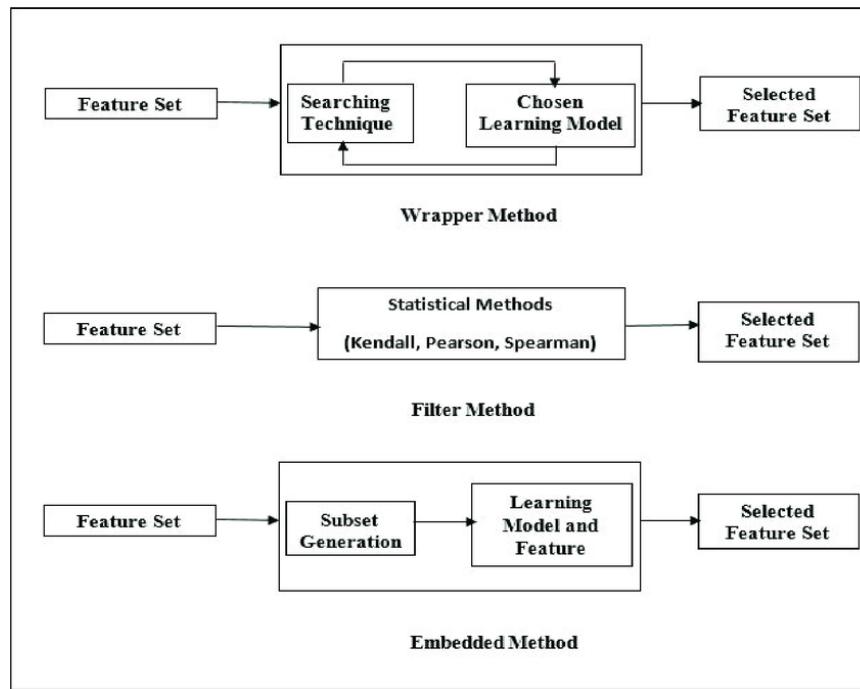


Figure 2–4: A comparative principle among feature selection techniques.

Filter techniques determine the importance of features based solely on their inherent properties. They generate a feature relevance score and eliminate features with low values. Filter techniques are computationally simple and fast, can handle high-dimensional datasets, and are independent of the classification algorithm. Once feature selection is done, different classifiers can be evaluated using the selected features.

Wrapper approaches combine the search for the best model hypothesis with the search for the optimal feature subset. They construct a search technique to generate and evaluate different feature subsets. This process is specific to a particular classification algorithm, as it trains and evaluates a unique subset of features using a specific model. Wrapper approaches allow for interaction between feature subset search and model selection, but they can be computationally intensive, especially if the classifier is expensive to construct.

Embedded techniques incorporate the search for an ideal feature subset into the process of building the classifier. They search in the combined space of feature subsets and hypotheses. Embedded methods, like wrapper approaches, are tailored to a single learning algorithm. They offer the advantage of interacting with the classification model while being less computationally costly than wrapper approaches [26].

These categories vary in terms of how they approach feature selection, including whether it is treated as a separate step or integrated into the learning algorithm. They also differ in evaluation metrics, computational complexities, and their ability to identify redundancies and interactions among features. Each methodological category possesses its own set of advantages and limitations, making them more appropriate for specific use cases. Table 2-2 presents a comprehensive classification of feature selection approaches, highlighting the significant advantages and disadvantages of each methodology. Additionally, it provides examples of the most influential examples within each category.

2.2.6 Why ANOVA from filter method in SPIRL V1?

In this particular task, the objective is to develop a classification predictive model using numerical input variables. This type of classification problem is frequently encountered across different domains. Traditionally, correlation-based techniques are utilized to tackle such problems. However, in the present work, it is crucial to consider the categorical nature of the target variable when applying these techniques [27].

You can see in the below figure (Figure 2-5) that we have two common correlation coefficients used when I have numerical input and categorical output which are ANOVA and Kendall's rank coefficient.

It is important to note that Kendall's rank coefficient assumes that the categorical variable is ordinal, as mentioned in reference [27] and so the best filter technique for our data is ANOVA feature selection technique.

In ANOVA, the selection of the optimal value for "k" is performed using GridsearchCV to maximize the mean accuracy with the best configuration. The aim is to identify the best subset of features that yields the highest performance. In ANOVA, the function "f_classify" is used in order to calculate the "F_test" feature importance.

2.2.7 Model selection

Machine learning algorithms can automate the intricate modeling of the IVF procedure for result prediction and patient counseling. Table A-2 in Appendix A presents various models used to predict the outcomes of IVF and ICSI. It is evident that no single model stands out as the best. Therefore, conducting experiments using different models on our data is necessary to achieve optimal results.

So, to predict pregnancy outcomes after IVF, we compared the performance of prediction models built using logistic regression (LR), Support Vector Machine SVM (SVM), Naïve Bayes Classifier (NB), K-Nearest Neighbor (KNN) and Random Forest (RF). In Figure 2-6, you can gain a comprehensive understanding of the algorithm used for each classifier by examining the overall view.

Table 2-2: A thorough classification of feature selection approaches, outlining the notable benefits and drawbacks associated with each methodology.

Feature selection techniques	Advantage	Disadvantage	Examples
Filter	<ul style="list-style-type: none"> • Fast • Independent of the classifier • Generalization ability is good 	<ul style="list-style-type: none"> • No feature dependencies • No interaction with the classifier 	<ul style="list-style-type: none"> • Chi2 • ANOVA • Fisher score • Gain information
Wrapper	<ul style="list-style-type: none"> • Simple • Feature dependencies • Interaction with the classifier 	<ul style="list-style-type: none"> • More expensive • More prone to overfitting • Classifier dependent selection 	<ul style="list-style-type: none"> • Boruta • Forward Feature Selection • Backward Feature selection • Recursive Feature Elimination • Exhaustive Feature selection
Embedded	<ul style="list-style-type: none"> • Feature dependencies • Interaction with the classifier • Lower computational cost than wrapper 	<ul style="list-style-type: none"> • Classifier dependent selection 	<ul style="list-style-type: none"> • Random Forest Importance • Extra Tree Importance • LASSO Regression

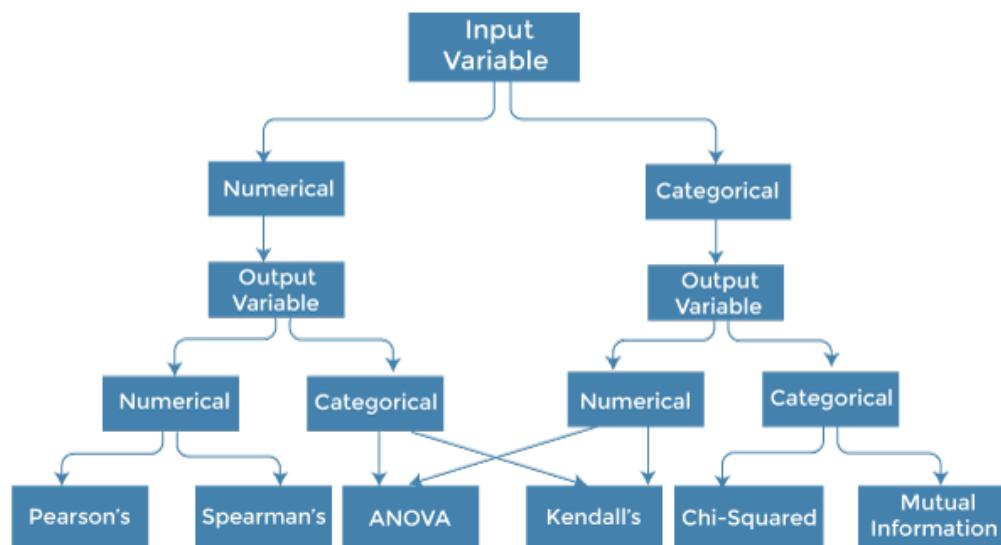


Figure 2–5: Choice the feature selection method based on the type of the data [27].

Logistic Regression is a machine learning model used to establish the relationship between a dependent variable and one or more independent variables. It is primarily utilized for classification analysis as it predicts the probability of an event occurring. If the probability is above 0.5, it is assigned a value of 1; otherwise, it is assigned a value of 0. For instance, logistic regression can be used to predict whether a student will pass (1) or fail (0) an exam.

Random Forest is a machine learning algorithm that operates similarly to a decision tree. The decision tree principle is described in “Figure 2-6 (a)”. However, it employs multiple decision trees to make predictions, reducing the risk of overfitting as you can see in “Figure 2-6 (d)”. The algorithm performs majority voting, where the class predicted by the majority of trees is assigned to an item. For example, if two trees predict a class of 0 and one tree predicts a class of 1, the item will be assigned a class of 0.

K-Nearest Neighbor is another simple machine learning algorithm that classifies new cases based on the category or class of the nearest data points as shown in “Figure 2-6 (b)”. The number of neighbors considered is determined by the assigned value of K. For instance, if K is set to 10, the algorithm will examine the 10 nearest neighbors of the item. The nearest neighbors are determined by calculating distances, often using measures such as Euclidean distance, and selecting those with the shortest distance.

Support Vector Machines divide data points using a hyperplane, which is a straight line as you can see in “Figure 2-6 (e)”. Points represented by blue diamonds belong to one class on the left side of the plane, while points represented by green circles belong to another class on the right side. When predicting the class of a new point, its position relative to the hyperplane and within the margin is considered. If it lies on the left or right side of the hyperplane, its class can be determined.

Naïve Bayes is a probabilistic machine learning model based on Bayes' theorem. It assumes that all features are independent of each other. Conditional probability refers to the probability of an outcome occurring given that another event has occurred. Naïve Bayes predicts the probability that an item belongs to a particular class and assigns it the class with the highest probability. This is shown in “Figure 2-6 (f)” [28].

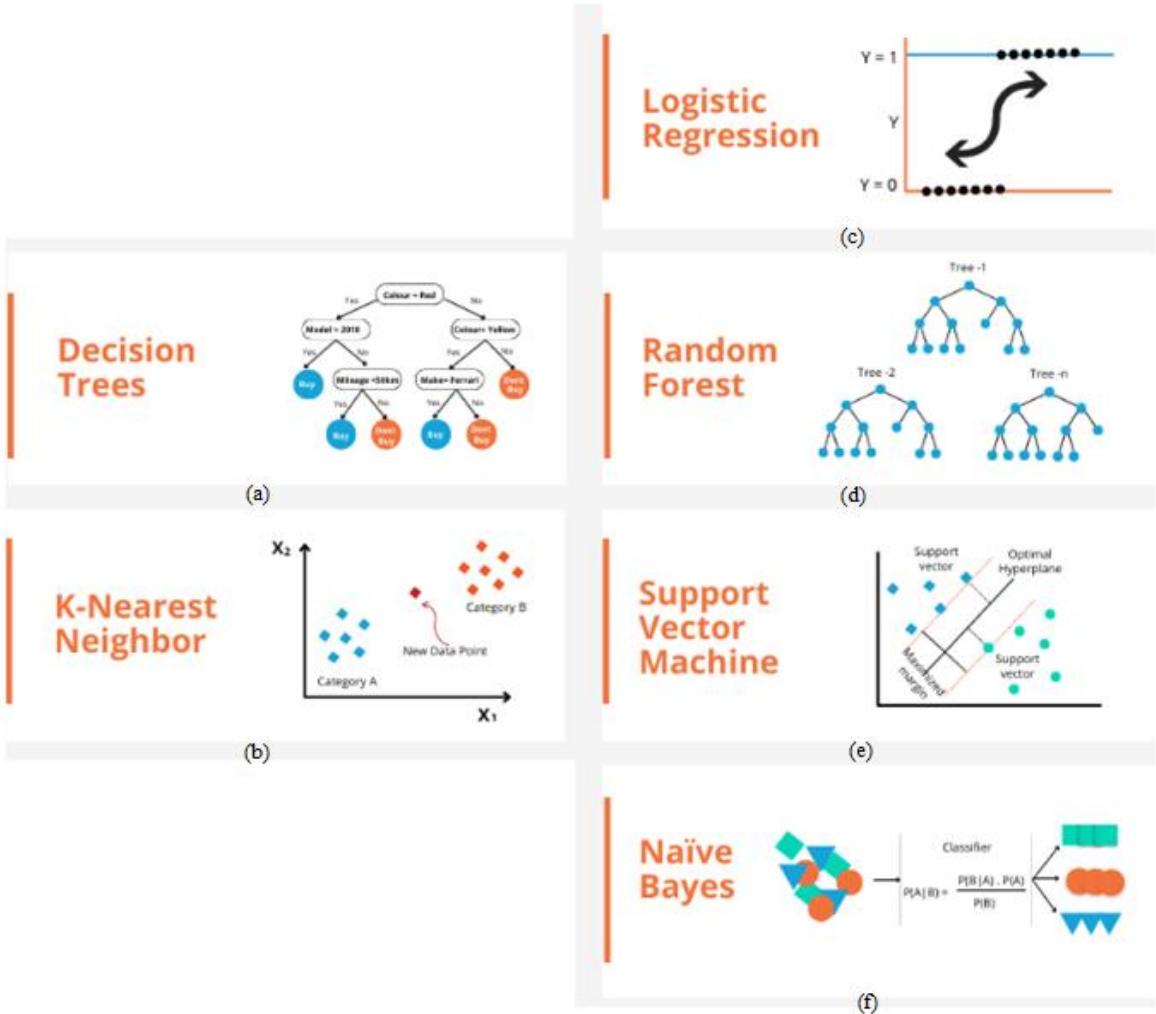


Figure 2–6: Machine learning algorithms [29].

2.2.8 Hyper-parameter tuning techniques:

To choose the best values of the hyper-parameters, we take 3 different types of optimization algorithm, and make a comparison between these three techniques for each classifier in order to obtain the best values for each classifier (Grid search, Bayesian search, and Random search). It is crucial to acknowledge that the efficacy of these methods may differ based on the particular problem, dataset, and hyperparameter space. Hence, it is advisable to experiment with various search techniques and assess their performance before reaching a final conclusion. A comparison is made between grid search, random search, and Bayesian search for each machine learning model. The choice of hyperparameter tuning techniques can be made by comparing the time taken in seconds and the corresponding scores. There are multiple hyperparameters associated with each classifier that I will incorporate into the search strategy to achieve the best possible outcomes [23].

2.3 RESULTS

The histogram presented in Figure 2-7 illustrates the performance of different models in predicting the outcome of ICSI using the ANOVA feature selection technique. It is evident that the Random Forest (RF) model outperforms the others, achieving an accuracy of 0.789 and an AUC of 0.8. On the other hand, the LR, SVM, KNN, and NB models show lower performance, with accuracies below 0.7.

Based on these results, it becomes crucial to improve the SPIRL model in the second stage to achieve more accurate results and enhance the success rate of ICSI outcomes. This can be accomplished by introducing new parameters (Power Doppler), as suggested in the introduction of this thesis, and incorporating additional models with different feature selection techniques. By doing so, we can determine the importance of these new parameters and verify if they play a crucial role in predicting the success of ICSI.

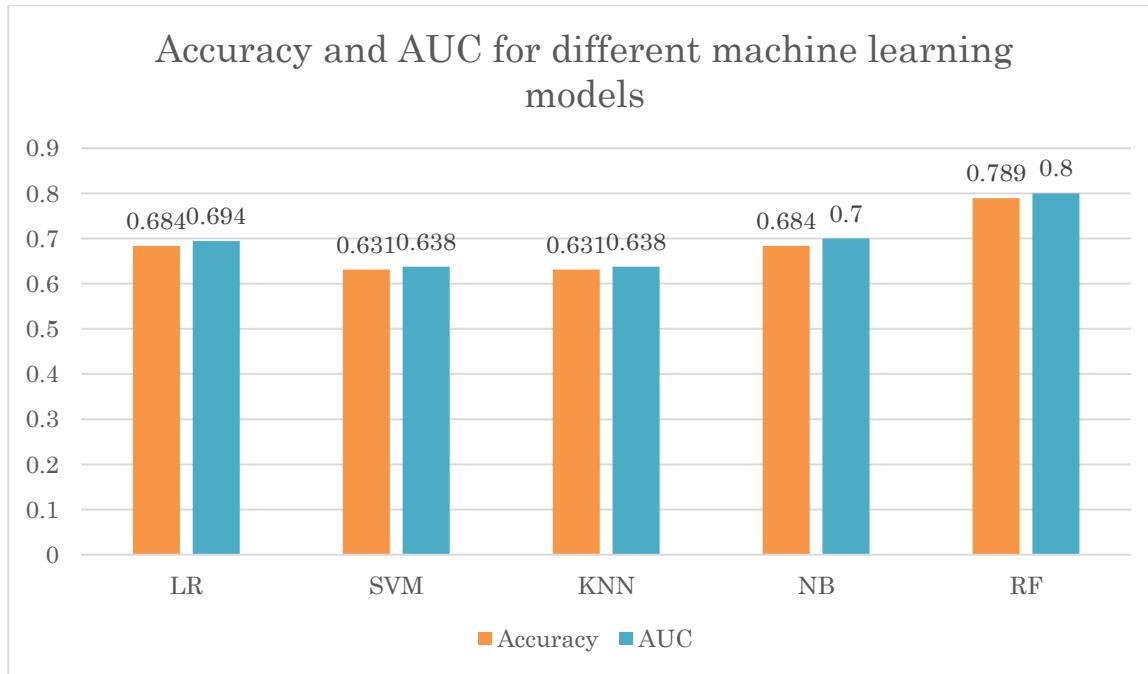


Figure 2-7: Accuracy and AUC for different machine learning models.

2.4 DISCUSSION

Machine learning has the potential to play a significant role in predicting the outcome of Intracytoplasmic Sperm Injection (ICSI). Predicting the success or failure of ICSI can be valuable for both healthcare providers and patients, as it can help guide treatment decisions and manage expectations.

Machine learning algorithms can leverage historical data on various factors associated with ICSI outcomes, such as sperm quality, egg quality, maternal age, hormonal levels, and other relevant clinical parameters. By analyzing patterns and relationships within this data, machine learning models can learn to make predictions about the success or failure of ICSI for individual patients.

The model can then be trained on the labeled data to identify patterns and develop prediction rules. Features such as sperm parameters, female reproductive health indicators, and treatment-related variables can be used to train the model. The features in SPIRL V1 was a clinical Data based on the women factors such as the age, tentative, the day of transfer, the number of embryo transfer and some characteristics of the oocytes in different stages. Researchers and clinicians consistently conduct studies on various factors that can enhance the success rate in their centers. In future endeavors, it is crucial to incorporate new parameters and examine their impact on the outcomes. This will shed light on the effectiveness of these additional factors and their influence on the results.

However, it's important to note that the accuracy and reliability of the predictions depend on the quality and quantity of the available data. Sufficient and high-quality data, including a diverse range of cases, are crucial for training robust machine learning models. To enhance the performance and accuracy of the model, it is crucial to increase the sample size by including a larger number of patients in future studies. Furthermore, it is essential to explore various classifiers and employ different feature engineering techniques. By doing so, we can improve the model's performance and ensure more accurate predictions. Ensemble learning is a powerful technique that merges multiple models to enhance the accuracy and generalization of predictions. By utilizing diverse base learners, ensemble learning captures various facets of the data, leading to more resilient predictions. Ensemble methods have demonstrated effectiveness across diverse applications and hold a pivotal position in advancing the field of machine learning. So, it is important to show the effect of ensemble learning on our prediction.

Overall, machine learning holds promise in predicting the outcome of ICSI, potentially assisting healthcare providers and patients in making informed decisions. However, further research, validation, and integration with clinical expertise are necessary to fully realize the potential of machine learning in this area.

In the next chapter, we will focus on expanding the patient sample size, incorporating additional feature selection and feature extraction techniques, and integrating ensemble learning models to forecast the outcome of ICSI.

CHAPTER 3

CLASSIFICATION BASED ON CLINICAL AND POWER DOPPLER DATA (SPIRL VERSION II)

3.1 INTRODUCTION

As mentioned earlier, to further improve the performance of the prediction system, we can include new data and test new classifiers (Gradient Boosting, Bagging Classifier, Radom forest, AdaBoost, Xtreme Gradient Boosting). Also, in this updated version of SPIRL, the Doppler parameters have been incorporated into the data entered into the classifiers. This addition aims to examine the impact of these Doppler parameters on the results and outcomes of the study. By including the Doppler parameters, the researchers can assess how these specific variables influence the performance of the classifiers and potentially provide valuable insights into their predictive capabilities. Figure 3-1 provides an overview of the SPIRL system, illustrating the sequential steps involved. Initially, data collection and parameter measurements are conducted as the first step. Subsequently, the collected data is organized into three separate files. The next phase involves the selection and analysis of the most significant features. Finally, the prediction of the outcome of ICSI is presented, providing couples with valuable information regarding the overall status of their treatment. This schematic representation showcases the systematic process followed by the SPIRL system, leading to informative predictions for the benefit of the couples undergoing ICSI treatment.

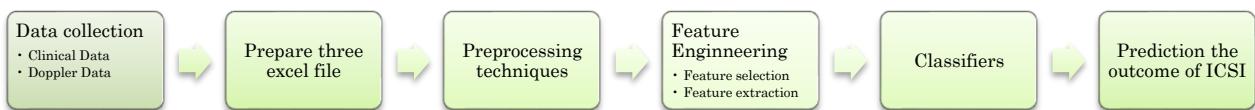


Figure 3-1: A general view of SPIRL version II.

3.2 MATERIALS AND METHODS

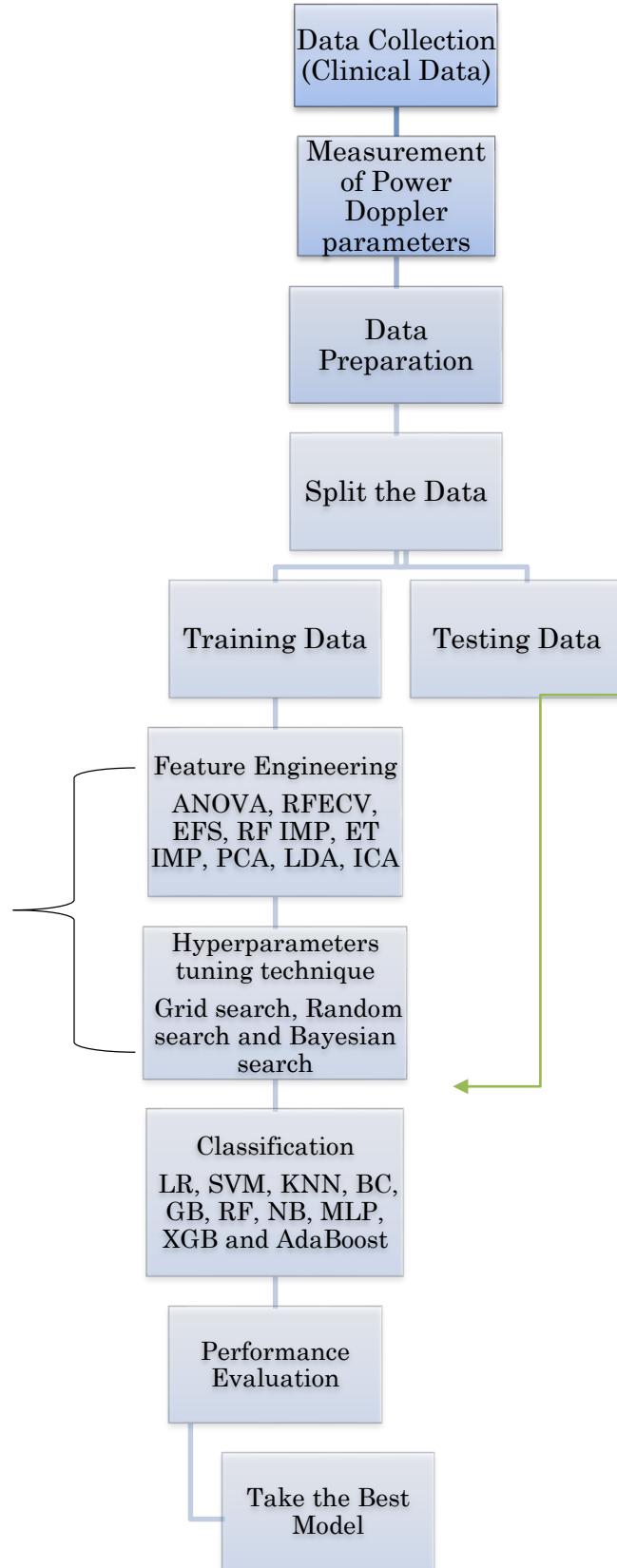
3.2.1 Flowchart

The workflow of this study is shown in Figure 3-2. After measuring 3D power Doppler parameters, raw data were preprocessed using data classification tools before being fed into the machine learning application. Handling missing values, normalization using Min-Max Scaling, a function from Scikit-learn. Also, in Scikit-Learn, we used One Hot Encoding which is a function that allowed to encode the features-based categorical data into numerical ones, such as the quality of the embryo

transfer which was the dependent variable encode 0 for embryo with bad embryo quality (grade 1, grade 2, and grade 3) and 1 for embryo with good quality (grade 0).

For training and testing, data was split to 80/20 using train-test-split function in Scikit-Learn. The remaining 80% of data was used solely for training, which includes the processes of hyper-parameter tuning, feature selection, and feature extraction procedures. To ensure an effective training result, stratified k-fold cross-validation was performed in the process. The remaining 20% was set aside until the end of both hyperparameter tuning and feature engineering and was exclusively utilized to evaluate the performance of the final model. This strategy is a frequent method for evaluating a prediction's model performance during training since it ensures nonbiased scoring of the prediction results [29].

Recall that the goal of feature extraction was to reduce the number of features in a dataset by producing new ones from existing ones. The new reduced set of features should then be able to summarize the majority of the information in the original set of features. Feature Selection was another method for reducing the number of features in a dataset [30]. So, the performance of the classifier can be enhanced through the selection of a combination of important features that represent the maximal separation between the classes. In this study, for each classifier, we applied ANOVA, RF Importance, Extra Tree Importance, Exhaustive Feature Selection (EFS) and Recursive Feature Elimination Cross-Validation (RFECV) as feature selection techniques and PCA, LDA, and ICA as feature extraction techniques. Then, to choose the best values of the hyper-parameters, we took 3 different types of optimization algorithms and made a comparison between these three techniques for each classifier in order to obtain the best values for each classifier (Grid search, Bayesian search, and Random search). These three types of tuning techniques differ from each other by the method of search using a grid search or using a Bayesian strategy or we can select a random search for the best hyperparameters. To help practitioners make informed decisions, machine learning algorithms can automate the intricate modeling of the IVF procedure for result prediction and patient counseling. The models are capable of exploring and analyzing the population pattern in a sizable dataset. These models also aid in the extraction of the dataset's latent knowledge. A variety of machine learning techniques, including Artificial Neural Networks (ANN), SVM, DT, RF, and many other classifiers are widely used for prediction [9]. A tree-based classifier is one of the earliest and most widely used types of classifiers for prediction tasks due to its effectiveness and low cost, and it has been utilized for a variety of activities, including IVF. Tree-based classifiers have a lower computing cost than more contemporary techniques while yet providing adequate or even good performance. The foundation of tree-based classifiers is a set of rules that determines the value of a target attribute based on a set of if-else situations. This strategy is deceptively basic, but it has been shown to be useful in a wide range of situations. Another advantage is that the resulting set of rules is understandable, as opposed to the whole black-box function of other more complex algorithms [29]. So, to predict pregnancy outcomes after ICSI, we compared the performance of prediction models built using (i) 5 models' tree-bases classifier such as Gradient Boosting (GB), Bagging Classifier (BC), RF, AdaBoost Classifier, and Extreme Gradient Boosting (XtremeGB), and (ii) 5 other models such as LR, SVM, NB, MLP, KNN).



After selecting the best Feature selection technique and the best Hyper-parameter tuning technique, we make the classification with 10 models

Figure 3–2: The workflow of the SPIRL V2.

Finally, the evaluation of the performance of each model using different metrics, to select the best model in order to predict the outcome of ICSI, such as accuracy, confusion matrix, Precision-recall, F1 score, and AUC-ROC. The calculation for this study was carried out entirely in the Python programming language, which makes use of a variety of existing libraries that could be particularly suitable for machine learning tasks. The commonly utilized libraries include sklearn.metrics, model_selection, ensemble, Decomposition, Discriminant Analysis, and feature selection. Additionally, widely-used libraries such as Numpy, Matplotlib, and Skopt are frequently employed in these contexts, among others.

In our work, we are interested in identifying all possible pregnant cases after treatment and so the recall, and also being sure that if we identify a pregnant case that one is actually a real pregnant case. Therefore, we have to focus on both precision and recall, that's why the F1 score was made for. The most commonly used parameter for evaluating classification models is classification accuracy. It is widely used because it is simple to calculate and explain, and it is a single number that summarizes the model's capacity. AUC represents the degree or measure of separability, while ROC is a probability curve. It indicates how well the model can distinguish between classes. The higher the AUC, the better the model predicts 0 classes as no pregnant and 1 class as pregnant.

3.2.2 Data Collection

First, in order to perform other machine learning models and other feature engineering techniques, I take the data between January 2018 and May 2018, Data was acquired from all patients undergoing ICSI treatments at ALHADI IVF Center (Center of Treatment by Intra Cytoplasmic Sperm Injection in Lebanon, Beirut, Lebanon). All medical acts were carried out in a standard manner, with no additional or typical procedures for diagnosis or treatment. As a result, the recorded analysis, in this case, is retrospective research with no impact on patient care. Each ICSI cycle was recorded in the "TrakMD platform" database. In this investigation, 94 ICSI patients were used.

Table 3-1 illustrates the 3D power Doppler parameters employed in our research, along with their corresponding definitions. Consequently, I conducted measurements of these parameters precisely as I detailed in Table A-2 in Appendix A. These parameters were added to the clinical parameters presented in Table 2-1.

Then, we took the data between January 2018 and December 2021, this part of the work included 572 patients undergoing ICSI outcome. Here, I repeat all the work on this larger data to get more precise results.

As you can see in Figure 3-3, in order to compare the results and show the importance of 3D Power Doppler parameters, we created three files with the same data patients. 3D Power Doppler and clinical parameters were included in the first and second files respectively. The third file included all Doppler and clinical parameters. For each file, I followed and repeated the steps described below.

3.2.3 Data Pre-Processing

These steps involve converting raw data into an appropriate format that machine learning algorithms can effectively work with. Several typical techniques are employed during the cleaning and preprocessing phase.

Table 3-1: Doppler parameters with their definitions.

3D Power Doppler Parameters	Definition
Volume of the Endometrium (EV) (Numerical)	The volume of the endometrium
The Mean Grey of the Endometrium (Endo MG). (Numerical)	The histogram of grey-level of the uterine-endometrium.
Vascularization Index of the Endometrium (Endo VI) (Numerical)	vascularization index which represents the vessel density in the endometrium.
Flow Index of the Endometrium (Endo FI) (Numerical)	flow index which represents the intensity of blood flow in the endometrium.
Vascularization Flow Index of the Endometrium (Endo VFI) (Numerical)	vascularization flow index which represents the endometrial perfusion.
Volume of the Myometrium (MYO V) (Numerical)	The volume of the Myometrium
The Mean Grey of the Myometrium (MYO MG) (Numerical)	The histogram of the grey-level of the uterine-myometrium.
Vascularization Index of the Myometrium (MYO VI) (Numerical)	vascularization index which represents the vessel density in the myometrium.
Flow Index of the Myometrium (MYO FI) (Numerical)	flow index which represents the intensity of blood flow in the myometrium.
Vascularization Flow Index of the Myometrium (MYO VFI) (Numerical)	vascularization flow index which represents the myometrium perfusion.

To maintain the integrity of our results, we chose to remove patient data for any other parameters that had missing values. As a result, our dataset decreased from 1,080 patients to 572 patients. Additionally, we normalize the values to bring them to a standard scale, which facilitates data comparison and analysis. Normalization also reduces the influence of outliers while improving the accuracy and stability of statistical models [31]. It's important to note that since the "embryo" variable is categorical, it needs to be converted into numerical representations. Finally, since our data is balanced, there is no need to use techniques for imbalanced data.

3.2.4 Feature Selection Technique

Several studies have compared the predictive performance of the different feature selection methods. These comparative studies have resulted in the widely held opinion that there is no such thing as the “best method” that is fit for all problem settings [32].

No feature selection method can be considered as the definitive best method. This statement applies not only on a universal scale but also to machine learning algorithms and sets of input variables. Instead, to determine the most effective feature selection approach for a specific problem, we need to conduct careful and systematic experimentation. This involves trying out a variety of models on different subsets of features selected using various statistical measures. Through this process,

we can discover the approach that yields the best results for the particular problem at hand [27]. For this reason, this section aims to incorporate additional feature selection techniques into our updated version, SPIRL V2.

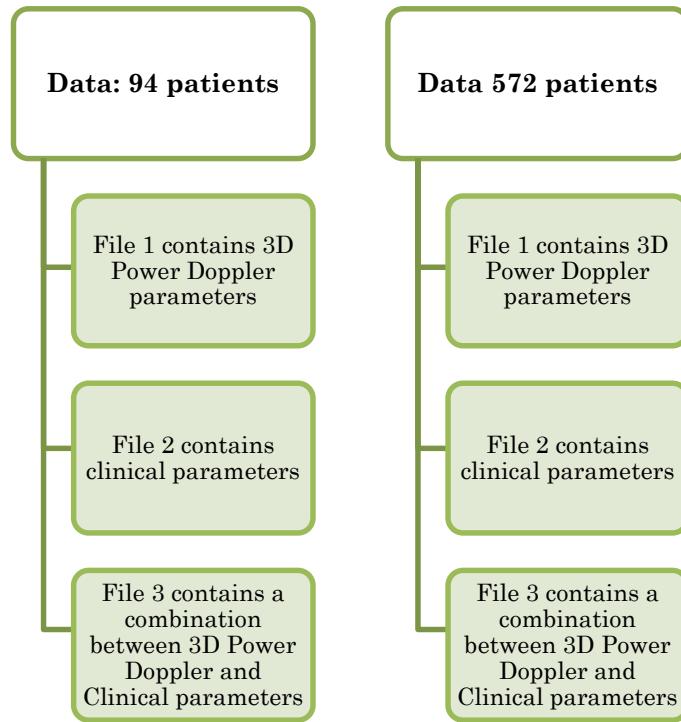


Figure 3–3: Data decomposition employed in our study.

So, different feature selection can be applied, as you can see in SPIRL V1 ANOVA, is the best technique from filter method for our data. In wrapper method, RFE and especially with cross validation RFECV and Exhaustive feature selection method (EFS) are the most used technique for the classification. In embedded method, Random Forest Importance and Extra tree importance are the most used technique. You can see in the below section why I choose these four techniques for SPIRL V2.

A. Why RFECV and EFS from Wrapper feature selection technique?

Exhaustive feature selection and recursive feature elimination have distinct advantages compared to backward and forward feature selection techniques. You can see in Table 3-2 below that forward and backward feature selection may not always result in the best subset of features. Exhaustive feature selection ensures reliable outcomes and is well-suited for datasets with a limited number of features. Additionally, Recursive Feature Elimination (RFE) offers various advantages, including accurate feature identification, robustness against overfitting, and noise resistance. For these reasons, I choose EFS and RFECV (RFECV employs cross-validation to assess the performance of various subsets of features, ensuring that the chosen features exhibit good generalization to unseen data and mitigating the potential for overfitting).

Table 3-2: A comprehensive categorization of wrapper feature selection methods, presenting a detailed overview of the significant advantages and disadvantages associated with each methodology.

Wrapper Feature selection method	Advantages	Disadvantages
Forward Feature Selection [33]	<ul style="list-style-type: none"> • Simple • Reduce the computational complexity 	<ul style="list-style-type: none"> • Prone to overfitting • Computationally expensive • Time-consuming • May not always result in the best subset of features
Backward Feature Selection [34]	<ul style="list-style-type: none"> • Simple • Can handle large feature sets 	<ul style="list-style-type: none"> • May not always result in the best subset of features
Exhaustive Feature Selection [35]	<ul style="list-style-type: none"> • Guarantees the optimal feature subset. • Suitable for small feature sets. 	<ul style="list-style-type: none"> • Time consuming
Recursive Feature Elimination [36]	<ul style="list-style-type: none"> • Can manage datasets with a large number of dimensions and accurately identify the crucial features. • Can handle relationships between different features, making it suitable for intricate datasets. • Can be utilized alongside any supervised learning algorithm. • Robust to noise and overfitting. 	<ul style="list-style-type: none"> • Could incur significant computational costs when dealing with sizable datasets. • May not be the optimal strategy for datasets containing correlated features. • Might not yield satisfactory results when confronted with noisy or irrelevant features.

B. Why RF Importance and Extra Tree Importance from Embedded feature selection technique?

You can see in Table 3-3 that LASSO regression encounter difficulties when dealing with correlated features and in general Tend to produce unstable estimates. RF Importance and Extra tree importance can be more accurate for our data.

3.2.5 Feature Extraction Techniques

The goal of feature extraction is to reduce the number of features in a dataset by producing new ones from existing ones (and subsequently discarding the original ones). The new reduced set of features should then be able to summarize the majority of the information in the original set of features. A summary version of the original features can thus be generated by combining the original set. Feature Selection is another method for reducing the number of features in a dataset. The distinction between Feature Selection and Feature Extraction is that Feature Selection seeks to rank the value of existing characteristics in the dataset and eliminate less significant ones (no new features are created) [37]. In 2021, [38] discusses popular feature extraction techniques in machine learning, focusing on PCA, ICA, and LDA.

“Yaameen” emphasizes the effectiveness of these feature extractors in capturing important features while minimizing any adverse impact on the overall performance. Here are the commonly used feature extraction techniques in machine learning. In [39], they employed PCA as a preprocessing step in combination with an Artificial Neural Network (ANN) as the classifier. The area under the receiver operating characteristic curve (AUC) ranged from 0.7670 to 0.9796. Additionally, in a study conducted by [40], PCA was utilized to analyze pregnancy data. The results showed that the classification accuracy of pregnancy varied between 61% and 80% across different groups analyzed in the obtained models. In a study conducted by [41], breast cancer classification was performed using an LDA feature extractor as a preprocessing step with an SVM classifier. The study reported an accuracy of 99.2%. Table 3-4 provides an extensive classification of feature extraction methods, offering a clear definition and an overview of the applications for each technique. In our work, I take the most used techniques in medical classification using machine learning techniques. So, for each model, I used PCA, LDA and ICA as some feature extractors in order to enhance the performance of the models.

Table 3-3: A comprehensive categorization of wrapper feature selection methods, presenting a detailed overview of the significant advantages and disadvantages associated with each methodology.

Embedded feature selection techniques	Advantages	Disadvantages
RF Importance [42]	<ul style="list-style-type: none"> highly accurate. generalize better Feature ranking 	<ul style="list-style-type: none"> Requires training a Random Forest model
Extra Tree Importance [43]	<ul style="list-style-type: none"> Robust to noise and multi-collinearity Computational efficiency Feature ranking 	<ul style="list-style-type: none"> Requires training multiple decision tree models.
LASSO Regression [44]	<ul style="list-style-type: none"> Automatic feature selection Reduces overfitting 	<ul style="list-style-type: none"> Struggle with correlated features. Generally unstable estimates.

In Figure 3-4, there are specific parameters associated with each feature selection and feature extraction technique. These parameters need to be adjusted in order to obtain optimal results. The figure illustrates the utilization of eight different feature selection and extraction techniques in machine learning models. The objective is to achieve optimal results by tuning the parameters associated with each technique.

- In **RFECV**, we can use LR (Logistic Regression) and RF (Random Forest) as estimators in classification tasks. I then assess and compare the performance of RFECV with different parameter variations to determine the most effective estimator. We import RFECV from “sklearn. feature selection”.
- In **EFS**, “mlxtend” package is used to provide the implantation of this feature selection technique. We change manually the number of minimum and

maximum features to obtain the best subsets. Min=1 and Max=5 gives the optimal number of subsets.

- In **RF and Extra Tree Importance**, we manually adjust the threshold value to sequentially eliminate features, aiming to achieve the highest performance by retaining the most important features.
- In **PCA and ICA**, to obtain the number of components we have to:
 - Standardize the range of continuous initial variables
 - Compute the covariance matrix to identify correlations
 - Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components
 - Create a feature vector to decide which principal components to keep
 - Recast the data along the principal component's axes

Table 3-4: An extensive classification of feature extraction methods, offering a clear definition and an overview of the applications for each technique.

Feature extraction techniques	Definition	Applications
PCA [38]	Statistical method used in modeling and data compression.	Image Processing, Medical Data Correlation, Facial Recognition, Time Series Prediction, Analyzing text or particular metadata fields
LDA [38]	Maximizes class separability to find a lower-dimensional representation.	for classification tasks.
ICA [38]	It is a technique for reducing the dimensionality of data by identifying a collection of statistically independent variable groups that collectively capture the maximum variability in the data.	clustering, classification, regression, and outlier detection.

So, as you can see in Figure 3-5 below for 572 patients with cut-off 95% of cumulative variance the number of components is 11.

In **LDA**, using GridSearch, we can fine tune the hyper parameters and define the grid solver, then compare the mean accuracy and get the best result. For example, the highest mean accuracy is 0.702 with the Singular value decomposition ‘svd’ solver.

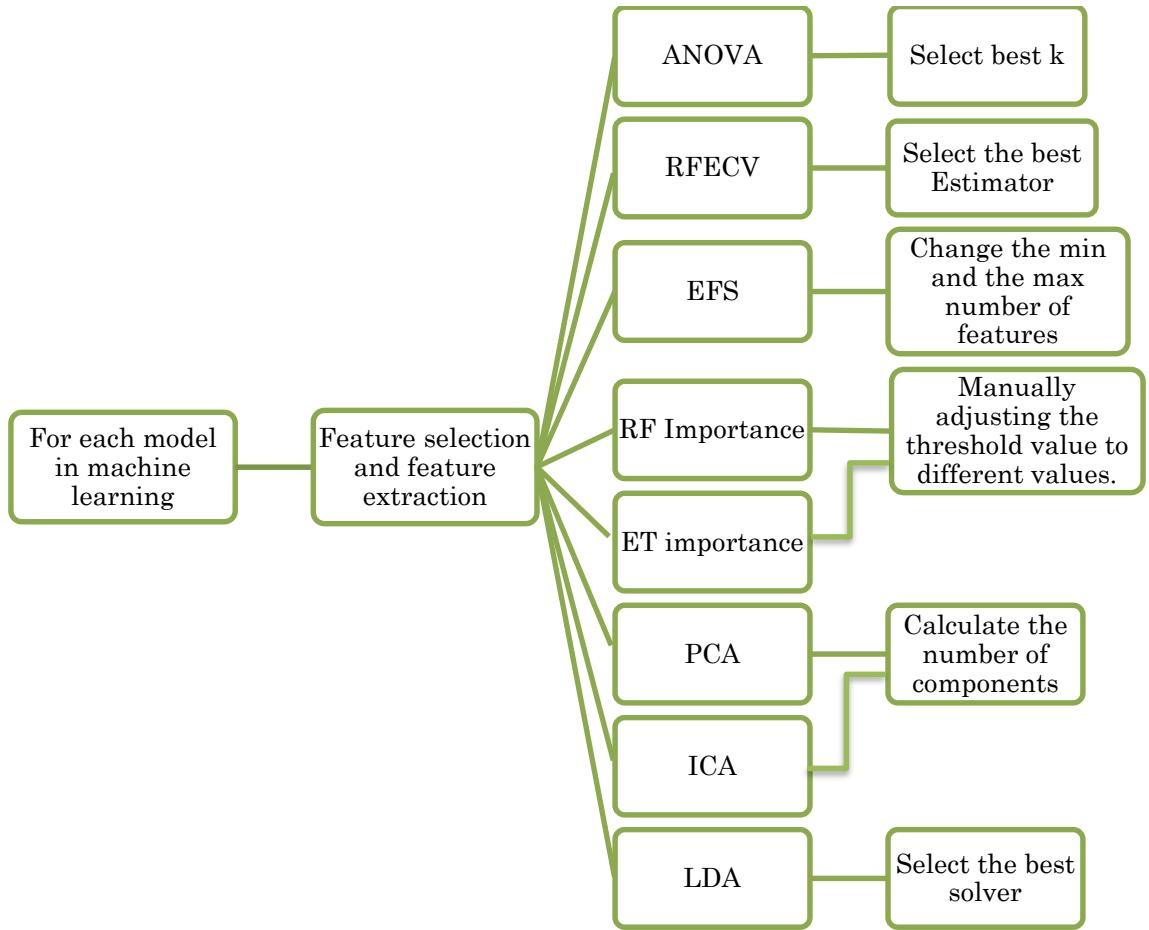


Figure 3–4: The feature selection techniques employed in our study, showcasing the parameters that were subjected to tuning

3.2.6 Model Selection

Applying complex deep learning models with high accuracy in practical settings is challenging due to their lack of explainability and the requirement for large datasets. In our case, ensemble learning methods, strike a better balance between accuracy and complexity, making them more suitable for our problem. So, it is important for SPIRL V2 to add more models in order to obtain a highest performance in the prediction of the outcome of ICSI. Ensemble learning encompasses various types of ensemble models, with boosting and bagging being two popular approaches. Figure 3-6 illustrates the key features and the underlying principles of two distinct learning processes.

Boosting involves training base models sequentially, with each subsequent model focusing on correcting the mistakes made by the previous models. It assigns greater weight to misclassified instances, enabling the ensemble to enhance its performance over iterations. Notable boosting algorithms include AdaBoost, Gradient Boosting, XGBoost, and LightGBM. Therefore, boosting:

- Follows sequential model in sequential
- Sequential weight, more weight is given to the model to better performance

- Reduce the bias.
- It generates the ultimate forecast by averaging the predictions from multiple learners.
- AdaBoost, Gradient Boosting, Extreme Gradient Boosting.

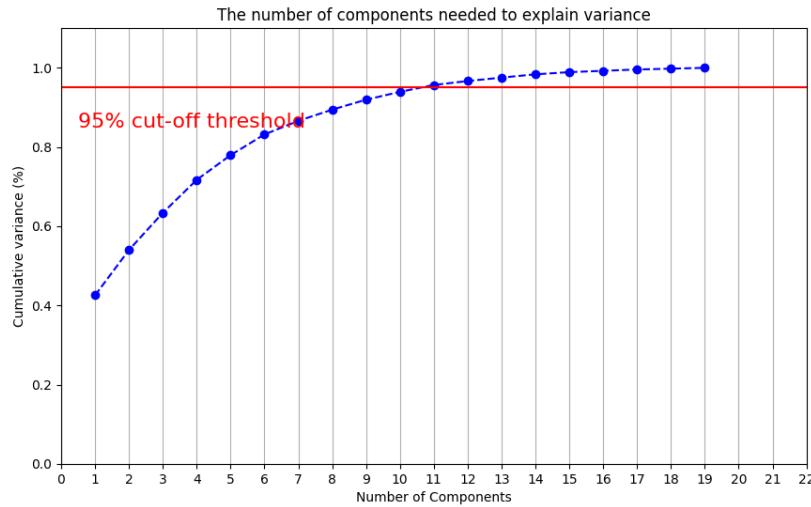


Figure 3–5 : Selection of number of components in PCA.

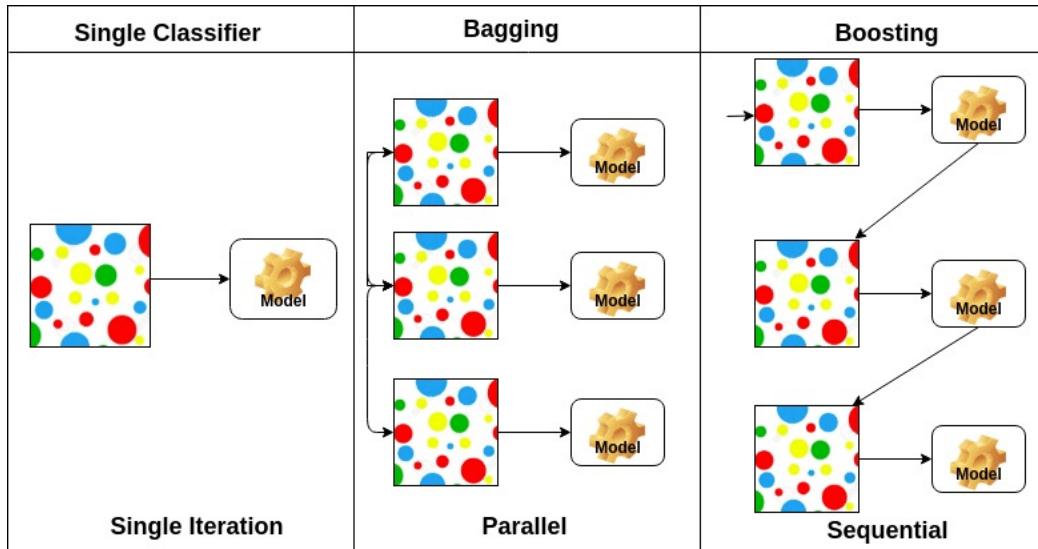


Figure 3–6: Algorithm comparison between Boosting and Bagging classifier.

Bagging, also known as Bootstrap Aggregating, entails training multiple base models independently on different subsets of the training data. These subsets are formed by randomly sampling the training data with replacement. The predictions of the base models are then combined, typically through majority voting (for

classification) or averaging (for regression). Random Forest, a well-known bagging algorithm, employs decision trees as the base models. Therefore, bagging:

- Follows multiple models in parallel
- Equal weight is given to all model
- Reduce the variance and solve the problem of overfitting
- It gives the final prediction by taking average of the learners.

Both boosting and bagging are robust ensemble techniques that endeavor to enhance the overall performance of machine learning models. They differ in their approaches to combining base models and handling errors, which can influence their performance across various datasets and problem domains. So, it is important to add Bagging classifier, AdaBoost, Gradient Boosting, XGBoost for the prediction of the outcome of ICSI in SPIRL Version 2.

However, it should be mentioned that MLP classifier is used to predict the outcome of ICSI. The MLP is chosen due to its prominence in state-of-the-art approaches for ICSI outcome prediction. For further details and information, please refer to the table provided in the Appendix.

3.2.7 Hyper-parameter tuning techniques

To optimize the performance of each model, hyper-parameters need to be adjusted. As we mentioned in the SPIRL version 1, a comparison between Grid search, Random search and Bayesian search for each model is achieved. The objective is to find the combination of hyper-parameters that yields the highest score while minimizing the computational time required for the search process [23].

3.3 RESULTS

This part devoted to the principal results in the classification part, serves as a comprehensive summary that highlights the primary findings and outcomes derived from our research. It acts as a comprehensive overview, presenting the most significant and impactful results obtained throughout the study in a concise and focused manner. The chapter provides a clear account of the discoveries, analyses, and conclusions that significantly contribute to the overall objectives and goals of the thesis.

Table 3-4 displays the outcomes of hyperparameter tuning strategies using Grid search, Random search, and Bayesian search for Bagging classifier. You can see that utilizing Bayesian search yields the greatest score "0.592" when compared to grid search "0.567" and Random search "0.486", but it also yields the most iterations to achieve the ideal parameters and thus takes the most time "106.7 s". Grid search takes relatively little time ("2.218 s"), thus you can select it as an optimal result.

3.3.1 Description of the innovated results for the final classification for 94 patients:

Table 3-5 specifically highlights the best feature selection and feature extraction approaches for 94 patients across the 10 models. In all models, RF Importance and Extra Tree Importance outperform other techniques. Only in

AdaBoost, RFECV also gives the same results with these two techniques. BC, GB, RF, AdaBoost, and XGB give an accuracy higher than 0.8. The highest results were with BC with 0.894 as an accuracy, 0.822 as an AUC, the overall precision of 0.91, an overall recall of 0.9, and an overall F1 score of 0.895.

Table 3-4: Techniques of hyperparameter tuning search with respect to the number of iterations, the number of iterations in order to obtain the optimal hyperparameters, the score and the time elapsed in second.

	Number of iterations	Iteration number of optimal hyperparameters	Score	Time Elapsed (s)
Grid Search	600	1	0.567	2.218
Random Search	100	1	0.486	5.115
Bayesian Search	50	27	0.592	106.7

Experimental results (Figure 3-7) showed that Bagging Classifier Algorithm, with Extra Tree importance feature selection technique with 0.04 as a threshold, outperformed all other algorithms with all different feature selection and feature extraction techniques in terms of AUC-ROC (0.822) with an accuracy 89.4 %, a sensitivity of 100%, and a specificity of 80% (red point).

Table 3-5: The precision, recall. F1 score and the accuracy for all models for 94 patients.

Model	Best FS or FE techniques	Accuracy	AUC	Precision		Recall		F1score	
				True	False	True	False	True	False
LR	Extra Tree Importance (0.04,0.035)	0.736	0.788	0.64	1	1	0.5	0.78	0.67
SVM	Extra Tree Importance (0.035)	0.789	0.711	0.69	1	0	0.6	0.82	0.75
KNN	RF Importance (0.04)	0.789	0.81	0.62	0.83	0.89	0.5	0.73	0.62
BC	Extra Tree Importance (0.04)	0.894	0.822	0.82	1	1	0.8	0.9	0.89
GB	Extra Tree Importance (0.04)	0.842	0.799	0.75	1	1	0.7	0.86	0.82
RF	RF Importance (0.04, 0.05)	0.842	0.794	0.75	1	1	0.7	0.86	0.82
NB	RF Importance (0.04,0.05)/ Extra Tree Importance (0.045,0.05)	0.736	0.833	0.64	1	1	0.5	0.78	0.67
MLP	RF Importance (0.05)	0.631	0.655	0.56	1	1	0.3	0.72	0.46
AdaBoost	RF Importance (0.03), Extra Tree Importance (0.035), RFECV (RF Estimator)	0.842	0.844	0.75	1	1	0.7	0.86	0.82
XGB	Extra Tree Importance (0.05)	0.842	0.827	0.75	1	1	0.7	0.86	0.82

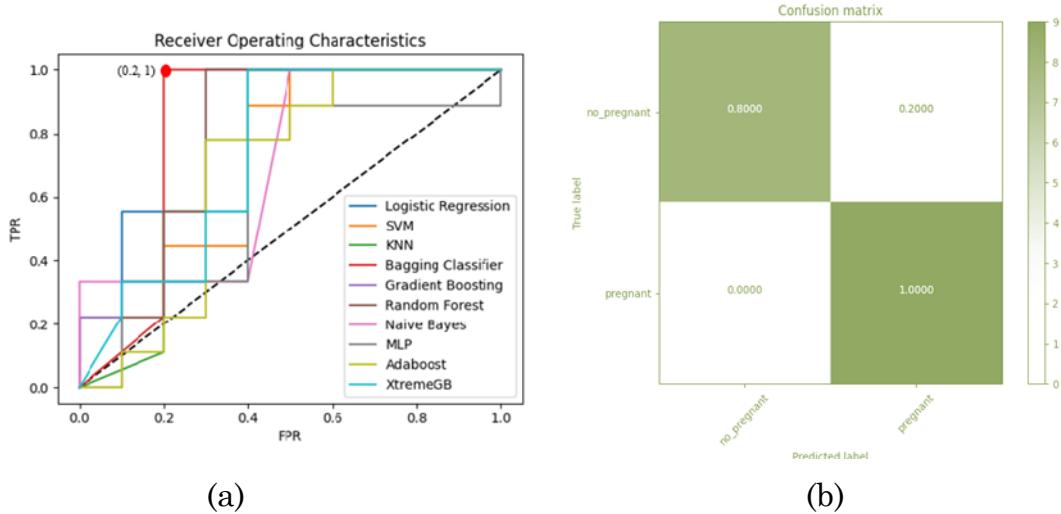


Figure 3-7 : (a) The ROC curve of all models, (b): The confusion matrix of the BC model (94 patients).

Figure 3-8 and 3-9, represent histograms that compare the accuracy and AUC for three files including 94 patients. The bars in dark grey color (in the below of both figures) represent the accuracy and AUC of file 1 that contains Doppler parameters. The bars in light grey color in the below both figures represent the accuracy and AUC of sheet 1 that contains Clinical parameters. The bars in middle grey color in the below both figures represent the accuracy and AUC of file 1 that contains the combination between Doppler and Clinical parameters. For all models, Doppler gives the lowest values of accuracy and AUC. It is very clear that when Doppler and clinical parameters are used together, the models give the highest performance with the highest accuracy and AUC. For the best classifier (BC), taking clinical and Doppler data into accounts leads to a 10-point gain in accuracy and a 4-point gain in AUC compared with clinical data alone. With MLP, the accuracy of clinical parameters was the same as with all parameters (Doppler and clinical) with a value of 0.526 but with higher AUC for the sheet combined Doppler and Clinical. The results demonstrated that combining Doppler and clinical parameters yielded a greater performance than with simply Doppler or clinical parameters. Therefore, the results of our study could be used to aid clinicians to assess the efficacy of ICSI treatment.

3.3.2 Description of the innovated results for the final classification for 572 patients:

Table 3-6 provides a comprehensive overview of the best feature selection and feature extraction methods for 572 patients across the 10 models. The analysis reveals that, with this sample size, no single feature engineering technique emerges as the definitive best choice. However, LR with ANOVA feature selection stands out as the top performer among the feature selection methods. On the other hand, when combining SVM with RFECV using RF estimator, it achieves the highest accuracy of 0.678. This combination proves to be particularly effective in maximizing predictive performance. When employing PCA feature extraction techniques, KNN, RF, and

AdaBoost models deliver the highest performance. Specifically, KNN and RF achieve an accuracy of 0.721, while AdaBoost achieves an accuracy of 0.704. Moreover, GB, NB, MLP, and XGB models demonstrate the highest accuracy when combined with Extra Tree Importance with 0.695, 0.669, 0.686 and 0.704 respectively. Finally, BC showcases superior performance among the considered models when utilizing RF Importance feature selection techniques with 0.05 as a threshold. It achieves an accuracy of 0.739, an AUC of 0.782, an overall precision of 0.8, an overall recall of 0.63, and an overall F1 score of 0.71. These results highlight the exceptional performance of BC compared to the other models. Table 3-6 presents also a comprehensive analysis of the best feature selection and feature extraction methods across various models. It demonstrates that different techniques excel in different scenarios, and BC emerges as the top performer, showcasing excellent accuracy and overall performance.

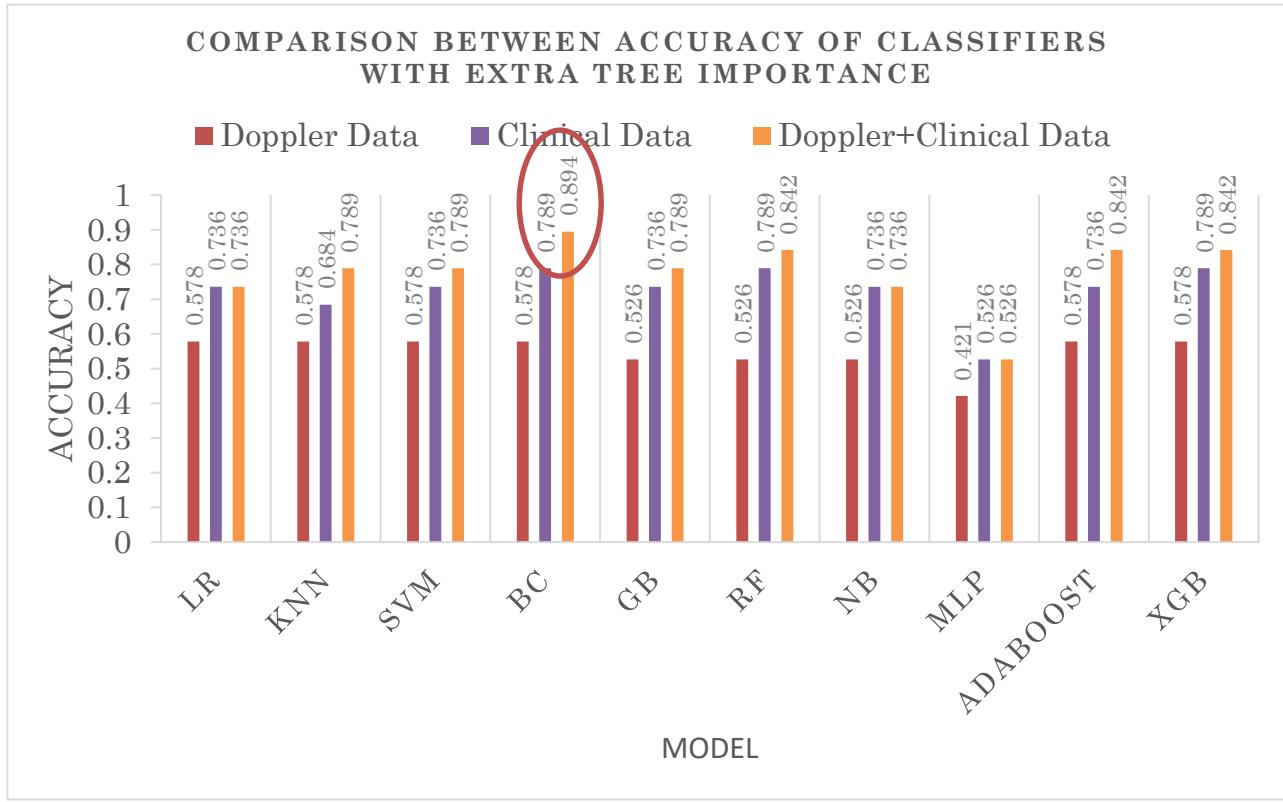


Figure 3–8: Comparison of the accuracy between Doppler, clinical and the combination of Doppler and clinical (94 patients).

COMPARISON BETWEEN AUC OF CLASSIFIERS WITH EXTRA TREE IMPORTANCE

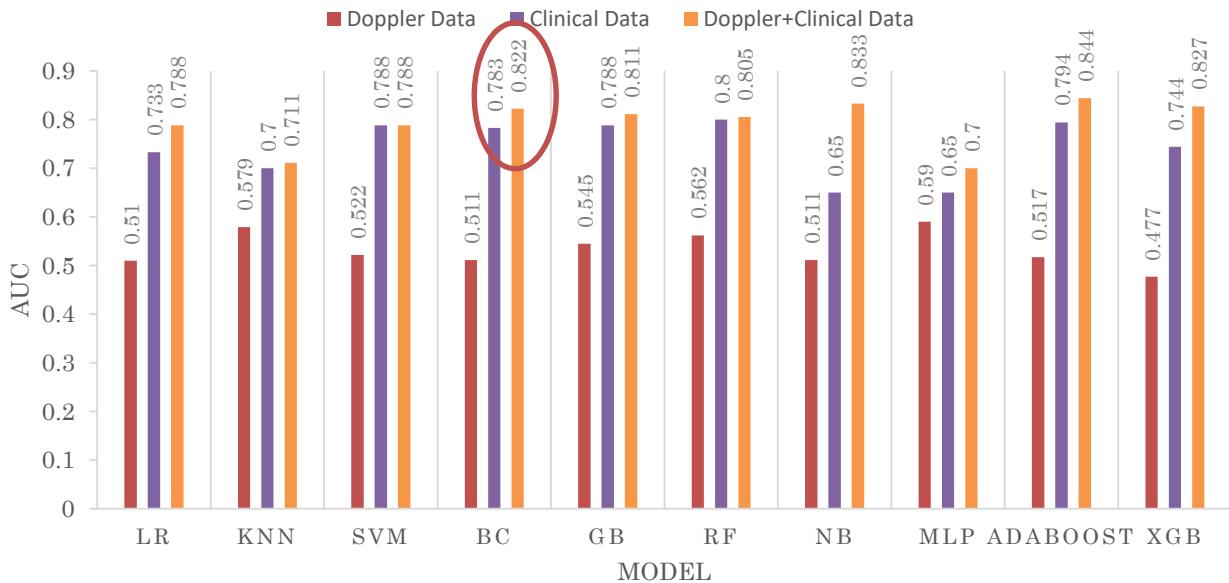


Figure 3–9: Comparison of the AUC between Doppler, clinical and the combination of Doppler and Clinical (94 patients).

Table 3-6: Representation of the precision, recall, F1 score and the accuracy for all models for 572 patients.

Model	Best FS or FE techniques	Accuracy	AUC	Precision		Recall		F1score	
				True	False	True	False	True	False
LR	With ANOVA (k=8)	0.695	0.728	0.72	0.68	0.63	0.76	0.67	0.72
SVM	With RFECV (RF estimator)	0.678	0.732	0.69	0.67	0.63	0.72	0.66	0.69
KNN	With PCA (components=11)	0.721	0.757	0.76	0.7	0.65	0.79	0.7	0.74
BC	With RF Imp (threshold=0.05)	0.739	0.782	0.8	0.7	0.63	0.84	0.71	0.77
GB	With ET Imp (threshold=0.044)	0.695	0.76	0.75	0.66	0.58	0.81	0.65	0.73
RF	With PCA (components=11)	0.721	0.75	0.75	0.7	0.67	0.78	0.7	0.74
NB	With ET (threshold=0.044)	0.669	0.725	0.68	0.66	0.63	0.71	0.65	0.68
MLP	With ET (threshold=0.044)	0.686	0.705	0.7	0.68	0.65	0.72	0.7	0.67
AdaBoost	With PCA (components=11)	0.704	0.717	0.76	0.67	0.6	0.81	0.67	0.73
XGB	With ET (threshold=0.044)	0.704	0.752	0.76	0.67	0.6	0.81	0.67	0.73

Figure 3-10 shows the analysis of the ROC curve results leads to the conclusion that Bagging classifier shows superior classification accuracy compared to other models. This judgment is based on the higher AUC value achieved by Bagging classifier (78.2%), which signifies its stronger discriminatory ability and overall predictive performance when compared to other models. Additionally, the curve shape of BC being closer to the top-left corner provides further evidence of its superior performance. Furthermore, examining the confusion matrix illustrated in Figure 3-10 (b), we observe that the Bagging classifier achieves a sensitivity of 83% and a specificity of 61%, resulting in an overall accuracy rate of 74%. In summary, when evaluating these two models using the ROC curve, bagging classifier exhibits better

classification performance by demonstrating higher accuracy in effectively distinguishing between positive and negative instances.

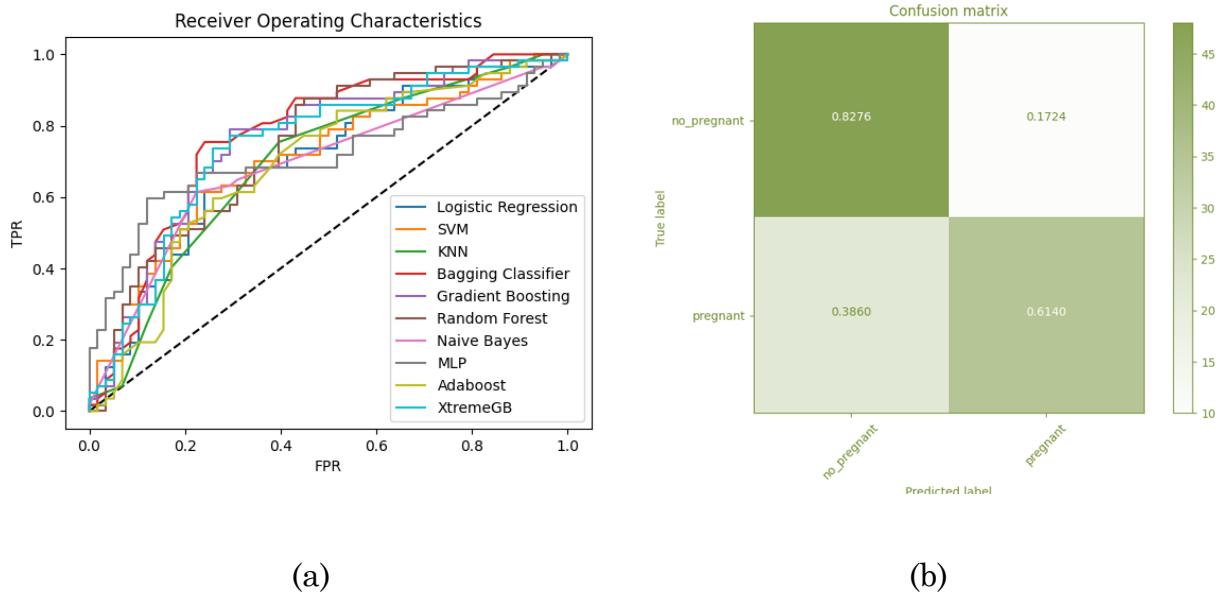


Figure 3–10: (a) The ROC curve of the 10 classifiers, (b) The confusion matrix for the highest model (572 patients).

By analyzing the boxplot illustrating the cross-validation performance, you can gain insights into the distribution, variability, and relative performance of the evaluated models. Comparing the medians, box lengths, and presence of outliers allows for comparisons between the models. Models with higher median performance and shorter box lengths are generally preferred as they demonstrate greater consistency and reliability. In Figure 3-11, it can be observed that LR exhibits the highest median performance. Additionally, LR, BC, RF, and NB show the shortest box lengths and lack any outliers. These characteristics indicate strong model performance. Consequently, these four models exhibit high performance during the validation phase. The RF Importance method selected thirteen variables with varying rankings to build the model. Table 3-7 and Figure 3-12 provide insights into the selected variables, highlighting their importance in predicting the outcome of ICSI. Notably, variables such as myometrium VFI, volume of the endometrium, myometrium mean grey, myometrium FI, myometrium VI, endometrium mean grey, and endometrium FI were identified as significant factors in the prediction. Furthermore, it is worth noting that Doppler parameters, particularly those associated with the myometrium, ranked among the top features in terms of their importance for prediction.

Figures 3-13 and 3-14 represent histograms that compare the accuracy and AUC for three files including 572 patients. The bars in red color (in the below of both figures) represent the accuracy and AUC of file 1 that contains Doppler parameters. The bars in violet color in the below both figures represent the accuracy and AUC of sheet 1 that contains Clinical parameters. The bars in orange color in the below both figures represent the accuracy and AUC of file 1 that contains the combination between Doppler and Clinical parameters. For all models, Doppler gives the lowest values of accuracy and AUC. It is very clear that when Doppler and clinical parameters are used together, the models give the highest performance with the highest accuracy and AUC. Using LR, NB, AdaBoost, and XGB, the accuracy achieved when considering only clinical parameters was the same as when including both Doppler and clinical parameters, with a value of 0.669. However, the combined Doppler and clinical parameters showed higher AUC compared to using only clinical parameters, with a value of 0.731 instead of 0.712 in SVM. It yielded slightly higher accuracy, with 0.686 for clinical parameters and 0.678 for the combined clinical and Doppler parameters. Similarly, the AUC values were also greater for the combined file, indicating improved performance with the inclusion of Doppler parameters. These results indicate that combining Doppler and clinical parameters resulted in superior performance compared to using either Doppler or clinical parameters alone. Consequently, our study's findings can potentially support clinicians in assessing the effectiveness of ICSI treatment.

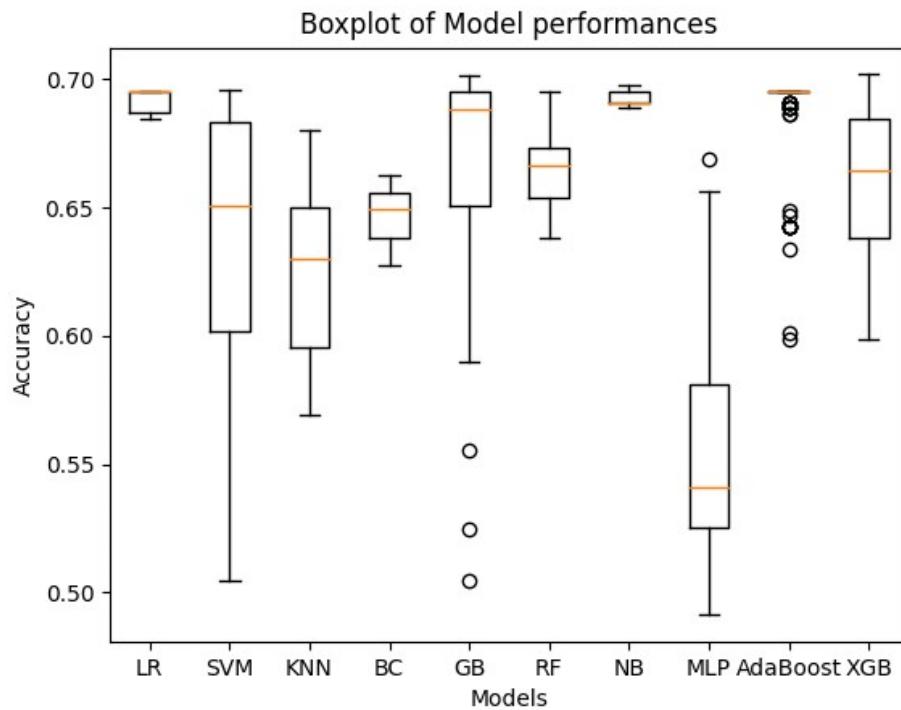


Figure 3–11:The Boxplot of Model performance.

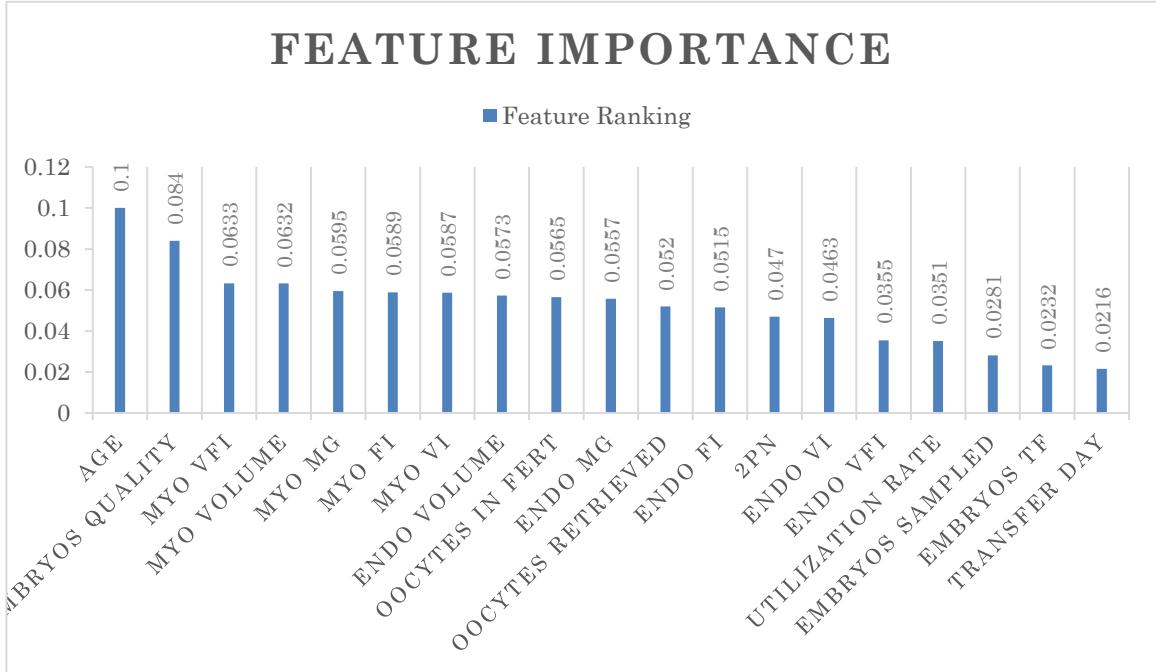


Figure 3–12: Feature ranking of the RF Importance feature selection technique.

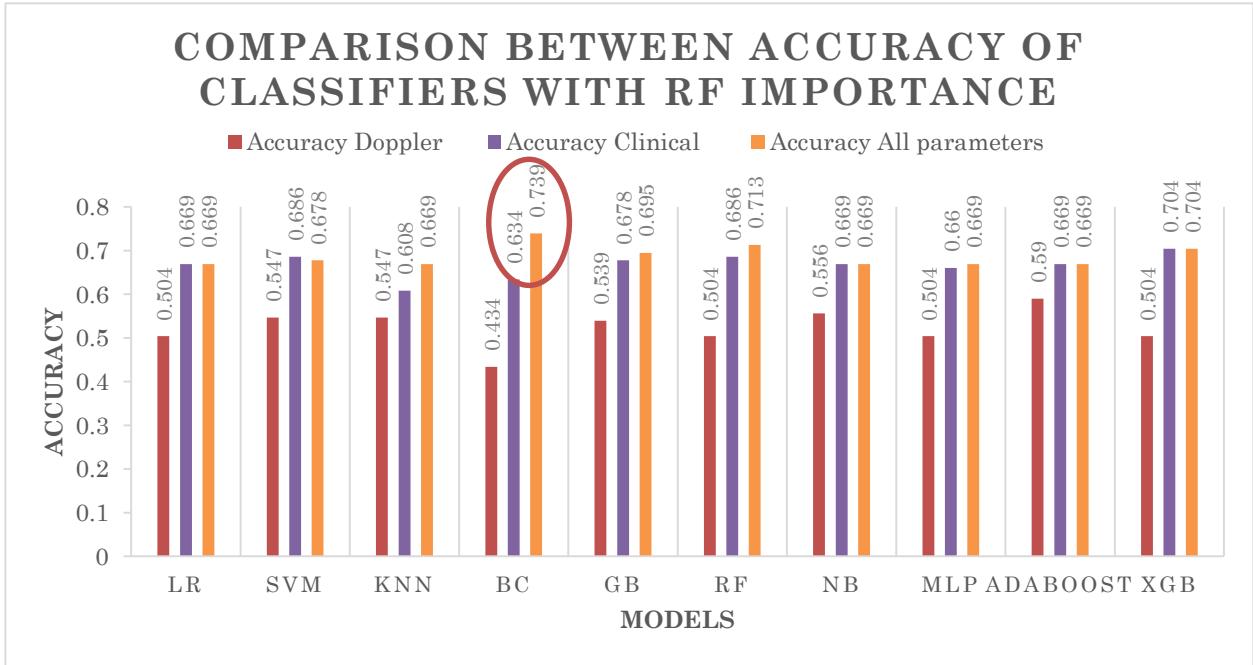


Figure 3–13: Accuracy of the classifiers on three files for 572 patients. (Doppler, clinical and all parameters)

COMPARISON BETWEEN AUC OF CLASSIFIERS WITH RF IMPORTANCE

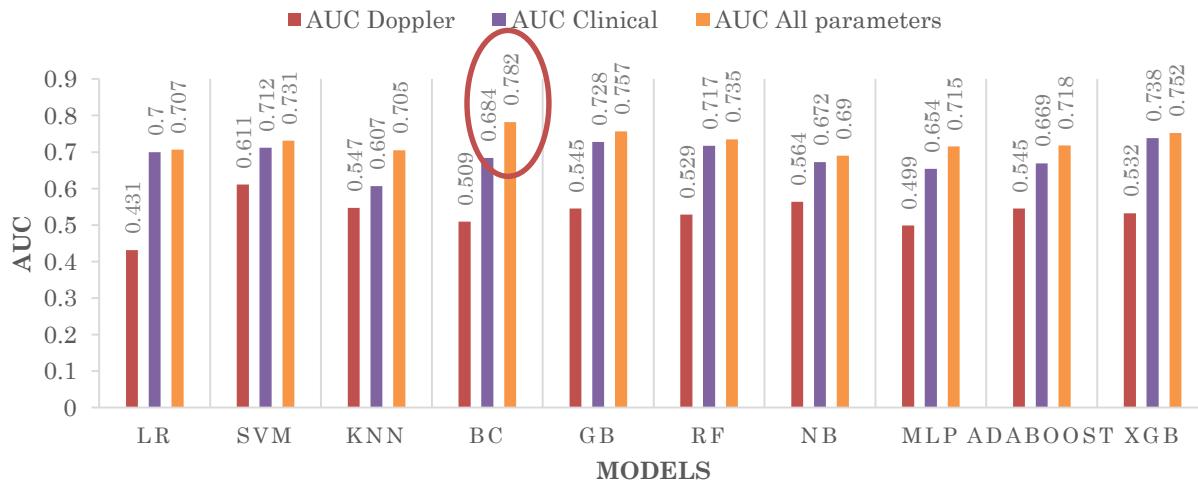


Figure 3–14: AUC of the classifiers on three files for 572 patients. (Doppler, clinical and all parameters)

3.4 DISCUSSION

This study showed the importance of the classifier and the feature selection technique that is based on the Tree learning algorithm. It showed that when the Extra Tree importance feature selection strategy was used to identify the significant features, the outcome of ICSI using the Bagging classifier can be predicted. This model outperformed all other tree-based learning algorithms and provided the best results for 94 patients. Additionally, when combined with RF importance feature selection techniques, the bagging classifier based on a tree learning algorithm surpassed the performance of all other models for the 572 patients. The significance of decision tree-based algorithms and ensemble learning, in comparison to other classifiers, can be attributed to various factors. Ensemble learning is highly valuable in scenarios where a single classifier may struggle to capture the intricate patterns within the data. By combining the predictions of multiple classifiers, ensemble methods enhance overall performance. They achieve better generalization and higher accuracy, mitigating the risk of overfitting and enhancing model robustness. On the other hand, decision tree-based algorithms provide interpretability by presenting a transparent decision-making process based on feature splits. This interpretability enables a better understanding of the decision-making mechanism of the model, facilitating the identification of important features and potential relationships among variables.

Ultrasound-based imaging techniques can reveal a variety of physiological and pathological uterine abnormalities. Personalized prediction techniques that have been systematically developed and verified can be utilized to assist infertile patients

[22]. We assessed the prediction of ICSI outcome using simple clinical factors, ultrasound measurement parameters, and a combination of all parameters. All contributing elements must be considered in order to provide convenient advice regarding a given couple's case for pursuing ICSI therapy. One reason contributing to such a low success rate could be the lack of adequate procedures for predicting infertile couples. Researchers attempted to develop the finest decision-making algorithms using data from ICSI patient records. However, Doppler measures affected the results of the prediction of the outcome of ICSI and it should be taken into consideration by the professionals. The results were very satisfactory in accordance with our hypothesis to forecast and improve the success of ICSI using additional features with an accuracy of 89.4% where Doppler parameters enhanced the models of prediction. Because access to the information used in this particular study with this specific dataset was limited due to confidentiality and other ethical concerns, making direct comparisons between our work and other studies can be challenging. Out of the mentioned studies, only one considered the use of endometrial and subendometrial blood flow as measures to establish a correlation with IVF. However, in the study by Zhang et al. [22], endometrial mean grey and myometrium receptivity was not specifically addressed. In this particular study, the authors combined clinical and Doppler parameters, such as flow index (FI), resistivity index (RI), and pulsatility index (PI), for the endometrium to predict the outcomes in embryo transfer cycles. The area under the curve (AUC) obtained during the training procedure was 0.698.

Indeed, by utilizing a Bagging classifier and a reduced feature set consisting of 19 IVF attributes selected through the Extra tree feature selection technique, we achieved a prediction accuracy of 89.4% and an AUC of 82.2% for a dataset of 94 patients. Furthermore, when expanding to a larger dataset of 572 patients, we obtained an accuracy of 74% and an AUC of 78.2% using a Bagging classifier with RF feature selection techniques. There is a noticeable decline in performance when comparing the results of 94 patients to those of 572 patients. This decline can be attributed to that with a very small number of patients, the classifier finds it easy to classify the features between pregnant and non-pregnant. However, as we increase the amount of data with more variations in features, it becomes more challenging for the classifiers to classify accurately. When we further increase the data, it will improve the performance since the classifier will have more training examples to learn from and improve its classification ability. The outcome of ICSI is notably influenced by Doppler parameters of the endometrium and myometrium, with a particular emphasis on the myometrium parameters. These Doppler parameters combined with clinical data play a significant role in determining the success of the ICSI procedure since they permit a 10-point gain compared with clinical data alone. The analysis of the selected variables from RF Importance feature selection technique in predicting the outcome of ICSI highlights the importance of various variables, including myometrium VFI, volume of the endometrium, myometrium mean grey, myometrium FI, myometrium VI, endometrium mean grey, and endometrium FI, in predicting the outcome of ICSI. Notably, Doppler parameters, particularly those related to the myometrium, emerged as influential features for prediction, ranking among the top variables in terms of their importance. The significance of myometrium and endometrium mean grey values lies in their capacity to provide

valuable insights into the characteristics and health of the uterus tissue. These variables play a crucial role in predicting the outcome of ICSI (intracytoplasmic sperm injection) and are considered relevant factors in this regard. The myometrium refers to the muscular layer of the uterus, while the endometrium is the inner lining of the uterus. The mean grey values associated with these tissues can indicate their composition, density, and overall condition. Deviations from the normal range of these values may suggest the presence of abnormalities or health issues that can impact the success of ICSI.

The inclusion of myometrium and endometrium mean grey values as important factors in the prediction model implies that alterations in these variables may be linked to the likelihood of a favorable outcome in ICSI. The knowledge gained will help the medical practitioners to design treatment plans in order to increase the success rate for each infertile couple.

Consequently, the 74% success rate achieved with SPIRL v2 compared with the 65% achieved at the Al Hadi center by the doctor without AI, leads us to believe that our strategy is paying off for assisting doctors in predicting ICSI treatment outcomes. The advantage of our study was the nature of the retrospective data collection to develop the prediction model, particularly the combination of Doppler measurements and clinical features. Doppler characteristics, particularly those linked to myometrium blood flow, and endometrium and myometrium mean grey were new parameters such as Myometrium VI, FI, VFI and mean grey as well as Endometrium VI, FI, VFI and mean grey used to build a prediction model based on machine learning techniques suited for the population receiving ICSI treatment. Therefore, it is important to increase the number of patients. Augmenting the patient sample size in a study offers numerous advantages. Firstly, a larger sample size enables more robust and dependable statistical analyses. By having a greater number of data points, researchers can derive more precise estimations of the associations between variables, thereby minimizing the potential influence of chance fluctuations and enhancing the overall applicability of the study findings.

Typically, this work conducted thus far relies on parameters measured through manual analysis using VOCAL software, which assesses the endometrium and myometrium. Therefore, overall, SPIRL can be considered a semi-automatic predictive model. However, our next objective, as part of the thesis, is to enhance and transform this model into a fully automatic one by incorporating an automatic segmentation technique. Specifically, our next focus is on automating the segmentation process for the endometrium. This advancement aims to streamline the analysis and improve the efficiency of the model.

Note that this chapter have been the subject of 2 articles, one in an international congress and the other submitted to IEEE Access.

Zeinab Abbas, Sébastien Ménigot, Jamal Charara, Zein Ibrahim, Chadi Fakih, Jean Marc Girault, predicting in vitro fertilization outcome from US measurements using machine learning techniques, ICABME, IEEE, Beyrouth, Liban, Oct 2023.

Zeinab Abbas, Sébastien Ménigot, Jamal Charara, Zein Ibrahim, Chadi Fakih, Jean Marc Girault, Machine learning approach to predict the pregnancy outcome undergoing In-Vitro Fertilization treatment using clinical and Doppler features. submitted to IEEE Access

CHAPTER 4

CLASSIFICATION BASED ON CLINICAL DATA AND DOPPLER WITH AUTOMATIC AND SEMI-AUTOMATIC ULTRASOUND FEATURE EXTRACTION (SPIRL VERSION III)

4.1 INTRODUCTION

With the help of the VOCAL software, it was possible for us to manually extract morphological parameters (B-mode) and vascularization parameters (Doppler mode), this part is described in the “Appendix”. In this chapter, our objective is to investigate the automatic and semi-automatic extraction of morphological parameters, specifically focusing on parameters like endometrial volume. However, due to time constraints and limitations in extracting and working on 3D Power Doppler images, we will not address the extraction of vascularization parameters in Doppler mode.

Image segmentation is a complex and captivating problem within computer vision, particularly in the realm of medical imaging. In the medical field, precise segmentation of medical images plays a pivotal role in providing valuable non-invasive information about the structures within the human body. This process aids radiologists in visualizing and studying anatomy, simulating biological processes, pinpointing pathologies, monitoring disease progression, and evaluating the necessity of treatments such as radiotherapy or surgeries.

Ultrasound images pose additional challenges due to factors like speckle noise, low contrast, a low signal-to-noise ratio (SNR), and artifacts. Moreover, there are notable variations in organ structures (such as the endometrium or breast) among patients, making it difficult to apply generalized anatomical knowledge. To overcome these challenges, researchers have developed and documented numerous segmentation techniques over the years. However, there is no one-size-fits-all approach that universally applies to all applications. Instead, a range of methods with varying levels of accuracy, speed, and complexity have been explored for each practical CAD problem.

In the “Appendix”, It was noted that the use of VOCAL software enables accurate endometrium segmentation through manual selection of six slices. However, this manual process is time-consuming. The VOCAL software allows us to extract various endometrial characteristics, including endometrial volume and mean grey value. Therefore, it is crucial to explore an automated segmentation method that can achieve high accuracy and effectively extract the desired characteristics while minimizing the time required for the process[45]. Several research articles and papers have been published on the topic of image segmentation in various journals. However, it is worth noting that none of these publications specifically concentrate

on the segmentation of the endometrium. Our project primarily aims to automate the segmentation of the endometrium (the inner lining of the uterus) to facilitate the classification and prediction of treatment outcomes in ICSI (Intra-Cytoplasmic Sperm Injection). Consequently, the aim is to achieve a completely automated prediction of the outcome.

Our first goal in this study is to extract relevant characteristics from 2D images that have been previously identified as influencing the outcome of IVF. These characteristics include endometrial volume, mean grey value of the endometrium, and the thickness of the endometrium, which have been extensively discussed in the existing literature.

In our segmentation work, we adhere to a comprehensive framework in Figure 4-1 that involves the following general steps: Firstly, 2D ultrasound images are gathered, followed by the application of diverse preprocessing techniques. Once we conduct an analysis of the endometrium criteria, we can choose between two segmentation strategies: fully automated or semi-automatic segmentation. A sequence of post-processing steps is then executed to obtain the final segmentation result. Lastly, we assess the performance of the segmentation through evaluation.

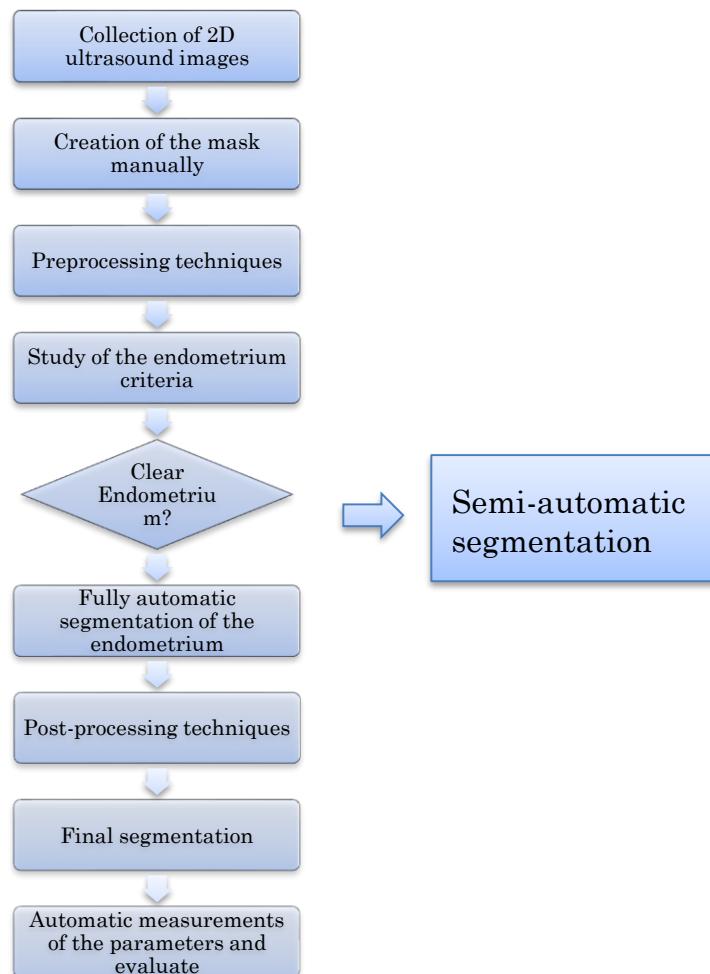


Figure 4-1: A Flowchart for the segmentation of the endometrium.

4.2 STATE OF THE ART

As previously mentioned, the IVF specialist utilizes uterine ultrasound and the VOCAL software to extract qualitative and quantitative information from the endometrium. This extraction process can be time-consuming, and an automated extraction method would be of great assistance, especially for aid in diagnosis and enhance objectivity and precision in assessments. Ultrasound imaging poses distinctive obstacles, including degraded image quality due to noise and artifacts, as well as heavy reliance on the experience of operators or diagnosticians. [46]. Advanced transvaginal ultrasound (TVUS) with high resolution and frequency is utilized to enhance the evaluation of endometrial receptivity in assisted reproductive treatment (ART) cycles. Various ultrasound measurements such as endometrial thickness, endometrial pattern, endometrial volume, Doppler analysis of uterine arteries, and assessment of endometrial blood flow are employed to evaluate the readiness of the endometrium for in vitro fertilization (IVF) procedures [47] [48] [49]. There are existing tools for analyzing images, such as the VOCAL Software, which can be used to estimate the volume of organs. Although the manual tracing and segmentation process in VOCAL software ensures accuracy, it is crucial to acknowledge that it can be a time-consuming task. The requirement to manually outline the boundaries of the endometrium on every ultrasound image demands careful attention and can consume a substantial amount of time, particularly when dealing with multiple images or conducting volumetric analysis on a sizable dataset. Adequate time and resources should be allocated when utilizing VOCAL software for precise measurements, as the manual approach can be more time-intensive in comparison to automated or semi-automated methods [50].

Segmentation is a critical task in medical image analysis, and many medical imaging tasks employ similar segmentation approaches. However, when compared to other imaging modalities the segmentation of endometrium images poses significant challenges. This is primarily due to two reasons:

- 1- Ultrasound images have inherently low quality due to factors such as speckle noise, low contrast, low signal-to-noise ratio (SNR), and the presence of artifacts. These characteristics make it particularly challenging to accurately segment the endometrium in ultrasound images.
- 2- Endometrium structures exhibit substantial variations among patients, making it difficult to rely solely on anatomical knowledge or pre-existing models for segmentation. The diverse and patient-specific nature of these features complicates the task of endometrium image segmentation.

In summary, the unique combination of poor image quality and patient-specific variations in endometrium structures makes the segmentation of endometrium images particularly challenging compared to other medical imaging modalities.

The available approaches for segmentation can be divided into two groups based on the level of human involvement in the process: semi-automated and fully automated methods. Semi-automated methods require the radiologist to provide input, such as defining a region of interest (ROI) that includes the lesion, specifying a seed within the lesion, or providing an initial boundary. On the other hand, fully

automated segmentation approaches do not require any user intervention. Various segmentation techniques have been employed for different organs in both semi-automated and fully automated approaches. When it comes to breast ultrasound imaging and segmentation, it poses a similar challenge to our images. In this study [51], they categorize breast cancer segmentation approaches into six main categories: graph-based approaches, deformable models, learning-based approaches, thresholding, region growing, and watershed. This paper serves as a review focusing on the distribution of automatic approaches for breast ultrasound image segmentation. The highest number of publications in this field utilize deformable models as the primary automatic segmentation technique. In the Appendix, "Table 1" provides a description of several studies that employ various segmentation techniques.

In the field of medical imaging, various techniques have been developed for the segmentation of different structures and organs. [52] applied a three-dimensional (3-D) active contour model to segment breast tumors in a 3-D ultrasonic data file, achieving a segmentation match rate of 95%. [53] presented a level set-based method for segmenting sparse freehand ultrasound data of the ovary. Their method utilized a speed function composed of a surface reconstruction term, a regularization term, and an image term. It required a reliable initialization for accurate segmentation. Learning based algorithms use training data to uncover underlying patterns, built models and make predictions based on the best fit model. Transformation of biomedical big data into valuable knowledge has been one of the most important challenges in bioinformatics. In 2010, [54] proposed a full-automatic learning-based approach for breast ultrasound (BUS) image segmentation. The approach consisted of two steps: ROI generation and ROI classification. The first step aimed to produce dependable regions of interest (ROIs) that closely covered the actual tumor regions, while the second step employed a robust feature extraction and classification strategy for accurate tumor classification. Then, [55] introduced a novel Chan-Vese model enhanced with neighborhood information for ultrasonic image segmentation. This model effectively captured regional details by utilizing neighborhood information and computing similarity to guide the contour towards the target. Experimental results showed an accuracy of 94.2% for this method. Deep learning has evolved rapidly since the early 2000s. It is a branch of machine learning, and has recently emerged based on big data, the power of parallel and distributed computing and sophisticated algorithm [56]. Recent applications of deep learning in medical US analysis have involved various tasks, such as traditional diagnosis tasks including classification, segmentation, detection, registration, biometric measurements and quality assessment, as well as emerging tasks including image-guided interventions of therapy. Of these, classification, segmentation and detection are the three most basic tasks [57]. There are widely applied to different anatomical structures (organ or body location) in medical US analysis, such as lymph node [58], thyroid nodule [59], urinary bladder [60], cardiac [61], kidney [62], breast, uterine [63], endometrium [64], fetus [65], placenta [66] and many more. [58] demonstrated the superiority of the U-net architecture, a deep learning-based approach for lymph node segmentation in ultrasound images. This method outperformed other state-of-the-art techniques when evaluated on a dataset of ultrasound images containing normal and diseased lymph nodes. Than [61] achieved automatic segmentation of the left ventricle on a

large ultrasound dataset using a deep convolutional neural network (CNN). The segmentation achieved a dice similarity coefficient of 0.86. Also, [67] proposed a fully automatic segmentation method for BUS images. The method was evaluated on a database with benign and malignant cases, and it exhibited higher accuracy and robustness compared to a recently published fully automatic method, with an APR (Area under the Precision-Recall curve) of 99.39%. Transfer learning has become integral to medical imaging, and [68] highlighted its effectiveness. In a study by [69], a segmentation network based on VGG16-UNet was developed for ultrasound artery scans, achieving good segmentation results with an average Jaccard index of 0.798. [70] obtained a 3D super-resolution reconstruction of the uterus using a tailored algorithm. Gabor filters were employed to enhance the discriminative power of statistical and GLCM features for placenta and vasculature segmentation, achieving Dice coefficients of 0.82 ± 0.02 and 0.81 ± 0.08 , respectively. [71] developed a fully automatic system for breast tumor detection, segmentation, and classification. The system utilized texture and location characteristics for ROI generation and employed a modified Ncut algorithm for tumor segmentation. An unsupervised clustering algorithm based on morphological and texture features differentiated between malignant and benign tumors, achieving an accuracy of 93.8%. [72] proposed an ovarian follicles segmentation network that incorporated edge information. By adding an edge detection branch and considering edge detection results as one of the network's loss functions, accurate segmentation of ovarian follicles in ultrasound images was achieved, particularly along the edges. Deep learning algorithms demand a substantial amount of memory, primarily due to the extensive data used for training. Deep learning models typically demand powerful GPUs or dedicated hardware like TPUs. Training intricate models can be computationally demanding and time-consuming. **Due to these factors, it is imperative to demonstrate the efficiency of alternative techniques other than deep learning.** Based on this state-of-the-art and for our segmentation work, we follow the following steps: Firstly, we collect 2D ultrasound images, then we apply various preprocessing techniques. Next, we perform segmentation using the Chan-Vese technique and the Split Bregman method with different strategies for automatic and semi-automatic performance. Finally, a series of post-processing steps are carried out to obtain the final segmentation. You can see in next section the description of the methods in details.

4.3 MATERIAL AND METHODS

4.3.1 Data Set collection

I gathered images captured by the "Samsung, Medison" (WS80A) ultrasound machine from January 2018 to May 2018. These images were obtained from patients who underwent ICSI treatments at ALHADI IVF Center in Beirut, Lebanon. We had a total of 172 images at our disposal. I removed 31 images with a gestational sac. Also, 14 had orientations other than sagittal, which required specific adjustments to their format during processing. Therefore, it was necessary to determine the image orientation beforehand in order to enhance the quality of the ultrasound. To maintain consistency, we decided to establish a standard and only considered images with a

sagittal orientation. Consequently, any images with orientations different from sagittal were excluded from the database.

Furthermore, we conducted a visual inspection to confirm if the target we were searching for, specifically the endometrium, was clearly visible in the images. If there were significant variations in the location of the endometrium between individuals or if it was not discernible to the naked eye, it would be impossible for a computer to accurately detect it. Consequently, based on this criterion, we removed 29 images from the database.

As a result of these selection criteria, our image bank now consists of 98 ultrasounds images. In the ongoing process, it is crucial to generate a ground truth mask for each image. To accomplish this, I utilized GNU Image Manipulation Program (GIMP) software and personally selected the endometrium mask for each image as you can see in Figure 4-2. I underwent training with Dr. Fakih to ensure accurate visual selection of the endometrium. Subsequently, a binary conversion was applied to each mask, resulting in a black and white image.

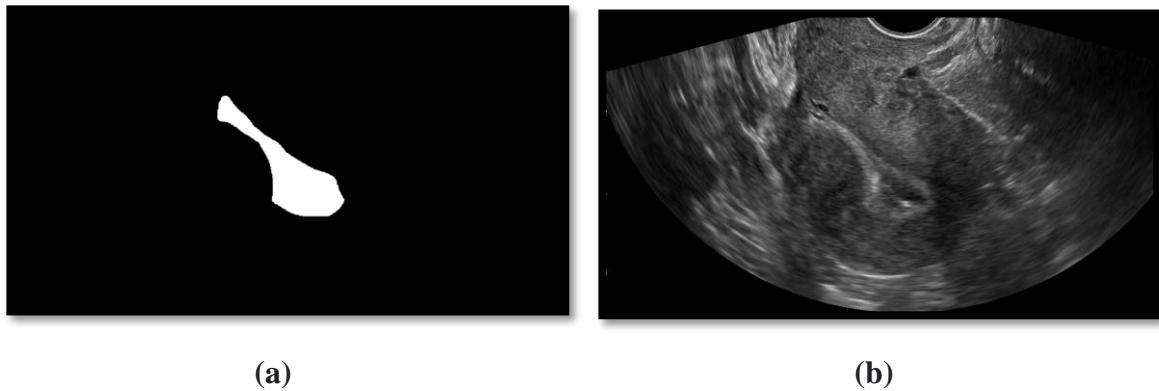


Figure 4-2: (a) Selection of the important part of the image. (b) Manually mask of the endometrium using GIMP software.

4.3.2 Preprocessing techniques

Speckle noise is a common issue in ultrasound imaging due to the coherent illumination and Rayleigh scattering caused by microstructures within the tissue. This type of noise poses a significant challenge in accurately segmenting endometrium images because speckle artifacts are specific to the tissue being imaged and cannot be effectively modeled. Various methods have been used to address speckle noise, including the mean filter, Gaussian low-pass filter, speckle reducing anisotropic diffusion (SRAD), nonlinear coherence diffusion (NCD), sticks filter, bilateral filter, fractional subpixel diffusion, and nonlocal means-based speckle filter. [51].

The study conducted by [73], compared the effectiveness of three filtering techniques: the median filter, Wiener filter, and unsharp filter. Among these, the unsharp filter was found to produce the most favorable results. The study employed a segmentation technique for each filter and compared the segmented images. Notably, the unsharp filter demonstrated the best segmentation performance. Based

on the analysis conducted by [74], another study revealed that among various filters used for ultrasound image denoising, the Gaussian Filter demonstrated superior performance in effectively eliminating Gaussian Noise compared to alternative filters. Additionally, the Mean Filter was identified as a suitable option for removing various types of noise, including Speckle Noise, Salt and Pepper Noise, and Poisson Noise. The study conducted by [75] aimed to optimize the denoising performance by analyzing and comparing different discrete wavelet transform (DWT) and filtering techniques. However, the results indicated that applying DWT and filtering techniques led to information loss and noise characteristics, and did not provide the most effective noise reduction performance. On the other hand, a technique for image fusion called Speckle Reduction Anisotropic Diffusion (SPRAD) was discovered to preserve crucial information from the original image while efficiently eliminating speckle noise. This approach demonstrated superior noise reduction capabilities compared to the DWT and filtering techniques analyzed in the study. [76] conducted a research study aimed to compare the performance of the non-local means (NLM) filter. The study concluded that the NLM filtering algorithm effectively reduces noise while preserving the inherent characteristics of the image.

It is evident from the findings that various filters can effectively reduce the noise present in ultrasound images. Table 4-1 highlights the key filters used in ultrasound imaging, each with its own set of advantages and limitations. Therefore, it is crucial to evaluate different filters to determine the most suitable option for enhancing the clarity of endometrium images, which can often be indistinct.

Table 4-1: The most popular filters to eliminate noise from ultrasound images

Type of filters	Advantages	Disadvantages
Gaussian filter [74]	Effectively reduces noise and smoothed the image.	<ul style="list-style-type: none"> May blur fine details
Median filter [73] [77]	<ul style="list-style-type: none"> Excellent at reducing impulsive noise. It preserves edges while reducing noise Increase the intensity value at a certain point in the image. 	<ul style="list-style-type: none"> May cause blurring in uniform regions.
Mean filter [74]	<ul style="list-style-type: none"> Effective and Simple 	<ul style="list-style-type: none"> Can blur details
Anisotropic diffusion [77] [75]	<ul style="list-style-type: none"> Preserves edges 	<ul style="list-style-type: none"> Requires carefully parameters
Unsharp filter [73] [77]	<ul style="list-style-type: none"> Increase the contrast of the image 	<ul style="list-style-type: none"> Enhancing high-frequency components in an image, this can amplify noise.
Non-Local mean filter	<ul style="list-style-type: none"> Effectively reduces noise 	<ul style="list-style-type: none"> Computationally intensive, especially for large images.

The study encompassed the application of various filters both individually and in combinations, with the objective of determining the most suitable filter or combination of filters for our specific set of images. The primary goal was to evaluate how each filter performed in terms of reducing noise and enhancing the quality of the

endometrium images. Furthermore, the study aimed to identify any potential synergies that could be achieved by combining two or three filters together.

The evaluation of filters involves the use of performance metrics such as PSNR, SNR, and MSE. These metrics provide quantitative measures to assess the effectiveness of speckle noise reduction and overall image quality. It is important to note that higher PSNR and SNR and lower MSE values correspond to better image quality. Table 4-2 displays the results of applying various filters to endometrium ultrasound images. The NLM filter exhibits the highest PSNR value of 44.93 but also has a higher standard deviation of 36.05. On the other hand, the Gaussian and median filters achieve the highest PSNR and SNR values while having the lowest standard deviation for both PSNR and SNR. Additionally, the median and Gaussian filters yield the lowest MSE values of 8.84 and 9.56, respectively. It is important to mention that the specific preprocessing steps required may vary depending on the chosen segmentation technique.

Table 4-2: The values of PSNR, MSE and SNR for each filter applied on our data.

Filter	PSNR		MSE		SNR	
	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
Gaussian filter (Kernel=3)	40.03	1.15	9.56	4.2	22.32	1.23
Mean filter (Kernel=3)	38.16	1.12	15.22	10.8	20.31	1.27
Median filter (Kernel=3)	40.25	1.22	8.84	5.61	22.72	2.72
Unsharp filter (Radius-3)	33.3	135	64.09	178.91	14.2	1.36
SPRAD filter (number of iterations =3)	38.04	2.08	10.72	9.69	22.08	2.78
NLM filter (H=1.15)	44.93	36.05	13.31	1348.63	28.96	33.62

4.3.3 Measurement of all features for the segmentation part.

Initially, we examined the endometrium in the available images with the aim of streamlining the segmentation process and identifying distinct markers or specific characteristics of this tissue.

Subsequently, we performed our analysis using MATLAB software on the images. Here are the findings:

- The average thickness of the endometrium is 12.2 mm.
- The endometrial profile is mainly uniform throughout.
- The average surface area of the endometrium is 3.90E+04 pixels.
- The average grayscale level for each endometrium is 77.2.

A. Endometrial Thickness:

The process of measuring the endometrial thickness involved manual analysis for each image. As Figure 4-3 shows, a line was manually drawn between two

echogenic borders of the endometrium, and the resulting measurement was obtained in pixels. Since the images had a resolution of 300x300 pixels per inch (ppi), it was straightforward to convert the measured thickness from pixels to millimeters.

B. Endometrial Pattern:

The endometrium generally presents itself in two distinct "forms":

1. The "triple-line" appearance, characterized by a hypoechoic endometrium with a hyperechoic line surrounding it.
2. The homogeneous appearance.

Therefore, we had to examine each image in the image database to determine the endometrial profile of each endometrium in the study.

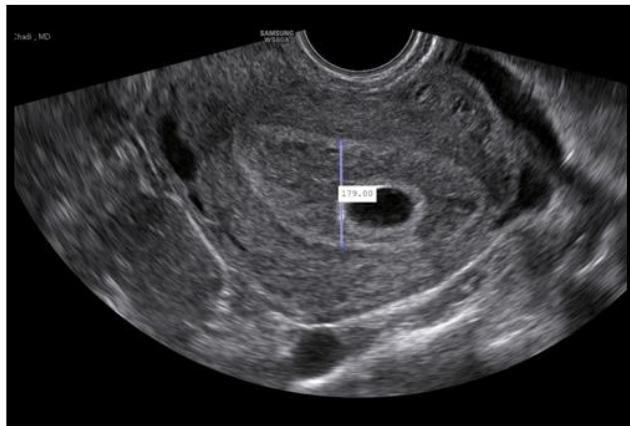


Figure 4–3: Measurement of endometrial thickness.

C. Endometrial surface:

Endometrial surface measurement was a crucial aspect of our study. Since determining the volume was not feasible, we focused on obtaining the surface area of the endometrium. To achieve this, we employed the manual masks makes using GIMP Software. Utilizing the surface calculation functions in MATLAB we were able to easily determine the surface area of each endometrium. In the ultrasound image below (Figure 4-4), you can see the manual mask overlaid on the image. The shaded blue area represents the surface area of our endometrium.



Figure 4–4: Measurement of endometrial surface area.

D. Mean grey of the endometrium:

We also examined the average gray level of the endometrium. Our objective was to identify a distinct value that could be utilized to enhance our images and effectively highlight the endometrial region of interest. The gray level values varied between 135 and 38, but they were primarily concentrated around the range of 80-100. To determine the average gray level, we employed Matlab, which automatically computed the gray levels and generated a histogram, as depicted below (Figure 4-5). By calculating the mean value of these levels, we obtained the average gray level of the endometrium.

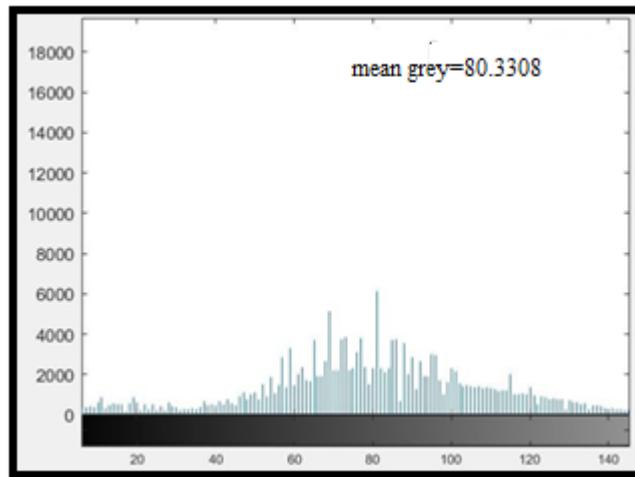


Figure 4–5: Measurement of gray levels of the endometrium.

4.3.4 Segmentation

In recent ultrasound image segmentation literature, various commonly used techniques have emerged. These methods can be classified into different categories, and we present an overview of some methods within each category such as: (1) Graph-

Based approaches, (2) Deformable models, (3) Learning-Based approaches, (4) Thresholding, (5) Region Growing, and (6) Other approaches. However, it is important to note that there is no universally superior technique that can be applied to all types of images. Each technique has its own strengths and limitations and is suitable for specific types of images and applications [78].

Table 4-3 compares the techniques of medical image segmentation with their Characteristics, advantages, and limitations. The selection of a segmentation technique relies on the unique attributes of the ultrasound images and the organ being targeted. It is advantageous to explore diverse techniques and adjust their parameters to optimize outcomes in a specific application. Each method in the table exhibits distinct advantages and limitations. However, considering the irregular shape and intensity of our organ structure, as well as the absence of a uniform criteria applicable to all patients, both deformable models and learning-based approaches have been extensively researched and proven to achieve precise segmentation. Deformable model is a flexible segmentation technique for different shape and size. Learning based approach can classify the organ based on a feature selected from each image.

Also, due to the similarity in challenges faced in segmenting breast ultrasound images and the endometrium, it can be beneficial to draw insights from research conducted on breast segmentation techniques. While there may be limited studies specifically focusing on endometrium segmentation using deep learning techniques, exploring the techniques employed in breast segmentation research can provide valuable guidance and inspiration for developing endometrium segmentation methods. The study [78] discusses the segmentation techniques employed in breast ultrasound (BUS) imaging. Figure 4-6 presents the distribution of BUS segmentation techniques based on their application frequency. The chart indicates that the prevailing techniques primarily rely on deformable models.

The active contour model, also known as Snake is one of the deformable model methods. It has gained popularity as a segmentation method for ultrasound images. However, these ACM methods are sensitive to noise and heavily rely on accurately defining the initial object contours. Traditional active contour models, such as snakes, primarily rely on edge-function and image gradients to detect objects. However, these models can only detect objects with gradient-defined edges, which can lead to the curve passing through boundaries, especially in ultrasound images with high noise levels. To address this issue, level sets approach (LS) integrates statistical information, boundary information, gradient information, phase information, and other factors [81].

The Chan-Vese Model (CVM) is an active contour model that does not require a stopping edge-function based on image gradients. Instead, it utilizes Mumford-Shah segmentation techniques for the stopping process. This model is capable of detecting objects with very smooth boundaries or even discontinuous boundaries, regardless of the presence or absence of a gradient [81]. The CVM is topologically flexible, less sensitive to noise and the initial position of the shape model and can segment structures with weak boundaries. The objective of the CVM is to divide an image into distinct regions by considering the similarity of intensity within each region. This

characteristic makes it well-suited for segmenting ultrasound images, enabling the identification of various anatomical structures or areas of interest within the images. However, one of the key limitations of the CVM is its high computational cost [82]. So, for these reasons you can see that CVM can be a good choice for our nature of the endometrium image segmentation.

Table 4-3: The segmentation techniques with their characteristics, advantages and disadvantages.

Segmentation technique	Characteristics	Advantages	Disadvantages
Graph-Based approaches [79]	<ul style="list-style-type: none"> Graph-based approaches have the capability to derive significant features from the graph structure and node attributes. 	<ul style="list-style-type: none"> An efficient and widely accepted method. Can attain globally optimal results for the energy function. 	<ul style="list-style-type: none"> Constraints in memory and CPU time can pose limitations. Segmenting thin and elongated structures can be challenging; they may struggle with segmenting highly complex or irregular structures in ultrasound images. Difficulty in handling topological changes.
Deformable models [51] [79]	<ul style="list-style-type: none"> It utilizes energy or cost functions to optimize their shape, incorporating image-derived information like gradients or intensity values. 	<ul style="list-style-type: none"> It has the ability to eliminate noise. Ability to adapt to differences in shape, size, and appearance, enabling them to effectively segment intricate structures in ultrasound images. 	<ul style="list-style-type: none"> Computational complexity
Learning Based approaches [80] [51]	<ul style="list-style-type: none"> The decision-making process is based on simulating a learning process. 	<ul style="list-style-type: none"> Robust to noise 	<ul style="list-style-type: none"> Takes time for training Memory consuming
Thresholding [80]	<ul style="list-style-type: none"> Determines threshold values by analyzing the peaks in an image's histogram. 	<ul style="list-style-type: none"> Simple and efficient. 	<ul style="list-style-type: none"> Influenced by the presence of noise and unclear boundaries. Perform well for images that have distinct and well-defined edges.
Region Based approaches [80]	<ul style="list-style-type: none"> Segments an image into regions that exhibit homogeneity. 	<ul style="list-style-type: none"> Performs effectively in situations where defining the similarity criteria is easy. 	<ul style="list-style-type: none"> Time and memory consuming. Challenges arise when dealing with intricate or complicated boundaries.

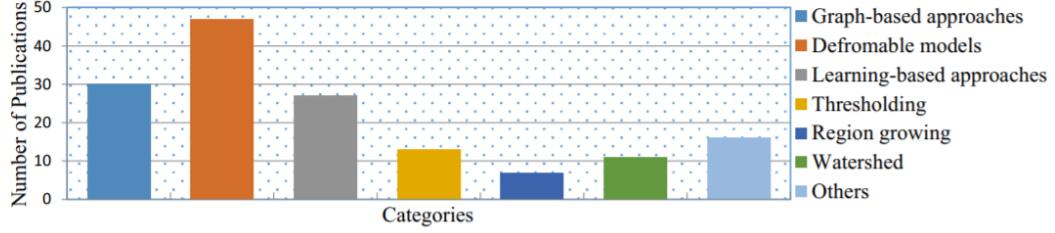


Figure 4–6: Distribution of automatic BUS image segmentation approaches [78].

Different studies talk about the use of split-Bregman to the CVM in order to obtain a segmentation more efficient and a model more robust to noise [83] [84]. Also, the use of the split Bregman method accelerates the minimization process of our model by reducing the computation time and iterative times [85]. This method that caught our attention which enables rapid minimization of l_1 -regularized functionals. Originally, regularization is a concept used in the context of learning. It serves to prevent overfitting by, for example, limiting the model's capacity. Its purpose is to intentionally stop the algorithm when the system shows signs of overfitting. Therefore, training is halted when the error starts to increase. This notion can be visualized in Figure 4-7 where the training set is represented in blue and the validation set in red. At the dashed line, it is advisable to stop the training since the error ceases to decrease and begins to increase again.

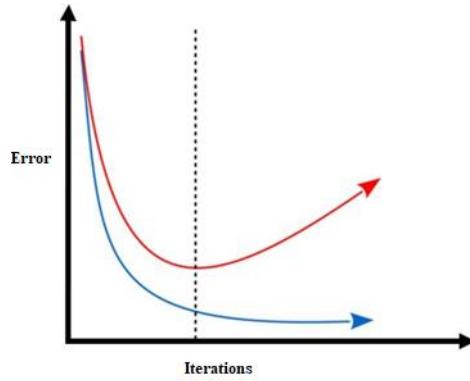


Figure 4–7: Demonstration of the principle of regularization.

Thus, l_1 regularization aims to minimize the sum of absolute differences between real values and predicted values. In our model, it will reduce the number of iterations and consequently decrease the computation time for the contour evolution. The Split-Bregman method is highly efficient and has fewer parameters compared to standard methods. The algorithm requires no initialization and converges rapidly. The l_1 regularization not only speeds up computation times but also makes the final contour independent of the initially chosen level, which generally leads to a good segmentation result [83]. There are primary parameters employed in the Split

Bregman algorithm such as “**max_iter**” that denotes the maximum number of iterations to achieve convergence. This parameter establishes the stopping criterion for the segmentation process (200). Also “**lambda**”, it represents the regularization parameter that influences the balance between data fidelity and the total variation regularization term. Larger values of lambda yield smoother segmentations, while smaller values enable more detailed boundaries (0.01). “**mu**” serves as the augmented Lagrangian parameter, determining the trade-off between the data fidelity term and the total variation regularization term (0.0001).

4.3.5 Segmentation using Chan Vese with Split-Bregman method and selection.

In order to carry out the segmentation, we initiated the process by selecting the specific area of interest. This initial selection is performed manually and plays a vital role in initializing our segmentation procedure. To determine the starting area, we opted for a shape that aligns well with the morphology of the endometrium. Through an examination of their unique characteristics, we determined that a polygon shape is the most appropriate choice, as it effectively encapsulates the majority of endometrium shapes. The actual drawing of the shape is accomplished using MATLAB's "drawrectangle" function. So, the blue rectangle is the first output in Figure 4-8.

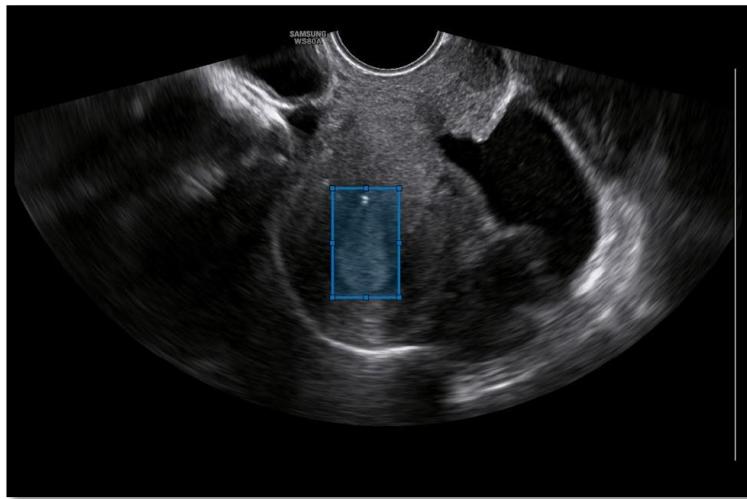


Figure 4-8: Selection of the area of interest.

The second step of our algorithm involves processing the selected area of interest. processing involves the application of a series of image processing techniques. The primary objective of these techniques is to improve the quality of the original image, making it better suited for the subsequent segmentation process. Without this crucial preprocessing phase, the segmentation results would be suboptimal. Therefore, we modify the image's contrast, brightness, and sharpness. Additionally, we employ Median and Gaussian filters, which were selected based on previous section evaluating their effectiveness on our images. Figure 4-9 represents

the original image with the last image after the successive treatments applied to the area of interest.

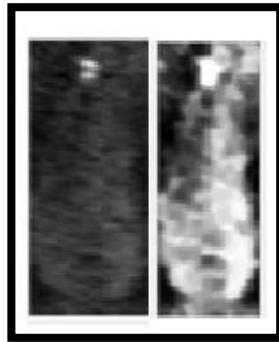


Figure 4–9: Successive treatments applied to the area of interest. The image on the left represents the original image, while the image on the right depicts the final preprocessed image.

After applying image processing techniques, we employ the Split-Bregman method to perform segmentation on our endometrium. The code we utilize for this purpose has been developed by Ander Biguri [86]. As evident from our research sources, this method relies on ℓ_1 regularization, which effectively showcases the attainment of the desired outcome.

Subsequently, we acquire an initial outline of the endometrium contours. However, many other small shapes have been detected. To address this, we will undertake a two-step process to eliminate them. The initial step involves employing the "imerode" function in MATLAB, which facilitates erosion as a means to remove the smaller shapes "Figure 4.10 (a)". The second stage involves the selection and preservation of the largest identified shape. To accomplish this, we utilize the "activecontour" function in MATLAB to extract the contours of our shapes. Subsequently, we assign labels to the connected components of these contours. After labeling, we determine the most frequently occurring label. The figure below illustrates the labeling process, providing a visual representation of the colors associated with each label. Notably, the predominant label is green, distinguishing it from cyan, yellow, or pink "Figure 4.10 (b)" Following that, we remove the labels that appear less frequently, assuming that the largest detected shape corresponds to our endometrium. Once we have correctly identified the shape, we carry out a dilation operation to fill in any remaining voids or openings within the shape (Figure 4-10).

After making these selections and utilizing the described techniques, it is evident that the detected endometrium contour, depicted in green in Figure 4-11, closely resembles the manually generated blue mask overlaid on the image. This similarity allows us to accurately determine the characteristics of the detected shape.

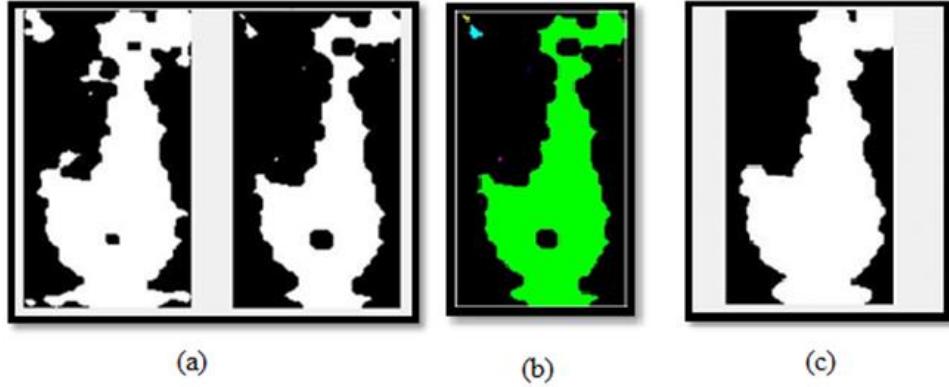


Figure 4–10: (a) Erosion applied to our detection., (b) Labeling of the detected shape, (c) Dilation applied to our detection.



Figure 4–11: Segmentation using the Split-Bregman method with the endometrium selected in green and visualized alongside the manual mask represented in blue.

4.3.6 Segmentation using the Chan-Vese model without any specific selection.

To segment our images without prior selection, we initially process the images through several steps.

Firstly, we resize the ultrasound image as typical ultrasound images contain irrelevant black borders and patient information. We focus on selecting the largest area that solely contains the ultrasound image. Next, we calculate the average gray level of the image, providing clarity information. This enables us to eliminate the black areas from the ultrasound image. We aim to remove any dark spots that could potentially mislead our algorithm. Following this, we apply histogram equalization and a median filter to the image. Additionally, we utilize K-means clustering, a commonly used technique in image segmentation. This algorithm is known as one of the most straightforward and widely used unsupervised methods. It iteratively assigns 'n' datasets to k clusters. By calculating the mean intensity for each cluster,

the algorithm classifies pixels based on their proximity to the closest mean values. The K-means clustering approach aims to minimize the number of clusters and cluster variability. It finds application in medical image segmentation, particularly in MRI images [87]. A Sobel filter is applied, followed by dilation. This initial processing phase enhances the visibility of the endometrium significantly (highlighted intentionally in red in the final image), as depicted in Figure 4-12.

This initial processing allows us to create a composite image from the treated images, which serves as a basis for the second stage of processing. The second stage (Figure 4-13) begins with the removal of gray areas. As seen in the first image, which is the result of combining the previous images, there is a grayish background. We eliminate this background to avoid its influence on subsequent processing techniques. After removing the background, we eliminate the bright structures associated with the image borders, preserving the central areas of the ultrasound image. We conclude the image processing by increasing the brightness to binarize the image while retaining the highlighted patterns.

After completing the processing, we examine the resulting image. In certain cases, this processing can lead to the loss of all ultrasound information, leaving us with only a few white pixels visible. In such instances, we already know that segmentation is not possible. Consequently, we categorize the image as part of the folder for images that cannot be segmented using the automatic algorithm.

If the processing appears suitable, we proceed with executing the algorithm. We utilize the "activecontour" function in MATLAB, similar to the segmentation method that involves prior selection of the endometrium. This enables us to obtain the selection of all the contours present in the image. To organize these shapes, we initially remove any shapes with fewer than 100 pixels since an endometrium is inherently larger. Subsequently, we perform a dilation operation on the identified shapes.

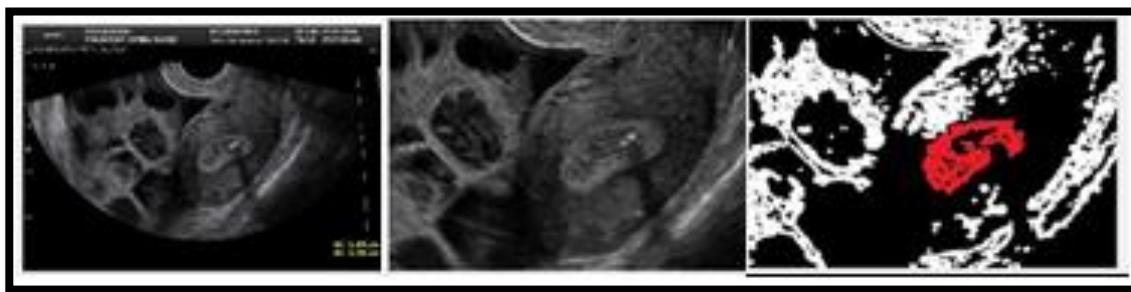


Figure 4–12: First stage of image processing in the context of segmentation using the Chan-Vese model without prior selection. The left image is the initial image, the middle image is the cropped image and the image in the right is the final image after

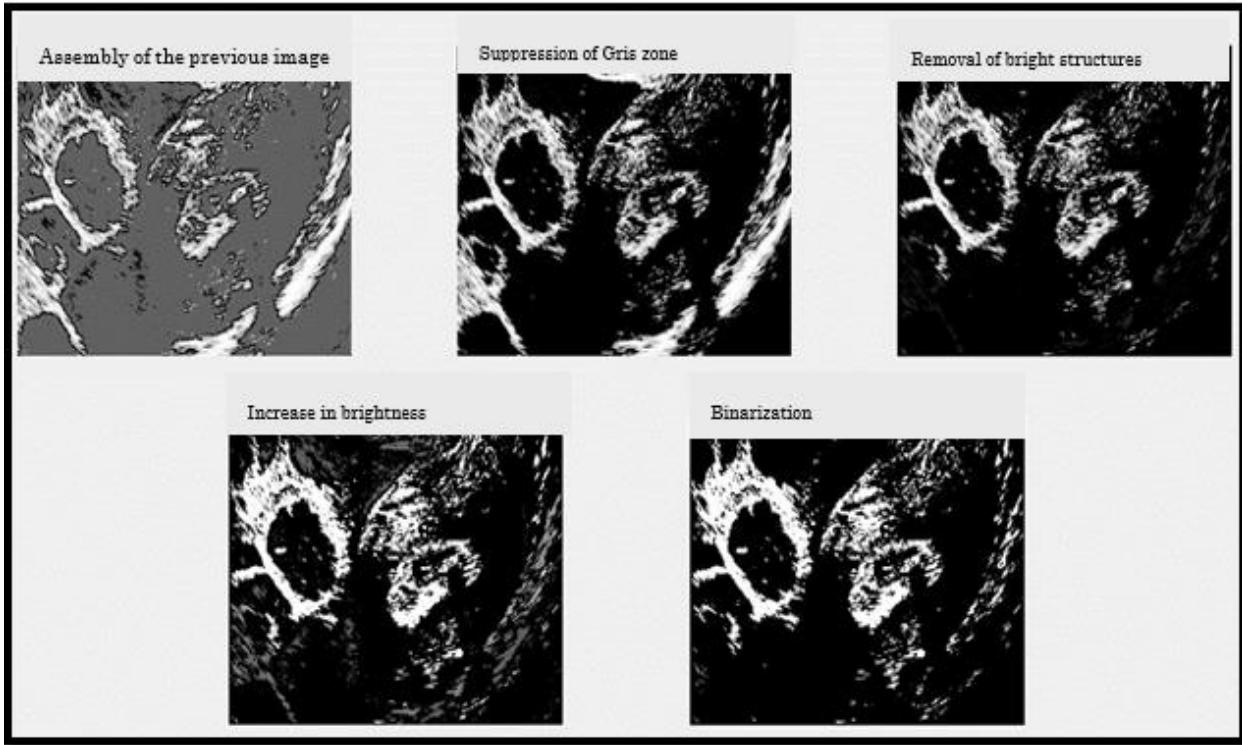


Figure 4–13: Second stage of image processing in the context of segmentation using the Chan-Vese model without prior selection.

To determine which shape corresponds to the endometrium, we start from the center of the image and search for a shape larger than 1000 pixels within a circle of radius 150 pixels, centered at the middle of the image. Given that the endometrium should be distinguishable following the processing, we have observed that it is typically the largest centrally positioned shape that represents the endometrium. To identify this shape, we employ connected component labeling, similar to the previous method. This assists us in detecting the most prevalent label and retaining it. If this shape is smaller than 1000 pixels, we gradually increase the circle's radius until we find a shape exceeding the threshold of 1000 pixels. This threshold was determined through a thorough examination of the database, where we observed that an endometrium typically consists of a minimum of 1000 pixels.

If we detect a shape meeting this criterion, we can select it and remove the other shapes. However, it is crucial to verify whether the circle has intersected the chosen shape.

Let's illustrate this using the image below. As depicted, the centered circle does not encompass the shapes entirely, with some portions being contained within it. Consequently, if the algorithm detects a shape larger than 1000 pixels inside the circle, it will retain only the portion inside and disregard the rest of the shape.

To address this issue, we create an index that records each pixel's association with each shape. Hence, when the algorithm identifies a shape comprising more than 1000 pixels, it will determine the initial shape stored in the index to which these

pixels belong. This allows us to select the complete shape. In case no shape larger than 1000 pixels is found, we opt for the largest shape, assuming it to be the endometrium. This situation arises when the processing has significantly reduced the image information but not to the extent of being classified as unsegmentable. Once the shape is chosen, we perform dilation and hole filling operations.

After this selection, we check if the circularity of the chosen shape is not too low. Indeed, according to our observations in the database, endometrium shapes tend to have a high circularity. Therefore, if any of the selected shapes appear too linear, we go back to the initial segmentation before shape selection and choose the largest shape found as you can see in Figure 4-14.

Once we have found our final shape, we again examine its topology and check for any holes. The example in Figure 4-15 allows us to visualize that the shape is not closed. To obtain a closed shape, we perform the skeletonization of our shape and gradually remove small branches until we obtain the overall contour and two endpoints.



Figure 4-14: Selection of the shape after segmentation.

After identifying the two points, we record their coordinates and determine the distance between each x-coordinate and y-coordinate. This allows us to calculate intermediate points along the line connecting the two endpoints, effectively closing the shape (red line in Figure 4-15). The resulting line can be observed in the leftmost image of the figure provided. By examining the skeleton of the shape and the inclusion of the line, it becomes apparent that the shape is now closed. To complete the process, we apply morphological closing, which fills the entirety of the shape that has been identified.

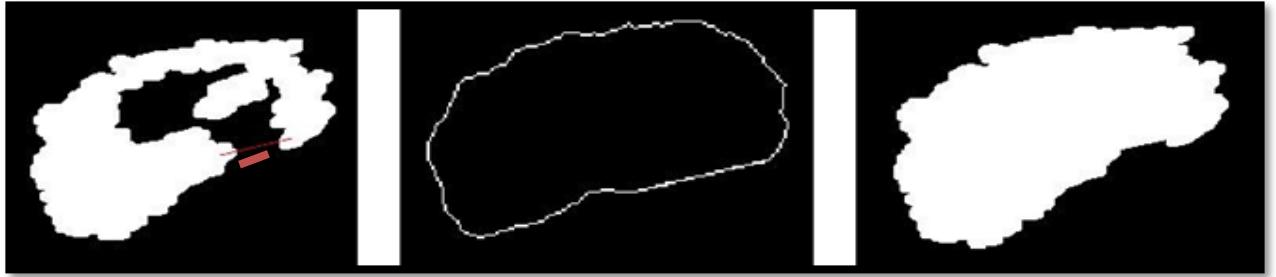


Figure 4–15: Closing the contours of the shape.

As a result of applying this function, we can see that the boundary of the detected endometrium, depicted as the green contour in Figure 4-16, bears a strong resemblance to the manually annotated mask represented in blue on the image.

4.3.7 Segmentation using Chan-Vese with Split-Bregman method and without selection.

The Split-Bregman method is an alternative technique that employs the level-set approach. It offers greater efficiency and time savings compared to the Chan-Vese model, which is why we found it intriguing. To implement this method, we utilized the codes developed by Ander Biguri, which are based on the method described in the article "Geometric Applications of the Split Bregman Method: Segmentation and Surface Reconstruction" by Tom Goldstein, Xavier Bresson, and Stanley Osher [86]. As a result, I won't delve further into the details of this particular aspect of our algorithm.



Figure 4–16: Segmentation with the Chan-Vese model without endometrium selection, visualized in green, compared to the manual mask in blue.

Similar to the Chan-Vese model, this function initiates with preprocessing the original image. We begin by reducing its resolution, akin to the previous function. However, we employ distinct treatments based on our observations during the development of our methods. We have noticed that preprocessing plays a crucial role in the segmentation's efficacy, and each method necessitates specific treatments for

optimal performance. Hence, we manipulate factors such as brightness, sharpness, and employ various filters within the scope of this method. The complete list of treatments is presented below (Figure 4-17).



Figure 4-17: Preprocessing of images before segmentation using the Split-Bregman method.

After applying image preprocessing, we proceed with the execution of Ander Biguri's code, enabling us to achieve image segmentation using the Split-Bregman method. As mentioned in the referenced literature, we have the capability to visualize the l1 regularization of the set of levels, providing insight into the ultimate outcome of the method.

The outcome, adjusted onto our initial image, is displayed in the leftmost image of the figure below (Figure 4-18). Numerous contours, including the one representing the endometrium, have been identified. Now, the goal is to inform the algorithm about the specific contour corresponding to the endometrium. To accomplish this, we apply an erosion operation to eliminate smaller structures and present the resulting binary visualization in the rightmost image.

In order to select the appropriate shape, we utilize the same techniques employed in the Chan-Vese model. Initially, we employ a circle with a radius of 150 pixels, positioned at the center of the image. Our objective is to identify a shape that covers an area of more than 1000 pixels. As mentioned previously, if a shape meeting these criteria is found, we determine which initial form it corresponds to. However, if no suitable shape is identified, we opt for the largest shape initially detected.



Figure 4-18: Shape selection and contour closure (identical to the Chan-Vese model).

Afterward, we apply the same technique discussed in the section covering segmentation with the Chan-Vese model without selection to close the shape. For detailed information on shape selection and the process of closing its contours, please refer to “Section V.3.d.ii” specifically addressing segmentation according to the Chan-Vese model without selection.

Upon executing this function, we can witness the outcome portrayed in Figure 4-19. It showcases the superimposition of a green contour onto the manually delineated blue mask.

4.3.8 Comparison of segmentation methods

At this point, we have two distinct segmentation methods with different functionalities. One method automatically segments the endometrium by processing the entire ultrasound image, while the other method focuses on segmenting only the user-selected region of interest.

In this section, our objective is to independently analyze the results obtained from both algorithms across the entire database. The aim is to identify a way to combine these methods effectively to optimize the performance of our code. To compare the segmentation methods, we will utilize the Jaccard index.

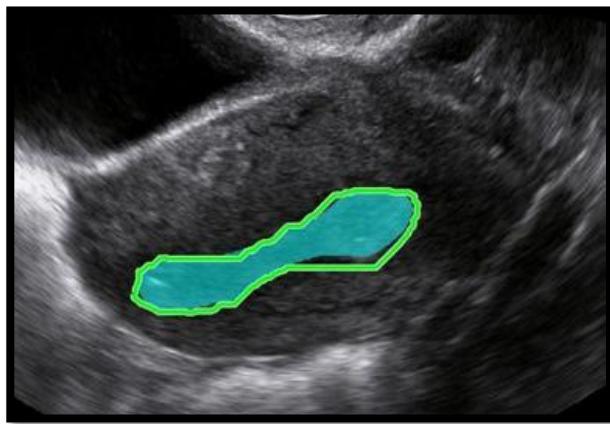


Figure 4-19: Segmentation of the endometrium using the Split-Bregman method, visualized in green, overlaid with the manual mask in blue.

To put it simply, the Jaccard index is a statistical metric used to measure the similarity between two samples. For our project, we consider a Jaccard index greater than 0.5 as an indication of good segmentation quality.

Initially, we will assess the performance of the automatic algorithm using all 98 images from the database, which corresponds to 57% of the initial image repository consisting of 172 images.

This automatic algorithm is only partially effective. Specifically, 25 images, which accounts for 26% of the database, are not processed. Additionally, in addition to the 26%, 33% of the Jaccard indices are below 0.5. Figure 4-20 is significant because it implies that for nearly two-thirds of the segmented ultrasounds, the result

is unreliable or even unusable. Next, we will analyze the semi-automatic algorithm on all 98 ultrasounds.

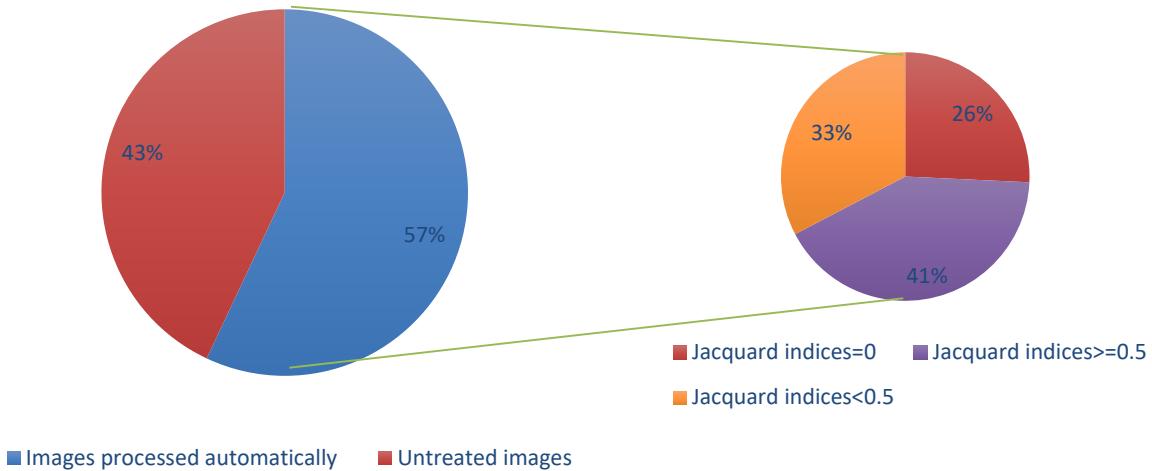


Figure 4–20: Results obtained for automatic segmentation.

In this case, we observe that the semi-automatic algorithm successfully processes the entire database without encountering any Jaccard indices equal to 0. Furthermore, it's worth mentioning that 65 images, which make up 66% of the segmented images, achieve a Jaccard index greater than 0.5 (Figure 4-21).

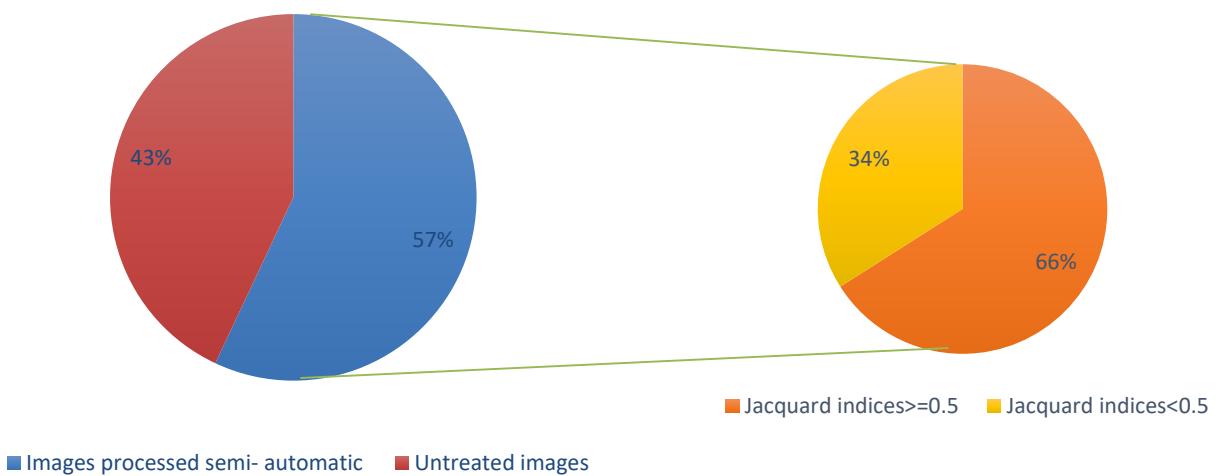


Figure 4–21: Results obtained for semi-automatic segmentation.

At first glance, this segmentation technique appears to be more effective than the previous one. However, contrary to initial expectations, the semi-automatic segmentation method does not outperform the automatic method. In fact, the accuracy of endometrium segmentation using the semi-automatic approach is

significantly lower. Only 18% of the images exhibit an increase in their Jaccard index when using the semi-automatic segmentation among the 41 images where the Jaccard index in the automatic method is greater than or equal to 0.5. Consequently, 82% of the images experience a decrease or unchanged Jaccard index.

Conversely, among the 32 images where the Jaccard index in the automatic method is less than 0.5, 63% of them demonstrate an increase in their Jaccard index with the semi-automatic segmentation. As a result, 37% of the images experience a decrease or unchanged Jaccard index.

Lastly, among the 25 images where the Jaccard index in the automatic method is zero, all of them are processed using the semi-automatic method. Among these images, 64% achieve a Jaccard index greater than or equal to 0.5.

To improve the results, it is necessary to combine both methods only for images where the Jaccard index is less than 0.5 and for images where the segmentation is not satisfactory at the endometrium level, resulting in a Jaccard index of zero.

To achieve this, it is necessary to establish one or more criteria to determine the quality and level of confidence regarding the segmentation performed, similar to the Jaccard index. As shown in the graph below, the ideal outcome for the 57% of images studied would be: no image with a Jaccard index equal to zero, 73% of images with a Jaccard index greater than or equal to 0.5, and 27% of images with a Jaccard index less than 0.5 (Figure 4-22).

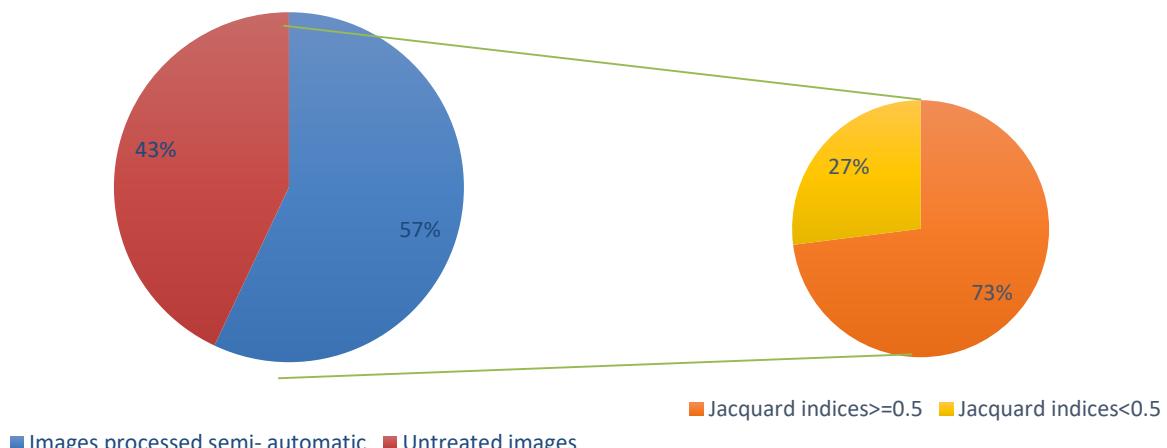


Figure 4-22 : Ideal results obtained by summing the two methods.

4.4 AUTOMATION

4.4.1 Automatic Algorithm

Segmenting images manually is a time-consuming and tedious task, which is why automation is crucial for the successful completion of the project. Automation enables the processing of a large number of images without requiring any manual

intervention. To begin with, we need to retrieve the images we want to process. In Matlab, we utilize the “uigetdir” function, which opens a dialog box allowing the selection of the folder containing the images to be segmented. The chosen folder's path is displayed in the dialog box and stored in a variable. To offer the user the flexibility to change the folder, we also employ the “questdlg” function. This function prompts the user to confirm the folder selection. If the response is negative, the “uigetdir” function is called again to enable the user to select a different folder, and the “questdlg” function intervenes once more. If the dialog box is closed, the code stops and needs to be restarted.

Once the selected folder choice is confirmed, the algorithm initiates. The number of ultrasound images in the folder is recorded in a variable called 'k,' which is used to process all the images. The processing time varies depending on the number of images to be processed. Ultimately, the segmented ultrasound images are saved in a subfolder within the previously selected folder containing the images for processing. The results are saved as JPEG images. These images consist of four areas that illustrate different aspects of the segmentation result.

Figure 4-23 provides visualization of these areas, (a) displays the original unmodified ultrasound image on the right, while in the middle area (b), it exhibits the contours of the segmentation on the original image, enabling a clearer view of the endometrial contours, which can be compared with the original image situated above. On the right side (c), the original image is displayed with a violet mask representing the obtained segmentation, facilitating the visualization of the endometrial surface.



Figure 4–23: Format of the saved results.

During the entire process, the algorithm will apply specific exclusion criteria to the segmented ultrasound images. If the images meet these criteria, they will be saved in the folder where the segmentations are stored. However, if the obtained result is abnormal, the images will be separated and placed in a different folder. These abnormal images will undergo further processing using the semi-automatic algorithm, which will be explained in “section V.4.c”. The steps of the automatic algorithm can be observed in the activity diagram provided (Figure 4-24).

4.4.2 Exclusion criteria

After running the automatic algorithm on the image database, we noticed that some images had inaccurate segmentation results. These anomalies occurred when the detected surface area of the endometrium was significantly larger than the reference values or when the localization of the endometrium didn't align with our expectations. To address this issue, we created a list of characteristics that indicated these aberrant results, with the intention of excluding them from the algorithm's processing. Our goal was to categorize these images separately so that users could later segment them using an alternative algorithm.

As I mentioned earlier, the endometrium is an organ that is relatively new in terms of segmentation, so initially, there is a lack of existing literature on the criteria for its segmentation. Additionally, the endometrium is a unique organ for each woman, which necessitates the empirical identification of criteria that can be reasonably applicable to all patients.

The first criterion we used to exclude images was related to the image processing itself. As mentioned earlier, the applied processing sometimes resulted in substantial loss of information, rendering the image unusable. Therefore, we established a threshold: if the processing produced a binary image with fewer than 12,000 white pixels, we would exclude that image from further processing. We also observed that this was particularly true for images with a low average gray level (below 40).

To find additional criteria, we focused on evaluating the homogeneity of the endometrium by examining the standard deviation. The standard deviation indicates the extent to which data points deviate from the mean. A high standard deviation suggests a wide dispersion of values, while a low standard deviation implies that the values are closely clustered around the average. Based on the values derived from the database, the average standard deviation for an endometrium was approximately 27. Consequently, we decided to exclude segmentations with standard deviations that deviated significantly from this average, whether too high or too low. Thus, we excluded images in which the detected endometrium had a standard deviation greater than 38 or less than 15.5.

However, we observed that when the standard deviations fell between 30 and 25, many segmentations were inaccurate. To avoid excluding a large number of images, we realized that additional criteria needed to be considered in conjunction with the standard deviation. Consequently, we examined circularity, surface area, and average gray level.

When the standard deviation exceeded 30 and the detected surface area was excessively large, we excluded the image because the resulting endometrium appeared aberrant. The endometrium was disproportionately large and exhibited highly dispersed pixel values, which contradicted our expectations of homogeneity. Similarly, when the standard deviation was high and the circularity was low, we deduced that the detected endometrium was false since it lacked roundness and displayed heterogeneity.

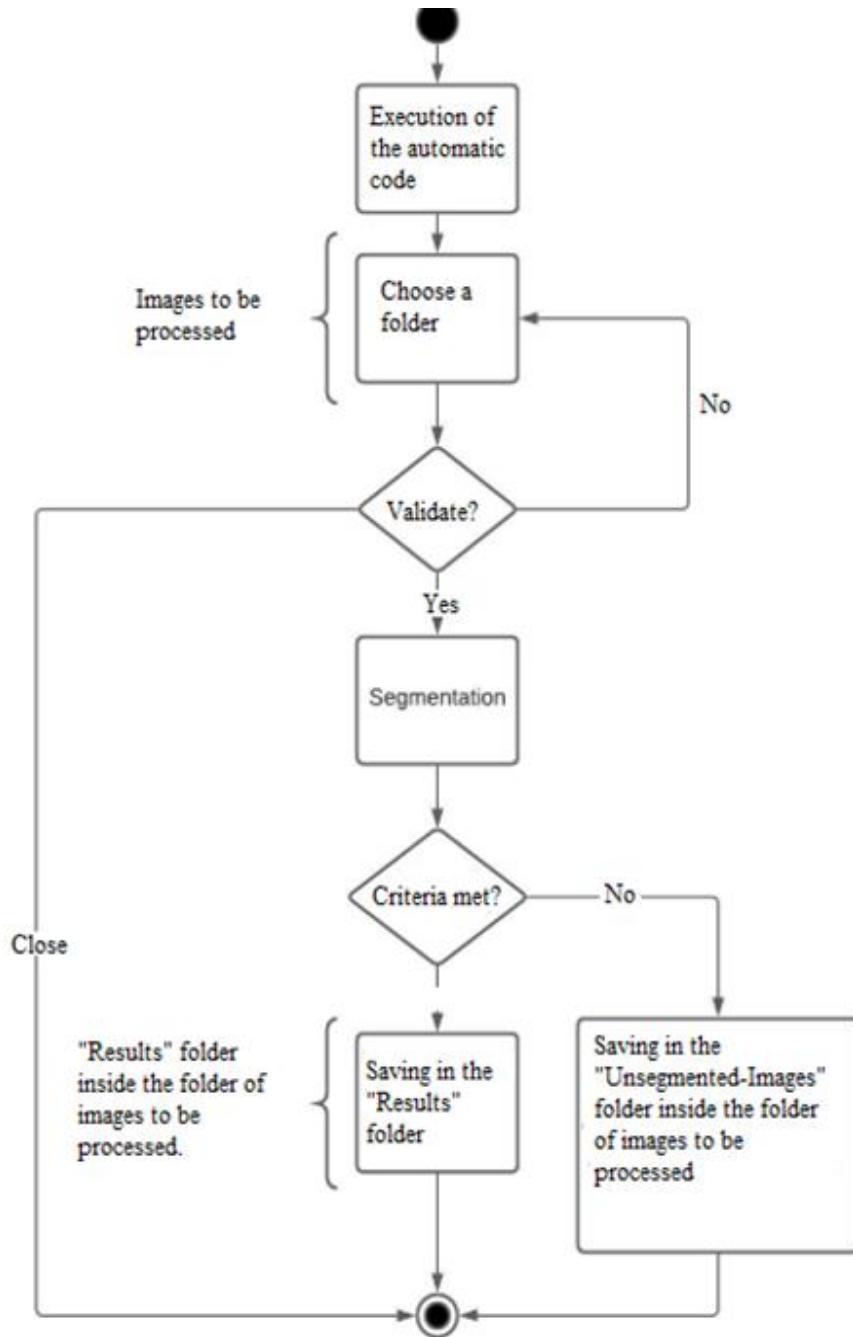


Figure 4–24: Workflow diagram of the automatic algorithm.

In such cases, we suspected over-detection. The same concept applied to smaller standard deviations (≤ 23) in terms of under-detection. Additionally, in instances where a high standard deviation indicated strong heterogeneity, we observed that a high average gray level resulted in an endometrium that appeared excessively "white," which was logically unsound. For cases where the endometrium exhibited a smaller standard deviation (≤ 23) and therefore greater homogeneity, we encountered a similar issue. However, since the values were less dispersed, we excluded those with lower average gray levels.

Lastly, we established a criterion based on the average gray level. Throughout the entire database, we observed that the calculated average gray levels generally seemed reasonable. However, values above 100 raised concerns given that the database mean was 74. After conducting further investigation, we discovered that some endometria could indeed have high average gray levels. Nonetheless, we set a threshold to filter out excessively high values as they would be considered aberrant. No observed endometrium was completely "white," so we established a high threshold that would still exclude detections with hyperintensity.

A summary of all the chosen criteria, including the number of excluded images for each criterion, is presented in Table 4-4. It's important to note that these exclusions may encompass images that could have been automatically segmented. However, as previously mentioned, we prioritized criteria that indicated aberrations. It's possible for certain endometria to fall into a category that would be considered aberrant under normal circumstances. Therefore, the criteria outlined above are not entirely reliable but serve as a guideline based on our analyses and observations. Therefore, all these criteria are subjective and should be approached with caution.

Table 4-4: Exclusion criteria chosen in the segmentation.

Chosen exclusion criteria	Number of excluded images that couldn't have been segmented by the automatic algorithm	Number of excluded images that could have been segmented by the automatic algorithm
Image processing yields fewer than 12,000 white pixels and the calculated average gray level is below 40	3	-
Detected endometrium's standard deviation is greater than or equal to 30 pixels	Detected endometrium surface area is greater than 25,000	4
	Detected endometrium's circularity is less than 0.29	3
	Detected endometrium's average gray level is above 105	1
Detected endometrium's standard deviation is less than or equal to 23	The circularity of the detected endometrium is less than 0.29	4
	Detected endometrium's average gray level is above 70 and below 90	8
Detected endometrium's standard deviation is less than 15.5 or greater than 38	7	-
Detected endometrium's average gray level is above 140	1	-

4.4.3 Semi-automatic algorithm

The semi-automatic algorithm is utilized only if the automatic algorithm has been executed beforehand. As explained in sections V.4.a (Automatic Algorithm) and V.4.b (Exclusion Criteria), the images that were excluded by the automatic algorithm based on the specified criteria are stored in a separate folder. This folder is the one processed by the semi-automatic algorithm. Similar to the automatic algorithm, the semi-automatic algorithm prompts the user to select a folder. Specifically, the user needs to choose the folder containing the images that were not processed by the automatic algorithm, which is typically named "no segmented images" by default. Once the folder is selected, the algorithm presents a pop-up window to the user, allowing them to confirm their choice or, if desired, change the selected folder. Upon confirming the folder selection, the image processing begins. Unlike the automatic algorithm, the semi-automatic algorithm involves user interaction. For each image, the algorithm requests the user to manually select a region where they believe the endometrium is located. After the user makes this selection, the segmentation process takes place. Once the result is saved, the user repeats this selection procedure for the subsequent image and continues until all the images in the folder have been processed.

The activity diagram below (Figure 4-25) provides a visual representation of the various steps involved in the semi-automatic algorithm.

4.5 FINAL RESULTS

The analysis of our automated or semi-automated segmentation relies on several key metrics. Initially, we utilized the Jaccard index, which is a statistical metric used to compare the similarity between two samples. We also examined two other metrics represented as percentages: sensitivity and specificity. Sensitivity evaluates the ability to produce a positive outcome when a hypothesis is confirmed, while specificity assesses the ability to generate a negative outcome when the hypothesis is not confirmed.

We collected a range of numerical values required to compute sensitivity and specificity, as outlined below:

$$\frac{TP}{(TP+FN)}, \quad \frac{TN}{(TN+FP)}$$

Indeed, it was crucial to gather counts for false negatives (FN), false positives (FP), true negatives (TN), and true positives (TP) in order to perform our calculations. Each of these terms was defined within the context of our specific case study in the part of the segmentation:

- False negatives represent the number of pixels that should have been detected by the segmentation algorithm but were missed, as they are part of the manual mask.
- True negatives represent the number of pixels correctly classified by the segmentation algorithm as not belonging to the endometrium, based on the manual mask.

- False positives represent the number of pixels incorrectly identified by the segmentation algorithm as belonging to the endometrium, despite not being part of the manual mask.
- True positives represent the number of pixels correctly classified by the segmentation algorithm as belonging to the endometrium, according to the manual mask.

These metrics enable us to evaluate the performance of our segmentation algorithms for each image and determine the reliability of the developed segmentation. Now, let's visualize this information across multiple processed images using both algorithms and delve into the results.

As indicated in Table 4-5, we can observe three different scenarios. The first scenario occurs when both segmentations produce nearly identical outcomes. In the initial row of the table, it is evident that the semi-automatic and automatic algorithms yield a Jaccard index of approximately 0.80. This signifies a significant similarity between the two sets. The figures accompanying the table illustrate this, with the blue region representing the overlapping area between our segmentation and the manual mask. The red pixels correspond to areas that our algorithm failed to detect, although they are part of the endometrium. Additionally, the green pixels indicate falsely detected areas.

Consequently, we can determine the sensitivity and specificity, both of which are relatively high for this image. It is important to note that the automatic algorithm exhibits a greater precision in segmentation compared to the semi-automatic algorithm, as explained in “Section IV.3.8”, Comparison of Segmentation Methods. This scenario perfectly exemplifies that point.

The second scenario arises when the automatic algorithm completely fails to detect the endometrium and erroneously locates it in an unrelated area. Conversely, the semi-automatic algorithm, with the assistance of user initialization, achieves a relatively accurate detection. Consequently, we obtain a Jaccard index of 0 for the automatic algorithm and a Jaccard index of 0.72 for the semi-automatic algorithm. Similar to the first scenario, we visualize the outcomes using the same color scheme and subsequently derive the values of sensitivity and specificity.

It is worth noting that, since sensitivity depends on true positives, which are absent in the case of the automatic algorithm, it is logical to obtain a sensitivity value of 0. The same logic applies to specificity, as it relies on the number of true negatives. However, in the case of automatic segmentation, numerous pixels that should not have been detected are missed, despite the presence of green regions. Hence, obtaining a specificity value of 0.97 for this scenario is considered normal.

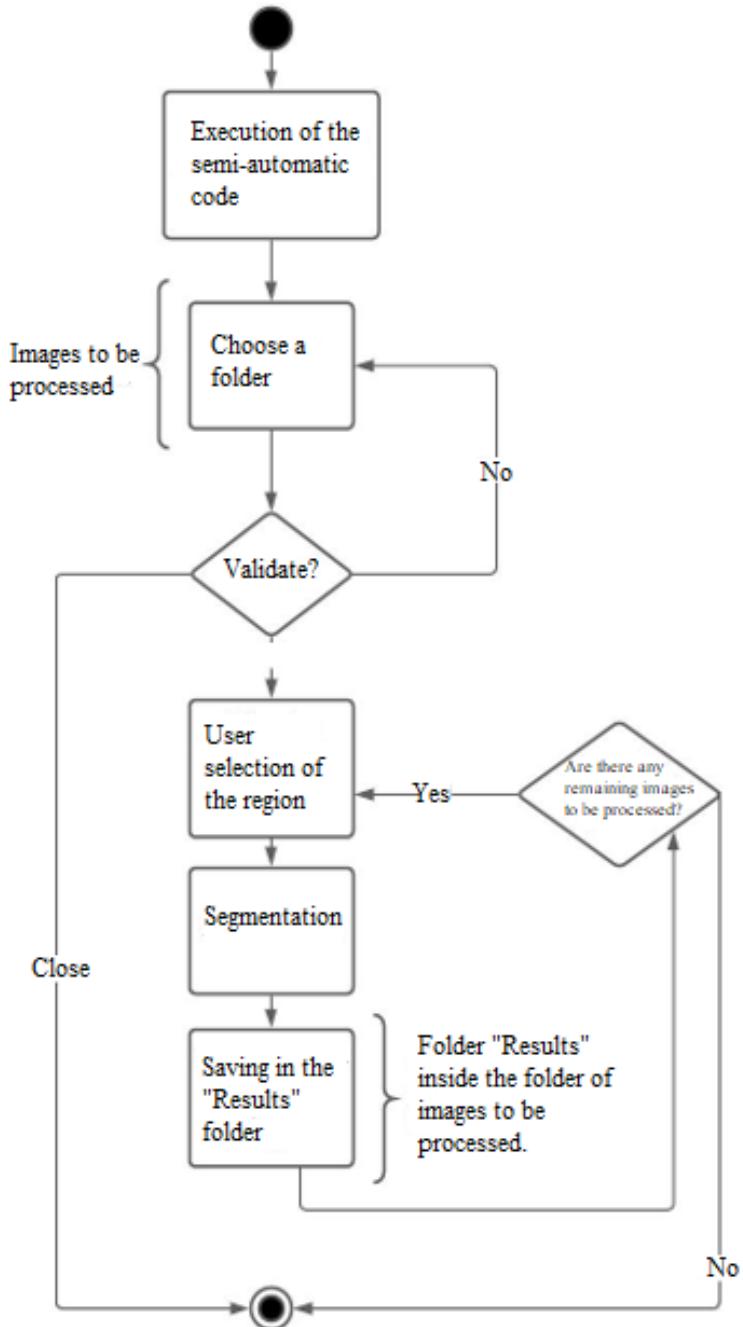
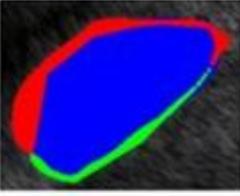
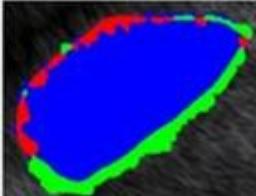
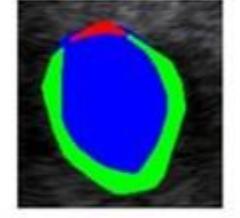
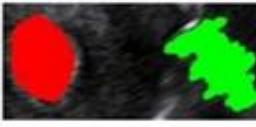
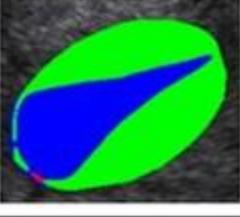
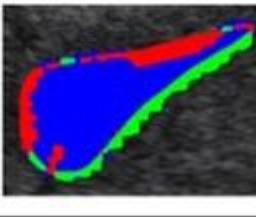


Figure 4–25: Activity diagram of the semi-automatic algorithm.

Table 4-5: The final results of the semi-automatic and automatic segmentation

Semi-automatic segmentation	Automatic segmentation
	Jaccard index : 0,81 Sensitivity : 0,85 Specificity : 0,99 
	Jaccard index : 0,72 Sensitivity : 0,97 Specificity : 0,99 
	Jaccard index : 0,45 Sensitivity : 0,99 Specificity : 0,96 

The final scenario occurs when the automatic algorithm achieves a more precise segmentation than the semi-automatic algorithm. We can observe the disparity in the Jaccard index between the two segmentation results, along with the resulting sensitivity and specificity values. Based on these findings, we favor a lower sensitivity (0.85 in the automatic algorithm) over a higher sensitivity (0.99 with the semi-automatic algorithm) because the segmentation obtained in the second case is less "faithful" compared to the first case. Although the semi-automatic algorithm yields a greater number of true positives, it also produces more false positives. Therefore, in this scenario, the outcome and similarity to the manual mask are more favorable with the automatic algorithm.

4.5.1 Analysis of the Results

The bank of ultrasound images must first be processed by the automatic algorithm, which excludes a certain number of images based on established criteria. These images will be segmented in a second step using the semi-automatic code. In this section, we will analyze the results obtained step by step.

Initially, we had 98 images to segment. 37 of them are excluded by our algorithm according to the established criteria. This corresponds to 37.75% of the Figure 4-26. Therefore, we are left with 61 images to process.

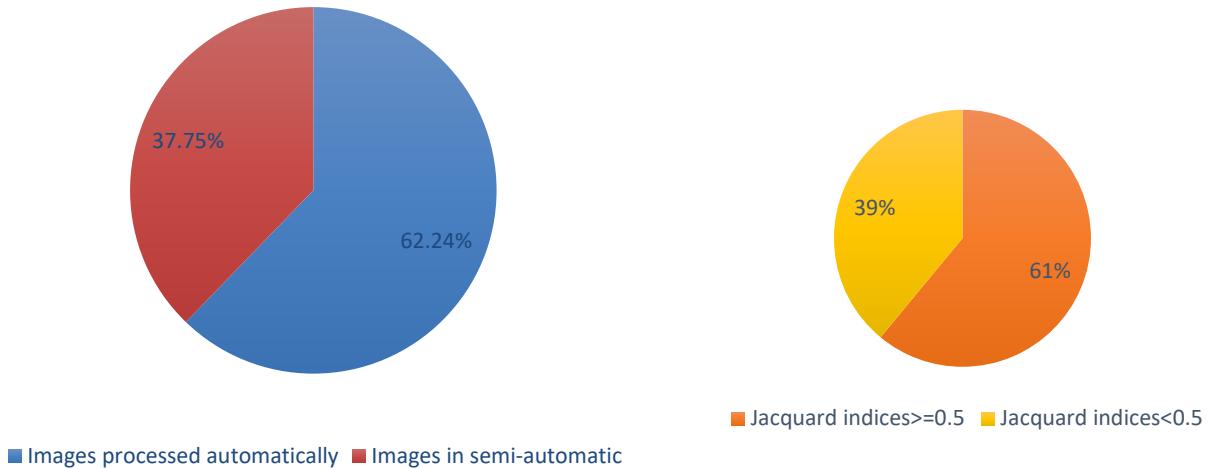


Figure 4–26: Results obtained for automatic segmentation.

Among these, 37 ultimately have a Jaccard index greater than or equal to 0.5. This indicates a relatively reliable segmentation based on the previous explanations of the metric. This represents 61% of the 61 images processed by the automatic algorithm. The remaining 24 are all segmented but have a Jaccard index lower than 0.5. Figure 4-27 represents a more detailed distribution of the Jaccard indices for automatic segmentation.

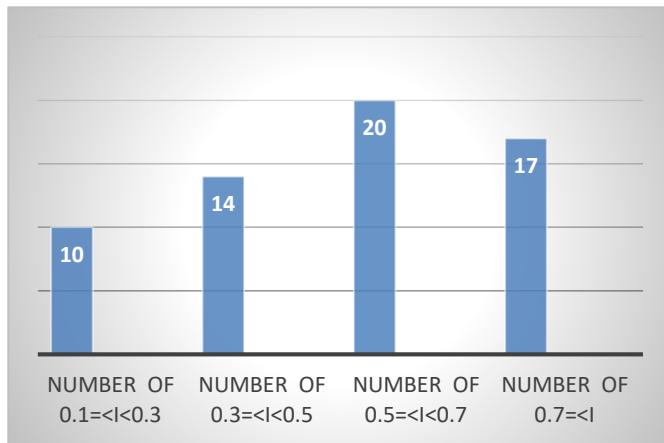


Figure 4–27: Distribution of Jaccard indices for automatic segmentation.

Following the execution of the automatic algorithm, we move forward with the semi-automatic segmentation of the 37 excluded images. Out of these images, 23 achieve a Jaccard index of 0.5 or higher, accounting for 62% of our total of 37 images. Conversely, 14 images obtain a Jaccard index below 0.5 (Figure 4-28).

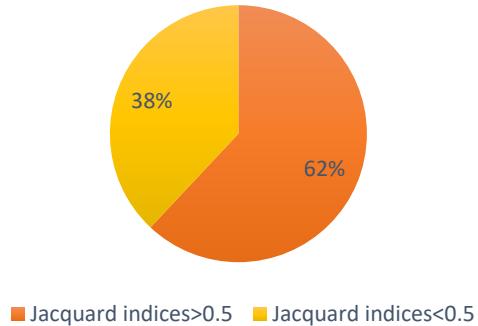


Figure 4–28: Results obtained for semi-automatic segmentation

Once again, Figure 4-29 provides a comprehensive breakdown of the Jaccard indices specifically for the semi-automatic segmentation.

The graph illustrates that the vast majority of images have a Jaccard index above 0.3, indicating that the semi-automatic segmentation is quite reliable. Taking into account the combined results of both algorithms on the initial set of 98 images, we can conclude that the segmentation process has been successful overall.

To begin with, it is noteworthy that all 98 images from the initial database have been processed. This achievement is significant as it ensures a plausible segmentation regardless of the specific outcome. Among the 98 ultrasounds, 60 of them ultimately exhibit a Jaccard index greater than or equal to 0.5, indicating a high-quality segmentation. This accounts for 61% of our initial database. The remaining 38 images are all segmented but have a Jaccard index lower than 0.5, suggesting a slightly less reliable segmentation. However, it is important to consider that during the analysis of manual masks, certain images obtained Jaccard indices lower than 0.3 despite being manually created. Therefore, these results can be put into perspective, and we can conclude that they are generally highly satisfactory. This result is described in Figure 5-30.

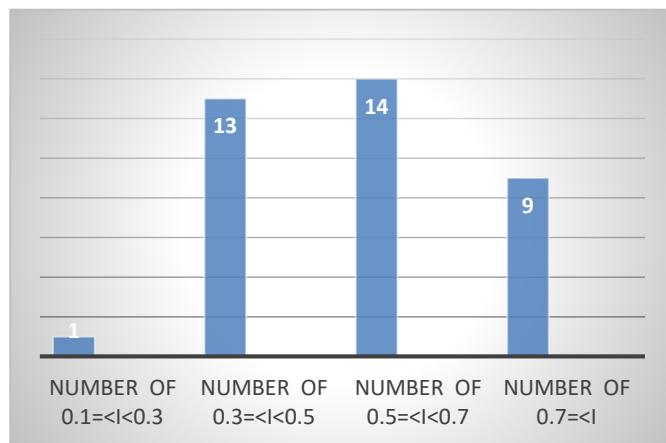


Figure 4–29: Distribution of Jaccard indices for semi-automatic segmentation.

4.6 DISCUSSION FOR THE SEGMENTATION PART

The segmentation of the endometrium, which involves extracting the specific region in an ultrasound (US) image, is a critical step in an Assisted Reproductive Technology (ART) cycle. The segmentation results are used to calculate quantitative features that describe the endometrial volume, size, echo pattern, thickness, and more. These features are then input into a classifier to determine the outcome of the treatment. Hence, the accuracy of the segmentation directly impacts the performance of the quantitative analysis and diagnosis of the treatment. Over the course of the segmentation work, we have developed two solutions for segmenting the endometrium in ultrasound images in order to calculate these parameters accurately and reliably. Obtaining precise segmentation was crucial to ensure the accurate calculation of endometrial characteristics. The first solution we implemented is an automated algorithm that performs an initial segmentation of the endometrium in each ultrasound image from a given dataset. Based on predetermined criteria established through the study of endometrial ultrasounds, the segmentation produced by this first solution is validated or rejected. In cases of rejection, the second solution comes into play, allowing for a result to be obtained. This second solution is semi-automatic, where the user manually selects the region of interest believed to contain the endometrium. The segmentation and calculation of endometrial characteristics remain automated. By combining these two solutions, we achieve a perfect compromise, as all the ultrasound images in our database are segmented. The results obtained thus far are highly promising for the future of the project. 61% of the endometrial segmentations are reliable, and even the remaining 39% are accurate in terms of endometrial localization, despite incomplete segmentation. Therefore, in the first case, the calculated endometrial characteristics are usable, while in the second case, these results may be distorted and cannot be relied upon by a physician for a definitive diagnosis.

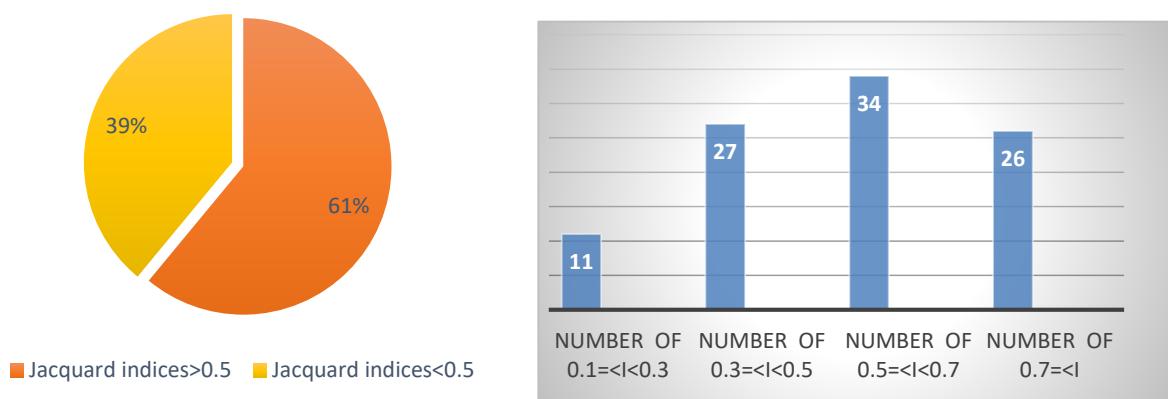


Figure 4-30: Final Results.

4.7 CLASSIFICATION USING OUR SEGMENTED MEASUREMENTS

Within this part, we enhance SPIRL Version II by transitioning it into a fully automated method as you can see in Figure 4-31. This is achieved by utilizing measurements obtained through automatic endometrium segmentation and inputting them into a classifier that predicts the outcome of ICSI without manual intervention.

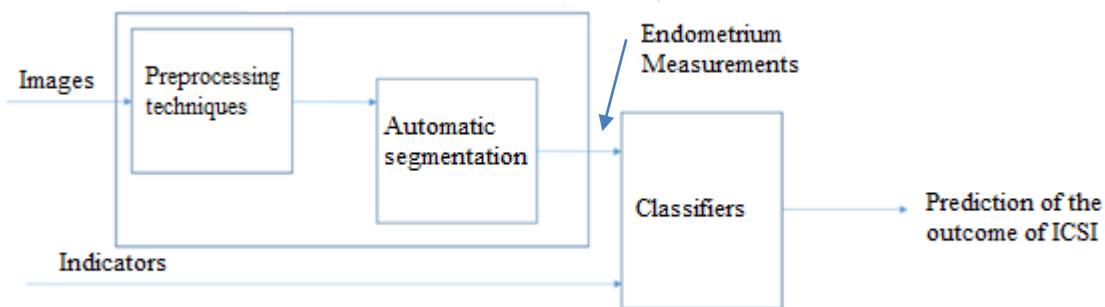


Figure 4-31: The Complete scheme of SPIRL Version III.

4.8 MATERIAL AND METHODS

4.8.1 Data collection and preparation:

In this study, data was gathered from a total of 61 patients undergoing ICSI. The images used for analysis were those that had been fully segmented in the previous chapter. The measurements of endometrial surface, thickness, and mean grey values were automatically obtained from the segmentation model. These characteristics were then combined with other features directly extracted from "TrakMD platform". In this case, a combination of Doppler parameters (measured ones) and clinical parameters are utilized as features to be inputted into the classifier. The choice of clinical parameters combined to our measurement parameters in this case is influenced by the absence of missing values in this small dataset (61 patients). Two files were prepared: the first file contained the actual measurements obtained using the Ultrasound machine, while the second file contained the features measured using our segmentation technique.

Eight features are taken into consideration in each file. Table 4-6 presents the features along with their corresponding numbers: age, tentative, day of transfer, number of embryo transfers, surface in mm, endometrial thickness, and endometrial Mean Grey.

Table 4-6: The selected features with their number.

Number of features	Definition of Parameters
Feature 0	Age of the patient
Feature 1	Tentative which is the number of ICSI trials
Feature 2	The number of the Embryo transferred to the uterus
Feature 3	Transfer day
Feature 4	Endometrial Thickness (mm)
Feature 5	Site of embryo (cm), which is the distance between the spot of the embryo and the fundus of the uterus.
Feature 6	Surface of the endometrium(mm)
Feature 7	Endometrium Mean Grey

4.8.2 Model Selection

The feature selection technique employed in this study is Extra Tree importance, which was selected as the optimal technique based on the SPIRL Version II chapter. Random Forest was chosen as the predictive model for determining the outcome of ICSI, considering it performed better in SPIRL Version I due to its reduced computational time compared with Bagging classifier the one he gives the highest performance in SPIRL Version II.

4.9 RESULTS

In this section, to compare SPIRL VERSION II and SPIRL VERSION III, it is crucial to perform classification on the measurements obtained using VOCAL software (following the same steps as in SPIRL VERSION II) and the measurements obtained through our segmentation technique. The identical clinical features are utilized, combined with the three measurement features from both VOCAL software (manual) and our segmentation technique. Subsequently, the same feature selection technique and classifier are employed, leading us to our results.

In Figure 4-32, you can observe that certain factors, namely age, tentative, embryo transfer, and transfer day, hold a similar level of importance. Age consistently remains the most crucial factor, as emphasized in the initial chapter. It is evident that there is no variation in relation to this factor between the two situations. Figure 6 in Page 18, shows that age is the most important parameter, confirming studies showing that the success rate decreases with female age. The 3 most important parameters are our 3 measurements derived from the segmentation (manual or automatic) of ultrasound images (blue colors). The values are fairly close, whether obtained manually or automatically. This is reassuring in view of the use of SPIRL V3 in real-life conditions. The orange bars in both histograms represent the same factors, namely the number of embryos transferred, tentative, site of the embryo and the transfer day. In other words, there is no a significant difference between the ranking of features for the real measurement compared with the SPIRL Version III.

Figure 4-33 displays the accuracy results for predicting the outcome of ICSI using features obtained directly from the ultrasound machine (semi-automated

machine) and the fully automated SPIRL Version III. For the semi-automated machine, the accuracy is 0.555, while for SPIRL Version III, the accuracy is 0.518. Additionally, the AUC is 0.5 for SPIRL Version III, compared to 0.561 for the semi-automated machine. Overall, it can be observed that there is not a significant difference between the two techniques.

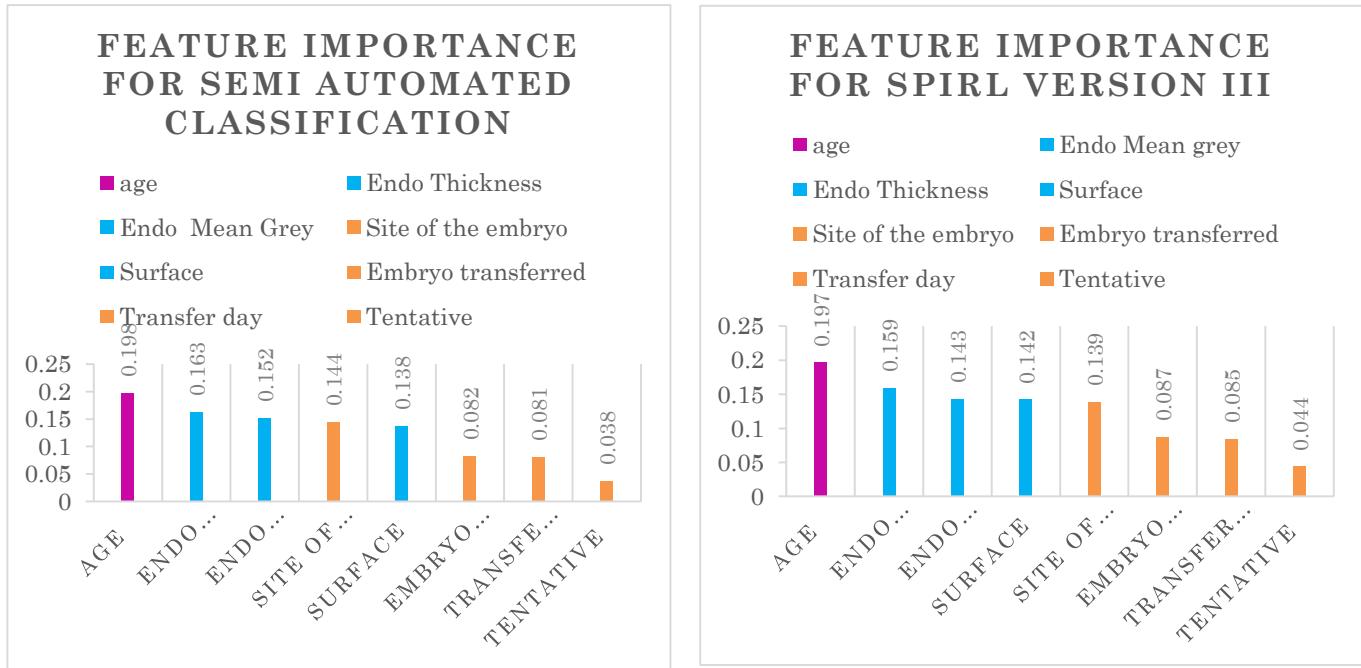


Figure 4–32: Histograms represent the feature Ranking after an Extra Tree Importance Feature selection technique for the two files.

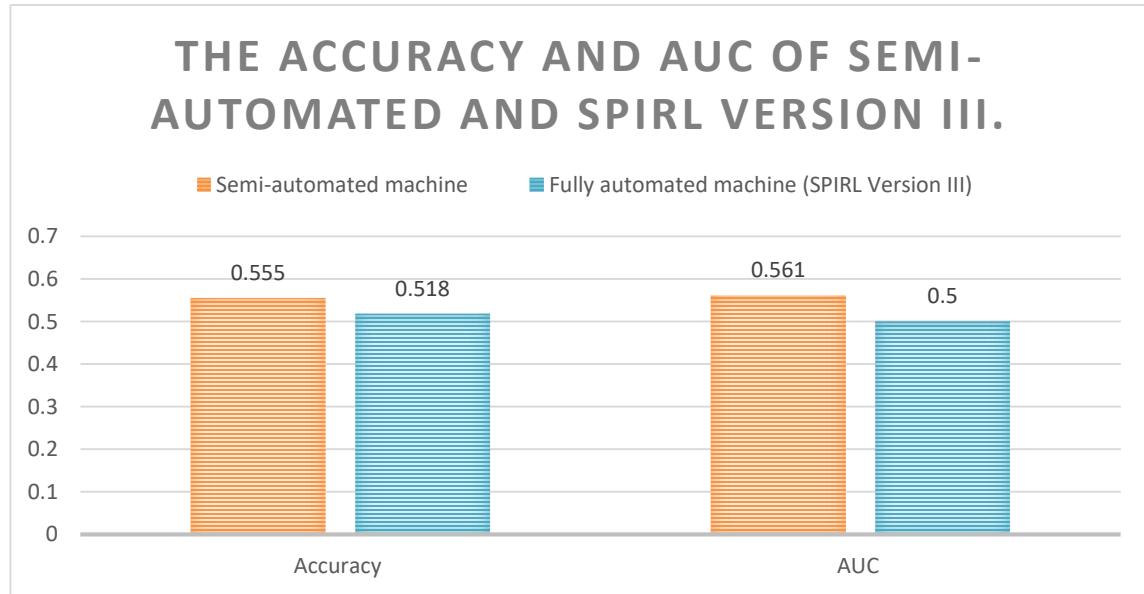


Figure 4–33: The accuracy and AUC of semi-automated and SPIRL Version III.

4.10 DISCUSSION

The chapter aims to enhance the prediction of the outcome of Intracytoplasmic Sperm Injection (ICSI) by combining the classification and segmentation components. The process involves utilizing a 2D ultrasound machine to measure various parameters. These parameters are then combined with additional features to predict the outcome of ICSI.

One of the advantages of this approach is its ability to provide accurate results without consuming excessive time for clinicians. By integrating the measured parameters with other relevant features, the prediction process becomes automatic, more efficient and less time-consuming.

Interestingly, the chapter compares the accuracy of predictions using two different methods: real measurements and a fully automated segmentation technique. The accuracy of the predictions based on real measurements is reported to be 0.55. On the other hand, when employing the fully automated segmentation technique, the accuracy slightly decreases to 0.51. Furthermore, when examining the ranking of features overall, it is evident that the most important factor remains unchanged, while the least three factors also remain the same. There is a slight variation in the order of surface, endometrial thickness, site of the embryo, and endometrial mean grey. However, the overall ranking of features is quite similar between the two files.

Despite the slight drop in accuracy, it is important to note that the difference between the prediction outcomes using real features and those obtained through the segmentation method is not significant. This suggests that the segmentation technique is a viable alternative, as it provides comparable results to the more time-consuming process of manually measuring the features.

However, it's crucial to consider the limitations of the study. The reported results are based on a relatively small dataset of only 61 images. To obtain more precise and reliable results, it is recommended to gather a larger and more diverse dataset. Increasing the sample size would likely improve the accuracy and generalizability of the predictions.

In summary, this chapter presents a promising approach to predict the outcome of ICSI by combining classification and segmentation techniques. The integration of parameters measured from a 2D ultrasound machine with additional features allows for efficient and accurate predictions. Although the fully automated segmentation technique shows a slightly lower accuracy compared to real measurements, the difference is not substantial. To achieve more precise results, future studies should focus on expanding the dataset to enhance the reliability of the predictions.

This particular aspect of the study is absent in the existing literature, primarily because of the originality of endometrial segmentation and its recent application. Previous studies have primarily focused on utilizing clinical features for both males and females, as well as some Doppler parameters, to predict the outcome of ICSI without incorporating segmentation into their research.

CHAPTER 5

GENERAL DISCUSSION, FUTURE WORK AND SOCIAL IMPACT

5.1 GENERAL DISCUSSION

IVF therapy is a complex and demanding process for both female and male patients. It involves the administration of various hormones and medications to stimulate follicle growth and prepare the uterus for embryo implantation. The treatment can be physically and emotionally challenging, and the success rates may vary depending on several factors such as age, underlying fertility issues, and overall health. The decision to undergo IVF treatment is not taken lightly by couples. The procedure is expensive, time-consuming, and emotionally taxing. Before embarking on the treatment, couples often consider the chances of success based on various factors, such as their medical history, age, and the underlying cause of infertility. If the chances of success are low, some couples may decide not to proceed with the treatment. In the context of IVF, machine learning techniques can be applied to analyze data and provide insights that can aid physicians and infertile couples in making informed decisions. By leveraging machine learning algorithms, researchers can identify patterns, relationships, and predictive factors that contribute to successful clinical pregnancy outcomes. This information can help physicians tailor treatment plans and provide couples with realistic expectations. The goal of this work was to employ machine learning algorithms to analyze new Doppler factors influencing clinical pregnancy outcomes, which could be important information for physicians and infertile couples. The real-world implementation of common machine learning necessitates careful examination of the input data and method employed. Choosing the most relevant pre-processing tasks improves performance. Our analysis showed that Bagging Classifier with Extra Tree Importance for feature selection as a pre-processing technique outperformed other models when predicting the outcome of the ICSI procedure with an accuracy of 89% and an AUC of 0.82 for 94 patients. Same model utilizing RF importance as a feature selection technique surpassed other models by achieving an accuracy of 74% and an AUC of 78.2% when applied to a dataset of 572 patients. The comparison between Doppler, clinical, and all parameters showed that the accuracy and the AUC of the classification with all parameters are higher than the classification with Doppler or Clinical parameters. Furthermore, a future study would require a substantial amount of data to comprehensively evaluate the significance of Doppler parameters with a better performance. Power Doppler parameters related to the endometrium and myometrium play a crucial role in the treatment. These parameters are currently measured manually using VOCAL software, which is a time-consuming process as it requires calculations based on six slices of the endometrium. we aimed to automatically segment the endometrium to calculate its volume, thickness, Doppler

parameters, and other relevant features. In the initial part of the study, we employed automatic and semi-automatic segmentation techniques, specifically the Chan-Vese and Split-Bregman techniques. Out of the 98 ultrasound images, 60 achieved a Jaccard index greater than or equal to 0.5, indicating high-quality segmentation. This accounted for 61% of our initial database. The remaining 38 images were segmented but had a Jaccard index below 0.5, suggesting slightly less reliable segmentation. Our overall goal had two main components. Initially, we aimed to segment the endometrium using an automated technique. Subsequently, our objective was to predict the outcome of ICSI using machine learning techniques based on the parameters obtained from our segmentation. This represents the final stage of our work, where we observed positive results as there was no significant difference in accuracy and AUC between the classification using RF for two sets of data. One set consisted of real measurements obtained using an ultrasound machine, while the other set involved parameters measured through fully automated segmentation. Additionally, the disparity in the importance ranking of features between the two sets is minimal. In summary, this section concludes that our automatic segmentation method performs well and provides the necessary measurements to accurately predict the outcome of ICSI.

However, it is evident that acquiring more data is necessary because our last phase of analysis was conducted using a limited sample of only 61 images. To ensure the reliability and applicability of our machine learning model's predictions, it is crucial to expand the dataset. By including a larger and more diverse range of data, we can better capture variations and improve the overall validity of our results.

The ability to predict the outcome of ICSI with a reasonable accuracy of 74% and an AUC of 78.2% has significant implications for stakeholders in the medical field and patients. ALHADI Laboratory, as a dedicated center for improving success rates in the market of ART, prioritizes advancements in this area. Our results empower healthcare providers to make more informed decisions about treatment options for couples undergoing ICSI, enabling them to assess the chances of success and customize treatment plans accordingly. Additionally, the inclusion of Myometrium Doppler parameters as important features in our analysis provides valuable information for clinicians, potentially improving success rates at our center.

For patients, knowing the predicted outcome of ICSI is crucial in managing expectations and making informed choices regarding their fertility journey. It allows healthcare providers to offer appropriate counseling and support, addressing emotional well-being and reducing stress associated with the uncertainty of the treatment process. Furthermore, predictive models can assist patients in making financial decisions by estimating the likelihood of success. This knowledge helps individuals and couples determine the number of treatment cycles they are willing to undergo, potentially reducing financial burden and facilitating better financial planning.

Overall, our objective of improving success rates in ICSI was achieved with satisfactory performance.

5.2 FUTURE WORK

The segmentation of organs and other substructures in medical images plays a crucial role in enabling quantitative analysis of clinical parameters such as volume and shape. This analysis is particularly important in fields like cardiac and brain analysis. Among the applications of deep learning in medical imaging, segmentation has received significant attention, resulting in a wide range of methodologies being proposed.

One of the most prominent and well-known convolutional neural network (CNN) architectures for medical image analysis is U-net, which was introduced by Ronneberger [88]. U-net was specifically developed for biomedical image segmentation, with the aim of predicting the class for each pixel in the image. It is an extension of the Fully Convolutional Network architecture, modified to improve segmentation performance in medical imaging applications [89]. In a study referenced as [90], an automatic segmentation technique achieved a dice similarity coefficient of 0.86 specifically for left ventricle segmentation on a large ultrasound dataset. Additionally, in the task of lymph node segmentation, the U-net architecture demonstrated superior performance compared to other state-of-the-art deep learning methods, as mentioned in [91]. With the rise of deep learning, transfer learning has become a crucial component in various applications, particularly in medical imaging. The current standard practice involves utilizing existing architectures designed for natural image datasets like ImageNet, along with their pretrained weights, and fine-tuning the models using medical imaging data [92]. In this study, the authors explored the effectiveness of Convolutional Neural Networks, including U-Net, combined with VGG16 encoder for segmenting ultrasound scans of arteries [93]. In order to facilitate the work of physicians and achieve accurate measurements, it is crucial to perform segmentation of the endometrium and myometrium on both 2D and 3D images. This can be accomplished by employing U-Net and transfer learning techniques. By accurately segmenting these regions, physicians can benefit from enhanced visualization and precise measurements, ultimately improving the overall diagnostic and treatment process. Once a segmentation is achieved, which closely matches the manual segmentation obtained using VOCAL software with an accuracy of over 90%, automated measurements can be performed using Python or MATLAB. This enables quick classification using machine learning techniques to determine whether the patient is likely to have a successful or unsuccessful Intracytoplasmic Sperm Injection (ICSI) procedure. The combination of accurate segmentation and automated measurements, followed by machine learning-based classification, allows for efficient and timely decision-making in assessing the potential outcome of ICSI for each patient. Due to the significant memory requirements of deep learning algorithms, our work faces a limitation in terms of memory resources. To address this issue, we initially focused on segmenting the endometrium using U-net and other deep learning architectures, but we were only able to work with a limited dataset of 20 images. The results of Deep learning segmentation are presented in the “Appendix” section. As a future endeavor, we intend to continue this work and explore ways to overcome the memory constraints, allowing for a larger dataset to be utilized in our segmentation process and continue our SPIRL Version III to enhance overall

performance. Furthermore, it is crucial to extend the prediction capabilities to include the forecast of live birth outcomes resulting from ICSI, rather than solely focusing on predicting pregnancy outcomes.

5.3 SOCIAL IMPACT

The market solutions for enhancing the success rates of IVF encompass a range of strategies and advancements within the field of ART. ALHADI Laboratory, as a center dedicated to improving the market, prioritizes the improvement of success rates. It is the first and only center in Lebanon that owns the techniques that permit to test the genetic information of the embryos before implantation. This innovative approach enables the examination of all chromosomes on-site, facilitating the identification of abnormalities, aneuploidy, and even gender determination, this test ameliorates the success rate of IVF as you can see in Figure 1-1. Also, Cryopreservation techniques have advanced the success rates of FET cycles. Embryos can be frozen and stored for future use, allowing for better synchronization with the woman's natural menstrual cycle and optimizing the chances of successful implantation. This technique is used in the Center. The clinicians in ALHADI laboratory continuously engage in research and exploration to identify new parameters or factors that can impact the success rate of treatments. This is our goal to discover additional variables that can assist in tailoring the treatment protocols for individual couples. By actively seeking out and studying these factors, clinicians strive to enhance the effectiveness of their treatment approaches and improve the chances of success for each couple.

In summary, the societal implications of utilizing machine learning to predict the outcome of ICSI are far-reaching. These include the potential to enhance success rates, reduce financial stress, offer individualized treatment approaches and improve emotional well-being, surrounding assisted reproductive technologies.

REFERENCES

- [1] P. Njagi, "Financial costs of assisted reproductive technology for patients in low- and middle-income countries:a systematic review," *Human Reproduction Open*, vol. Vol.2023, no. 2, p. 16, 2023.
- [2] J. F. Aderaldo, B. h. D. Rodrigues de Albuquerque, M. T. F. Câmara de Oliveira, M. De Medeiros Garcia Torres and F. L. D. C., "Main topics in assisted reproductive market: A scoping review," *National Library of Medicine*, vol. 18, no. 8, 2023.
- [3] C. Fakih, F. Fakih, F. M. and Y. Mourad, "<http://lab-alhadi.com/>," [Online].
- [4] R. Guh, T. C. J. Wu and S. P. Weng, "Integrating genetic algorithm and decision tree learning for assistance in predicting in vitro fertilization outcomes," . *Expert Systems with Applications*, pp. 4437-4449, 2011.
- [5] P. Amini, F. Ramezanali, M. Parchehbaf-Kashani, S. Maroufizadeh, R. Omanisamani and A. Ghaheri, "Factors associated with in vitro fertilization live birth outcome: a comparison of different classification methods," *Int J Fertil Steril*, pp. 128-134, 2021.
- [6] C. W. Wang, C.-Y. Kuo and C.-H. Chen, "Predicting clinical pregnancy using clinical features and machine learning algorithms in in vitro fertilization," *PLOS ONE*, 2022.
- [7] P. R. Supramaniam, M. Mittal, E. McVeigh and L. N. Lim, "The correlation between raised body mass index and assissted reproductive treatment outcomes: a systematic review and meta analysis of the evidence," *Reproductive Health*, 2018.
- [8] A. K. Bartmann, M. Faria, G. P. d. Barros, L. S. d. Paula and N. R. Bettini, "The number of embryos obtained can offset the age factor in IVF Results According to an Artificial Intelligence," *Women's Health & Gynecology*, p. 307, 2016.

- [9] M. Babitha, "A Survey on the Machine Learning Techniques used in IVF Treatment to Improve the Success Rate," in *RTESIT*, 2019.
- [10] S. J. Kaufmann, J. L. Eastaugh, S. Snowden, S. W. Smye and V. Sharma, "The application of neural networks in predicting the outcome of in-vitro fertilization," *Human Reproduction*, pp. 1454-1457, 1997.
- [11] A. Uyar, A. Bener, H. N. Ciray and M. Bahceci, "Predicting Implantation Outcome from Imbalanced IVF dataset," in *Proceedings of the word congress on engineering and computer science*, San Francisco, USA, 2009.
- [12] A. Uyar, A. Bener and H. N. Ciray, "Predictive Modeling of Implantation Outcome in an In Vitro Fertilization Setting: An Application of Machine Learning Methods," *Med Decis Making*, p. 714, 2015.
- [13] M. a. K. R. N. Durairaj, "Data Mining Application on IVF Data For The Selection of Influential Parameters on Fertility," *International Journal of Engineering and Advanced Technology (IJEAT)*, pp. 262-266, 2013.
- [14] P. Hafiz, M. Nematollah, R. Boostani and N. J. Bahia, "Predicting implantation outcome of In Vitro fertilization and intracytoplasmic sperm injection using data mining techniques," *Fertility and Sterility*, pp. 184-190, 2017.
- [15] J. Qiu, P. Li, M. Dong, X. Xin and J. Tan, "Personalized prediction of live birth prior to the first in vitro fertilization treatment: a machine learning method.,," *Journal of Translation Medicine*, p. 17, 2019.
- [16] R. Hassan, S. Al-Insaif, M. I. Hossain and J. Kamruzzaman, "A Machine Learning Approach for Prediction of Pregnancy Outcome following IVF Treatment," *Neural Computing and Applications*, 2020.
- [17] R. Liu, S. Bai, X. Jiang, L. Luo, X. Tong, S. Zheng, Y. Wang and B. Xu, "Multifactor Prediction of Embryo Transfer Outcomes Based on a Machine Learning Algorithm," *ORIGINAL RESEARCH article*, 2021.
- [18] N. C. N. Barreto, G. Z. Castro, R. G. Pereira, F. A. N. Pereira, F. M. Reis, W. M. Junior, I. K. D. Cavallo and K. B. Gomes, "Predicting in vitro fertilization success in the Brazilian public health system: a machine learning approach," *Med Biol Eng Comput*, 2022.
- [19] J.-Y. Wen, C.-F. Liu, M.-T. Chung and Y.-C. Tsai, "Artificial intelligence model to predict pregnancy and multiple pregnancy risk following in vitro

fertilization-embryo transfer (IVF-ET)," *Taiwanese Journal of Obstetrics and Gynecology*, pp. 837-846, 2022.

- [20] M. Ivanovski, "The Role of Ultrasound in the Evaluation of Endometrial Receptivity Following Assisted Reproductive Treatments," in *In Vitro Fertilization - Innovative Clinical and Laboratory Aspects*, Macedonia, www.Intechopen.com, 2012, pp. 31-68.
- [21] J. Zhao, Q. Zhang and Y. Li, "The effect of endometrial thickness and pattern measured by ultrasonography on pregnancy outcomes during IVF-ET cycles," *Reproductive Biology and Endocrinology*, 2012.
- [22] Q. Zhang, X. Wang, Y. Zhang, H. Lu and Y. Yu, "Nomogram prediction for the prediction of clinical pregnancy in Freeze-thawed Embryo Transfer," *BMC Pregnancy and Childbirth*, 2022.
- [23] D. a. B. M. Cournapeau, "Scikit-Learn," 2007. [Online].
- [24] M. Alhamid, "Medium," 24 December 2020. [Online].
- [25] S. Kumar, "<https://towardsdatascience.com/understanding-8-types-of-cross-validation-80c935a4976d>," 13 september 2020. [Online].
- [26] V. Bolón-Canedo, N. Sánchez-Marcano and A. Alonso-B., "A review of feature selection methods on synthetic data," *Springer*, vol. 34, no. 3, p. 1007, 2012.
- [27] Jason B., "Machine Learning Mastery," 20 August 2020. [Online].
- [28] A. Wahab, "datasciencedojo," 25 October 2022. [Online].
- [29] N. Handayania, C. M. Louisa, A. Erwin, T. Aprilliana, A. A. Polim, B. Sirait, A. Boediono and I. Sino, "Machine Learning Approach to Predict Clinical Pregnancy Potential in Women Undergoing IVF Program," *Fertility and Reproduction*, pp. 77 - 87, 2022.
- [30] B. Ghojogh, M. Samad, S. Mashhadi, T. Kapoor, W. Ali, F. Karray and M. Crowley, "Feature Selection and Feature Extraction in Pattern AnalysisA Literature Review;" *arXiv*, 2019.
- [31] A. Bhandari, "Analytics Vidhya," 3 April 2020. [Online].

- [32] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr and J. O'Sullivan, "A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction," *Frontiers in Bioinformatics*, 2022.
- [33] Mohammed A., "FinsliQ," 12 2016. [Online].
- [34] M. Z. I. a. T. T. C. Chowdhury, "Variable selection strategies and its importance in clinical prediction modelling," *Family medicine and Community Health*, 2020.
- [35] K. D. and M. G. J., "Comparative Study on Heart Disease Prediction Using Feature Selection Techniques on Classification Algorithms," *Applied Computational Intelligence and Soft Computing*, 2021.
- [36] Avcontentteam, "<https://www.analyticsvidhya.com/blog/2023/05/recursive-feature-elimination/>," 17 May 2023. [Online].
- [37] P. Ippolito, "Towards Data Science," 10 October 2019. [Online].
- [38] YAAMEEN C., "Techie Cub," 30 September 2021. [Online].
- [39] A. Mostaar, M. Sattari, S. Hosseini and M. Deevband, "Use of Artificial Neural Networks and PCA to Predict Results of Infertility Treatment in the ICSI," *PubMed Central*, pp. 679-686, 2019.
- [40] A. Milewska, J. Dorota, C. Dorota, D. C. D. W. T. Jankowska, B. Acacio, R. Milewski and B. a. M. R. Acacio, "The Use of Principal Component Analysis and Logistic Regression in Prediction of Infertility Treatment outcome," *DE GRUYTER*, pp. 978-83, 2014.
- [41] O. Egwom, M. Hassan, J. Tanimu, M. Hamada and O. Ogar, "An LDA–SVM Machine Learning Model for Breast Cancer Classification †," *BioMedInformatics*, pp. 345-358, 2022.
- [42] Akash D., "Medium," 15 December 2018. [Online].
- [43] A. Gupta, "GeeksforGeeks," 18 May 2023. [Online].
- [44] C. Ellis, "Crushing the Data," 23 May 2022. [Online].

- [45] e. a. Min Xiana, "Automatic breast ultrasound image segmentation: A survey," *ELSEVIER*, vol. 79, pp. 340-355, 2018.
- [46] V. Rovei, p. Dalmasso and G. Gennarelli, "IVF outcome is optimized when embryos are replaced between 5 and 15 mm from the fundal endometrial surface; a prospective analysis 1184 cycles," *REPRODUCTIVE BIOLOGY AND ENDOCRINOLOGY*, 2013.
- [47] A. Firoozeh, F. Akhbari, M. Zamani, F. Ramezanali and R. Cheraghi, "Value of Endometrial Echopattern at HCG Administration Day in predicting IVF outcome," *Archives of Iranian Medicine*, 2017.
- [48] M. Ivanovski, The Role of Ultrasound in the Evaluation of Endometrial Receptivity Following Assisted Reproductive Treatments, 2012.
- [49] J. Zhao and Q. a. L. Y. Zhang, "the effect of endometrial thickness and pattern measured by ultrasonography on pregnancy outcomes during IVF-ET cycles," *Reproductive Biology and Endocrinology*, 2012.
- [50] M. Seonwoo, L. Byunghan and Y. Sungroh, "Deep Learning in Bioinformatics," *briefings in bioinformatics*, vol. 18, no. 5, pp. 851-869, 2016.
- [51] M. Xian, Y. Zhang, H. Cheng, F. Xu, B. Zhang and J. Ding, "Automatic Breast ultrasound image segmentation: Asurvey," *ELSEVIER*, vol. 10, no. 1016, pp. 340-355, 2018.
- [52] C. Ruey-Feng, W. Wen-Jie, M. Woo Kyung, C. Wei-Ming, L. Wei and C. Dar-Ren, "Segmentation of breast tumor in three-dimensional ultrasound images using three-dimensional discrete active contour model," *Ultrasound in Medicine & Biology*, pp. 1571-1581, 2003.
- [53] M. Gooding, S. Kennedy and J. Noble, "Volume Segmentation and Reconstruction from Freehand Three-Dimensional Ultrasound Data with Application to Ovarian Follicle Measurement," *Ultrasound in medicine & biology*, pp. 183-95, 2008.
- [54] B. Liu, H. Cheng, J. Huang, J. Tian, X. Tang and J. Liu, "Fully automatic and segmentation-robust classification of breast tumors based on local texture analysis of ultrasound images," *ELSEVIER*, vol. 43, no. 1, pp. 280-298, 2010.

- [55] L. W. and H. F., "Segmentation of Ultrasonic Images Based on Modified Chan-Vese," in *6th International Conference on Electronic, Mechanical, Information and Management*, 2016.
- [56] L. Shengfeng, W. Yi, Y. Xin, L. Baiying, L. Li, Z. L. Shawn, N. Dong and W. Tianfu, "Deep Learning in Medical Ultrasound Analysis: A Review," *ELSEVIER*, vol. 10, no. 1016, pp. 261-275, 2019.
- [57] S. Kaufmann, J. Eastaugh, S. Snowden, S. Smye and V. Sharma, "The application of neural Networks in predicting the outcome of in vitro fertilization," *ReasearchGate*, vol. 12, no. 7, pp. 1454-1457, 1997.
- [58] Y. Zhang, M. T. C. Ying, L. Yang, A. Ahuja and Z. Chen, "Coarse-to-Fine Stacked Fully Convolutional Nets for lymph node segmentation in ultrasound images," in *IEEE International Conference on Bioinformatics and Biomedicine*, Shenzhen, China, 2016.
- [59] J. Ma, F. Wu, J. Zhu, D. Xu and D. Kong, "A pre-trained convolutional neural network based method for thyroid nodule diagnosis," *ELSEVIER*, vol. 10, no. 1016, pp. 221-230, 2017.
- [60] K. Cha, L. Hadjiiski, R. K. Samala, H.-P. Chan and E. M. a. C. R. Caoili, "Urinary bladder segmentation in CT urography using deep-learning convolutional neural network and level sets," *US National Library of Medicine*, vol. 10, no. 1118, p. 1882–1896., 2016.
- [61] E. Smistad, A. Ostvik, B. O. Haugen and L. L., "2D left ventricule segmentation using deep learning," *IEEE*, pp. 1-4, september 2017.
- [62] A. Lemay, "Kidney Recognition in CT using YOLOv3," *arxiv*, 2019.
- [63] Y. Li, R. Xu, J. Ohya and H. Iwata, "Pregnant Uterine Ultrasound Image Segmentation by Encoding-Decoding Convolutional Neural Network," in *Proceedings of the Fifth IIEEJ International Workshop on Image Electronics and Visual Computing 2017*, Da Nang, Vietnam,, 2017.
- [64] N. Singhal, S. Mukherjee and S. Perrey, "Automated Assessment of Endometrium From Transvaginal ultrasound using deep learning snake," *IEEE*, 2017.

- [65] A. J. a. V. S. Vidhi Rawat, "Automated Techniques for the interpretation of fetal abnormalities: A review," *Applied bionics and biomechanics*, vol. 10, no. 1155, 2018.
- [66] P. N. S. C. S. L. Looney, "Fully automated, real-time 3D ultrasound segmentation to estimate first trimester placental volume using deep learning," *JCI insight*, vol. 3, no. 11, 2018.
- [67] K. K. W. L. S. S. M. Annupan R., "Automatic initialization of active contours and level set method in ultrasound images of breast abnormalities," *sciencedirect*, pp. 172-182, 2018.
- [68] M. Raghu, C. Zhang and J. a. B. S. Kleinberg, "Transfusion: Understanding Transfer Learning for Medical Imaging," in *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, Canada, 2019.
- [69] S. D. a. S. S. Balakrishna, "Automatic detection of lumen and media in the IVUS images using U-Net with VGG16 Encode," *arXiv*, 2018.
- [70] J. Torrents-Barrena, "Fully automatic 3D reconstruction of the placenta and its peripheral vasculature in intrauterine fetal MRI," *ELSEVIER*, vol. 54, pp. 263-279, 2019.
- [71] e. a. Yanni Su, "Automatic Detection and Classification of Breast Tumors in Ultrasonic Images Using Texture and Morphological Features," *The Open Medical Informatics Journal*, vol. 13, pp. 26-37, 2019.
- [72] C. Z. Z. L. J. Y. ,. D. Zhong Ch., "Automatic segmentation of ovarian follicles using deep neural network combined with edge information," *frontiers*, 2022.
- [73] X. Min, Y. Zhang, H. D. Cheng, F. Xu and B. Z. a. J. D., "Automatic Breast Ultrasound Image Segmentation: A Survey," *Elsevier*, pp. 340-355, 2018.
- [74] N. Mahmood, M. Razif and M. Gany, "Comparison between Median, Unsharp and wiener filter and its effect on ultrasound stomach tissue image segmentation for Pyloric Stenosis," *International Journal of Applied Science and technology*, pp. 218-226, 2011.
- [75] G. Manoj, T. Heena and C. Laxmi, "Performance Enhancement and Analysis of filters in Ultrasound Image Denoising," *Science Direct*, pp. 644-652, 2018.

- [76] C. Hyun, J. L. and S. K. a. S. P., "Speckle noise reduction in ultrasound images usind a discrete wavelet transform-based image fusion technique," *Bio-Medical Materials and Engineering*, pp. 1587-1597, 2015.
- [77] L. a. J. L. Baozhong, "Overview of image noise reduction based on non-local mean algorithm," in *MATEC Web of Conferences*, China, 2018.
- [78] I. NJEH, O. B. S. K. CHTOUROU and A. BEN HAMIDA, "SPECKLE NOISE REDUCTION IN BREAST ULTRASOUND IMAGES: SMU (SRAD MEDIAN UNSHARP) APPROCH," in *8th International Multi-Conference on Systems, Signals & Devices*, Tunisia, 2011.
- [79] H. T. Yu Sh., "A hybrid approach of tumor segmentation in ultrasound images based on Chan-Vese model and information theorem," in *IEEE International Conference on Signal Processing, Communications and Computing*, Kunming, China, 2013.
- [80] R. Saru Meena, R. Muthaiah, K. Kannan, J. Jaikanth, C. Panagiotis and C. Nachiappan, "A method to improve the computational efficiency of the Chan-Vese model for the segmentation of ultrasound images," *Biomedical Signal Processing and Control*, p. 1746, 2021.
- [81] Y.-Y. C. J.-S. Y. J.-Y. L. Z.-Y. C. Xiao-Wen L., "Update on thyroid ultrasound: a narrative review from diagnostic criteria to artificial intelligence techniques," *Chinese Medical Journal*, pp. 1974-1982, 2019.
- [82] J. Saeed, "A SURVEY OF ULTRASONOGRAPHY BREAST CANCER IMAGE SEGMENTATION TECHNIQUES," *Academic Journal of Nawroz University (AJNU)*, 2020.
- [83] B. W. Yunyun Y., "Split Bregman method for minimization of improved active contour," *Elsevier*, pp. 351-366, 2012.
- [84] D. T. W. J. X. S. B. W. Yunyun Y., "Split Bregman method based level set formulations for segmentation and correction with application to MR images and color images," *sciencedirect*, pp. 50-67, 2019.
- [85] Y. Sh. and T.-Z. Huang, "A hybrid approach of tumor segmentation in ultrasound images based on Chan-Vese model and information theorem," in *IEEE International Conference on Signal Processing, Communications and Computing*, 2013.

- [86] A. Biguri, "Split Bregman - Globally Convex Segmentation," [Online].
- [87] C. Tam, "Machine Learning towards General Medical Image Segmentation," Western University, ontario, 2020.
- [88] O. Ronneberger and P. a. B. T. Fischer, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *arxiv*, p. \, , 2015.
- [89] A. Qayyum, "Medical Image Analysis using Convolutional Neural Networks: A review," *journal Medical Systems*, vol. 10, no. 1007, 2018.
- [90] E. Smistad, A. Ostvik, B. O. Haugen and I. Lovstakken, "2D left ventricule segmentation using deep learning," *IEEE*, p. 10.1109/ULTSYM.20178092573, september 2017.
- [91] M. T. C. Y. L. Y. A. T. A. D. Z. C. Yizhe Zhang*, "Coarse-to-Fine Stacked Fully Convolutional Nets for lymph node segmentation in ultrasound images," in *IEEE International Conference on Bioinformatics and Biomedicine*, 2016.
- [92] M. Raghu, C. Zhang and J. a. B. S. Kleinberg, "Transfusion: Understanding Transfer Learning for Medical Imaging," in *33rdConferenceonNeuralInformationProcessingSystems(NeurIPS2019)*, Canada, 2019.
- [93] C. Balakrishna, S. Dadashzadeh and a. S. S., "Automatic detection of lumen and media in the IVUS images using U-Net with VGG16 Encode," Department of Computing Science, University of Alberta, {cbalakri, dadashza, soltanin}@ualberta.ca, Canada, 2018.
- [94] B. Liu, H. Cheng, J. Huang, J. Tian, X. Tang and J. Liu, "Fully automatic and segmentation-robust classification of breast tumors based on local texture analysis of ultrasound images.," *ELSEVIER*, vol. 43, no. 1, pp. 280-298, 2010.
- [95] R. Milewski and A. J. Milewski, "Classification issue in the IVF ICSI/ET data analysis," *Studies in logic, Grammar and Rhetoric*, pp. 75-85, 2012.
- [96] M. Durairaj and P. Thamilesevan, "Application of Artificial Neural Network for IVF data analysis and prediction," *Journal of Engineering, Computers & Applied Sciences*, pp. 11-15, 2013.

- [97] H. A. Guvenir, G. Misirli, S. Dilbaz, O. Ozdegirmenci, B. Demir and B. Dilbaz, "Estimating the chance of success in IVF treatment using a ranking algorithm," *Springer*, pp. 911-920, 2015.
- [98] A. J. Milewska, D. Jankowska, U. Cwalina, D. Citko, T. Wiesak, B. Acacio and R. Milewski, "Significance of Discriminant analysis in prediction of pregnancy in IVF treatment," *DE GRUYTER*, 2015.
- [99] R. Milewski, D. Jankowska, U. Cwalina, A. J. Milewska, D. Citko, T. Wiesak, A. Morgan and S. Wolczynski, "Application of Artificial Neural Networks and Principle Component Analysis topredict Results of Infertility treatment using the IVF method," *De Gruyter*, p. 33, 2016.
- [100] L. Shen, Y. Zhang, W. Chen and X. Yin, "The Application of Artificial Intelligence in Predicting Embryo Transfer Outcome of Recurrent Implantation Failure," *ORIGINAL RESEARCH article*, 2022.
- [101] K. Fu, Y. Li, H. Lv, W. Wu, J. Song and J. Xu, "Development of a Model Predicting the Outcome of In Vitro Fertilization Cycles by a Robust Decision Tree Method," *Front Endocrinol*, 2022.
- [102] D. Pham, C. Xu and J. Prince, "Currents Methods in Medical Image Segmentation," 2Department of Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, Maryland, 2000.
- [103] R. O. and a. T. B. Philipp Fischer, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *arxiv*, 2015.
- [104] S. Agrawal, "Analytics Vidhya," 20 july 2021. [Online].
- [105] M. S. Pathana, B. Avishek Nag, M. M. Pathan and S. Dev, "Analyzing the impact of feature selection on the," *arXiv*, 2022.
- [106] F. Tannock and Z. J. D. T., "A review of neural networks for statistical process control," *Springer*, vol. 9, no. 3, pp. 209-224, May 1998.
- [107] P. a. P. S. Bahad, "Study of AdaBoost and Gradient Boosting Algorithms for Predictive Analytics," in *International Conference on Intelligent Computing and Smart Communication 2019*, Singapore, 2019.

- [108] J. Wu, X. Chen, H. Zhang, L. Xiong, H. Lei and S. Deng, "Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimizationb," *Journal of Electronic Science and Technology*, pp. 26-40, 2019.
- [109] D. Dey, "GeeksforGeeks," 23 January 2023. [Online].
- [110] N. V. G. A. S. A. B. G. K. A. A. D. T. A. Nersisyan S, "ExhauFS: exhaustive search-based feature selection for classification and survival regression," *National Library of Medicine*, 2022.
- [111] A. Gupta, "GeeksforGeeks," 18 May 2023. [Online].
- [112] E. Lee, "Medium," 5 June 2019. [Online].
- [113] S. Talebi, "Medium," 17 March 2021. [Online].
- [114] R. Hemalatha, V. Vijaybaskar and T. Thamizhvani, "Performance Evaluation of Contour Based Segmentation," *Hindawi*, vol. 10, no. 1155, p. 8 pages, 2018.
- [115] Meigarom, "Medium," 28 March 2017. [Online].
- [116] H. Park, "An Introduction to Logistic Regression : from basic concepts to interpretation with particular attention to nursing domain," *J Korean Acad Nurs*, April 2013.
- [117] C. Manning, P. Raghavan and H. Schütze, "Introduction to Information Retrieval," Cambridge University Press, 2008.
- [118] Z. Jaadi, "Built In," 29 March 2023. [Online].
- [119] A. Dubey, "Medium," 15 December 2018. [Online].
- [120] N. Donges, "built in BETA," 22 July 2021. [Online].
- [121] A. Mustaqim, A. Sumarni, Y. Pristyanto and Y. Astuti, "The Effect of Recursive Feature Elimination with Cross-Validation (RFECV) Feature Selection Algorithm toward Classifier Performance on Credit Card Fraud Detection," in *International Conference on Artificial Intelligence and Computer Science Technology (ICAICST)*, 2021.

[122] A. a. G. M. Durkin, "30th European Symposium on Computer Aided Process Engineering," Italy, 2020.

[123] M. Babitha, "A Survey on the Machine Learning Techniques used in IVF Treatment to Improve the Success Rate," in *RTESIT*, 2019.

APPENDIX A

A.1 THE STATE OF THE ART ON IVF

Table 5-1: Some studies for the segmentation of an organ from ultrasound images

References	Year	Number of images/cases	Segmentation techniques	Additional comments
[52]	2003	-	Active contour model (snake)	Breast tumor segmentation is satisfied with 95% average match rate.
[53]	2008	-	Level set based method	Segmentation and reconstruction of the ovarian follicle, they discovered that both the computer-based measurement and the clinical measurement consistently underestimated the volume of follicular aspirate due to a clinical misinterpretation of the follicular boundary.
[94]	2010	112 images	SVM	Breast ultrasound segmentation with an accuracy = 93.75%
[55]	2016	-	Chan-Vese	Capture regional details from ultrasonic images
[58]	2016	80 images	Unet	lymph node segmentation (IoU=0.798)
[61]	2017	80 images	CNN	automatic segmentation to the left ventricle (Dice similarity= 0.86)
[67]	2018	180 cases	Active contour and level sets	Segmentation of ultrasound images of breast cancer (TPR= 92.78%)
[69]	2018	326 test images	Unet, VGG16-Unet	segment ultrasound scans of arteries (jaccard index = 0.798)
[70]	2019	44 images	SVM (Learning based segmentation) and Region growing	Classify the placenta then reconstruct the blood vessels with Dice coefficient=0.82
[71]	2019	132 cases	Normalized Cut (Ncut) algorithm and AP clustering	Breast tumor detection, segmentation and classification (Accuracy=93.8)
[72]	2022	1050 images	Unet with canny edge	Automatic segmentation of the ovarian follicle (jacquard index =0.679)

A.2 STUDIES ON THE CLASSIFICATION AND PREDICTION THE OUTCOME OF IVF AND ICSI USING MACHINE LEARNING TECHNIQUES

TABLE A-2: Studies on the classification and prediction the outcome of IVF and ICSI using machine learning techniques.

Authors	Sample Size	Machine learning technique	Features selection technique	Selected features	Findings
[10]	455 patients	ANN	T-test	Age, Previous live births, Previous abortions, Duration of infertility, Diagnosis, Sperm motility, Number of eggs recovered, Number of viable eggs, Proportion of viable eggs, Proportion of fertilized eggs, proportion of embryo cleaved, Number of embryos transferred, Freezing.	Accuracy = 59%
[11]	2275 case study	NB	No feature selection technique	Woman age, infertility factor, treatment protocol, Follicular stimulating hormone dosage, Peak estradiol level, early cleavage morphology, early cleavage time, number of cells, Nucleus characteristic, Fragmentation rate, Equality of blastomeres, Appearance of cytoplasm, transfer day, Physician performing embryo transfer, Difficulty of transfer.	Sensitivity = 64.4% Specificity = 30.6%
[95]	1445 patients	RF, SVM, K-fold cross validation	No feature selection technique	108 clinical features.	Highest accuracy with RF = 79%. (SVM= 71.4%)
[96]	250 patients	ANN	No feature selection	Endometriosis, tubal factors, and follicles in the ovaries, and the physiological factors (27 attributes)	Accuracy = 73%
[97]	1456 patients	SERA, NB, RF	No feature selection	64 independent features => 52 for female and 12 for male (table)	SERA (Accuracy=0.844) NB (Accuracy=0.783) RF (Accuracy =0.792)
[98]	610 patients	DA	No feature selection	Female age, number of follicles before retrieval, number and maturity of retrieved oocytes (GV, MI, MII, MII*), number and quality of embryos (2PN, NEF 2PB, non2PN, cleavage on day3, >=7 cells on	Sensitivity= 0.512 Specificity= 0.74

				day3, number blastocysts on day5 and day6), number of transferred embryos, and number of pregnancies.	
[12]	2453 embryos transferred using ICSI	NB, SVM, DT, KNN, RBF, MLP	Information gain feature weighting and forward feature selection methods	18 clinical features women and man factors/ Age of women, gravidity, infertility factor, treatment protocol, utilized sperm, duration of stimulation, FSH amount, Peak E2 level, Endometrium thickness, Early cleavage inspection time, Early cleavage morphology, Number of cells, Nucleus characteristics, Fragmentation rate, Equality of blastomeres, Appearance of cytoplasm, Thickness of zona pellucida, Transfer Day.	NB: Accuracy = 80.4% Sensitivity = 63.7%
[99]	1995 patients	PCA-ANN, ANN	Chi square	Age, number of cells and embryos at various development stages, and sperm characteristic, the type of IVF method, causes of infertility and type of applied ovulation stimulation.	ANN : AUC = 0.663 ANN+PCA : AUC = 0.666
[14]	486 patients	SVM, RPART, RF, Adaptive Boosting, 1NN	No Feature selection	Age of woman, Age of man, Body mass index, Secondary, fertility, Tubal factor, Pelvic factor, Ovulatory factor, Uterine factor, Male factor, Infertility duration, Experience of IVF treatment, Sperm count, Sperm morphology, Sperm motility, Follicle stimulating hormone, Anti-mullerian hormone, Antral follicle counts, Number of gonadotropin ampoules, Number of follicles in ultrasound, Serum E2 level on the day of hCG administration, Number of retrieved oocytes , Number of oocytes of GV quality, Number of oocytes of MI quality, Number of oocytes of MII quality, Type of treatment, Embryo grade, Number of developed embryos, Embryo transfer day, Number of transferred embryos	RF and RPART outperform the other comparable methods. AUC-ROC = 84.23 and 82.05%, respectively.
[16]	1048 patients	MMLP, SVM, CART, RF.	Hill climbing wrapper algorithm to select the best subset of features	25 attributes Age, indication of fertility factor, Antral follicle counts, NbreM2, method of sperm collection, chamotte, Fertilization rate in vitro, Follicles on day 14, and Embryo transfer day.	SVM: Accuracy = 98.38%
[17]	401 patients	LR, RF, SVM	Pearson's correlation	age, body mass index (BMI), endometrial thickness (EMT) on the day of progesterone treatment, good-quality embryo rate (GQR), and type of infertility (primary or secondary), serum estradiol level (E2) on the day of embryo transfer, and serum progesterone level (P) on the day of embryo transfer.	RF: Accuracy = 0.61.

[5]	6071 cycles	SVM, XGBoost , LR, RF, NB, LDA	No feature selection	The collected demographic and clinical variables comprised women's ages, source of infertility (female factor, male factor, combined male-female factor infertility, unexplained), infertility type (primary, secondary), body mass index (BMI), infertility duration (years), number of previous abortions, polycystic ovary syndrome (PCOS), number of previous IVF attempts, total number of retrieved oocyte, number of injected oocytes, number of embryos, number of transferred embryos, spermogram, fertilization rate after intracytoplasmic sperm injection (ICSI), number of two-pronuclear embryos to number of metaphase II (MII) oocytes (2PN/MII ratio), and data on embryo quality (number of compact, blastocysts, grade A, grade AB, early blastocysts, A compact, and AB compact), as well as the day of the embryo transfer (ET)	RF: Accuracy = 0.81
[18]	771 patients	LR, RF, XG BOOST, SVM	No feature selection	Age of the women and the man, Number of embryos with classification A, B and C, total number of viable cleavage stage embryos, number of blastocysts, total of fertilized pro-nuclei, Antral follicle count, sperm source, Number of follicles greater than 17 mm, Number of MII oocyte, indication for IVF, Number of type A embryos transferred, Number of oocytes, total dose of gonadotropins, Number of embryos transferred, Number of IF oocytes, Catheter model used for embryo transfer.	RF: AUC= 0.73
[6]	17,288 cycles	LR & RF	T-test	Male age, women age, duration of infertility, number of IVF cycles performed, Number of oocytes retrieved, number of embryo transferred, the total number of embryo of frozen embryos, cause of infertility, Fertilization method, Micromanipulation technique, source of sperm and oocytes, the use of fresh/freeze-thaw, ovarian hyperstimulation syndrome, ovarian stimulation protocole.	RF: AUC=0.72 Accuracy = 64.7%
[100]	45,921 cycles	RF, GBDT, AdaBoos t, MLP	No feature selection	44 features/ age, date of the last pregnancy, type of infertility, cause of infertility, stimulation used. Specific treatment type, elective single embryo transfer, fresh cycle, frozen cycle, eggs thawed, fresh eggs collected, fresh eggs stored, total eggs mixed, eggs mixed with partner sperm, total embryos created, eggs micro-injected, embryos from eggs micro-injected, total embryo thawed, embryos transferred from eggs micro-injected, embryos stored for	Single Embryo Transfer => GBDT with 85.06% Double Embryo Transfer => AdaBoost with 76.16%

				use by patient, date of eggs collection, date of egg thawing, date of egg mixing, date of embryo thawing, date of embryo transfer.	
[22]	2457 cycles	LR, Nomogram		Female age, BMI, type of embryo transferred, number of embryos transferred, number of sub endometrial blood flow, RI, PI, level of baseline FSH, The level of AMH, proportion of blastocyst, endometrial thickness on transplantation day, VI, FI and the number of blood flow branches of endometrial and sub-endometrial blood	AUC=0.699
[29]	4.570 IVF cycle	DT, RF, GB	Genetic Algorithm	42 variables/ Age, stimulation method, number of previous failed IVF, type of infertility , duration of infertility, history of miscarriage, Female BMI, IVF indication (male and female factors), Female prognosis, Basal FSH, Basal LH, Basal estradiol, Basal progesterone, AMH, AFC, estradiol level on trigger day, Progesterone level on trigger day, Type of ganadropine, starting dose of ganadropine, type of suppression drug, type of maturation trigger drugs, Number of oocyte(s) retrieved Number of mature oocyte(s) following injection, Maturation rate (%)- Sperm quality, Number of fertilization(s), Number of cleavage(s) Number of top-quality cleavage(s), Number of blastocyst(s), Number of top-quality blastocyst(s), Day of embryo transfer, Number of top-quality ET(s), 337Total number of embryo(s) transferred, All top-quality, Mix quality ET, Female smoking status, Male smoking status, Male alcohol drinking history,	Both the decision tree and random forest showed similar performance that was much better than the gradient boost. Accuracy= 0.62
[101]	37062 cycles	GBDT	Variables were selected through the significance of 1-way analysis between the clinical pregnant group and the non-pregnant group.	Age, number of 2 PN, Number of oocytes retrieved, Endometrial thickness, Basal AMH, Number of embryo transfer, stimulation days, LH on day HCG, Fertilization procedure, Embryo culture time, Duration of infertility, Age of man, Blastulation, BMI, History of ART, Basal LH, AFC, E2 before gonadropine, Treatment strategy, Basal E2.	AUC = 0.704

A.3 ULTRASOUND MEASUREMENTS

Basically, the high resolution of ultrasound machine (WS80A) from "Samsung, Medison" enables us to examine clearly the different implantation aspects: endometrial/ myometrium volume and endometrial/ myometrium mean grey. Pulsed and color Doppler is used to indicate different variables of uterine and endometrial/ myometrium perfusion which is VI, FI, VFI that are also used as receptivity factors. The parameters are measured by utilizing three sections.: coronal, sagittal and transverse of 3D sonography. A shell histogram was computed and displayed the gray value distribution of an object's 2D and Power Doppler pictures for which VOCAL was executed. The shell histogram allowed to compute Mean Gray (MG), Vascularization Index (VI), FI, and Vascularization Flow Index (VFI). The measurements of this part is done for the classification work.



Figure 5A-5-1 : Image of 3D ultrasound

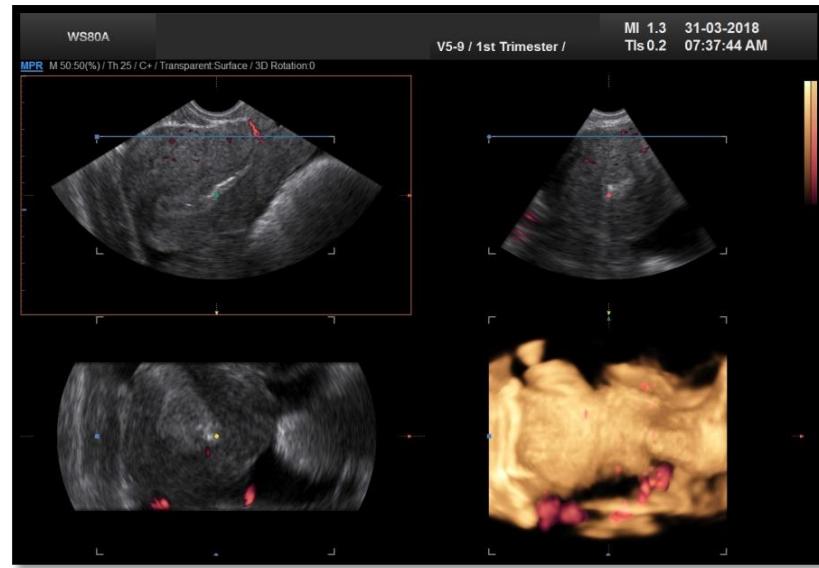


Figure A-5-2: Image of pulsed and color Doppler

Endometrial volume: the best method used to calculate the endometrial volume is the Virtual Organ Computer-aided Analysis (VOCAL) by applying the rotational techniques with 30° axes and take 6 endometrial slices if the measurement is manual.



Figure A-5-3: Image shown the measurement of endometrial volume and the endometrial mean grey.

Endometrial Mean Grey: After the measurement of the endometrial volume, we can display the histogram of the endometrial mean grey.

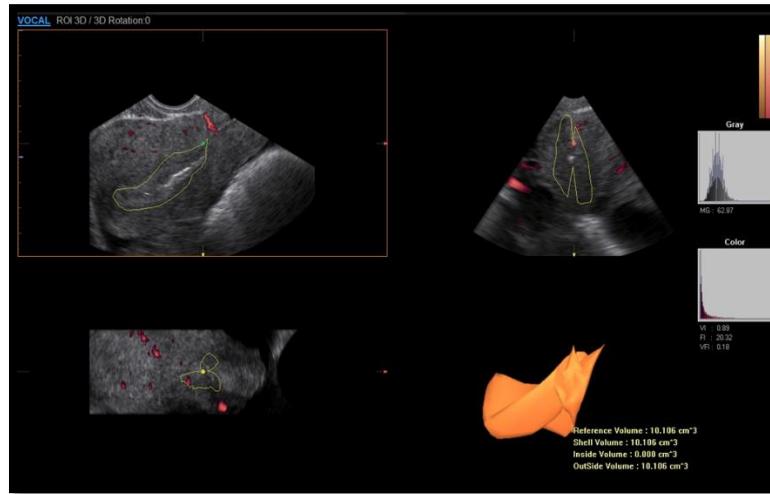


Figure A-5-4: This image represents the measurement of endometrial power Doppler indexes.

Endometrial thickness: is measured from echogenic border to echogenic border across the endometrial cavity on a sagittal midline image.



Figure. A -5-5: Measurement of endometrial thickness

Endometrial pattern: endometrial pattern is defined as the relative echogenicity of the endometrium as demonstrated on a longitudinal ultrasound axis. Two types of echogenicity exist: triple line and homogeneous. The triple line configuration means that the endometrium contains a hyperechoic (usually displayed as light) line in the middle surrounded by two more hyperechoic (darker) lines.



Figure A-5-6: A triple line endometrium represents in this image.

Myometrium volume: Furthermore, the VOCAL technique can also be used to measure the volume of the myometrium. In this case, six slices are taken for the myometrium using VOCAL manual.

Myometrium mean grey: Similar to how the mean grey of the endometrium can be represented, we have the capability to visualize the histogram of the mean grey values of the myometrium.

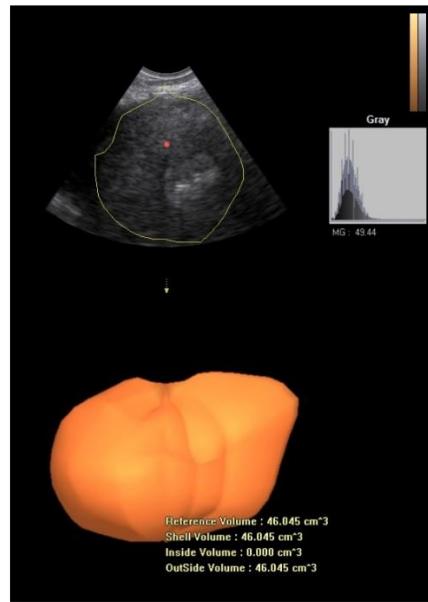


Figure A -5-7: Image shown the measurement of myometrium volume and the myometrium mean grey.

Myometrium VI, FI and VFI: These indexes represent the vascularization indexes, flow indexes, and volume flow indexes respectively, but in this case

we represent the power Doppler indexes for the myometrium. To measure the three indexes, we take 3D Doppler image and repeat the measurement of volume, then VOCAL software should calculate the index of vascularization.

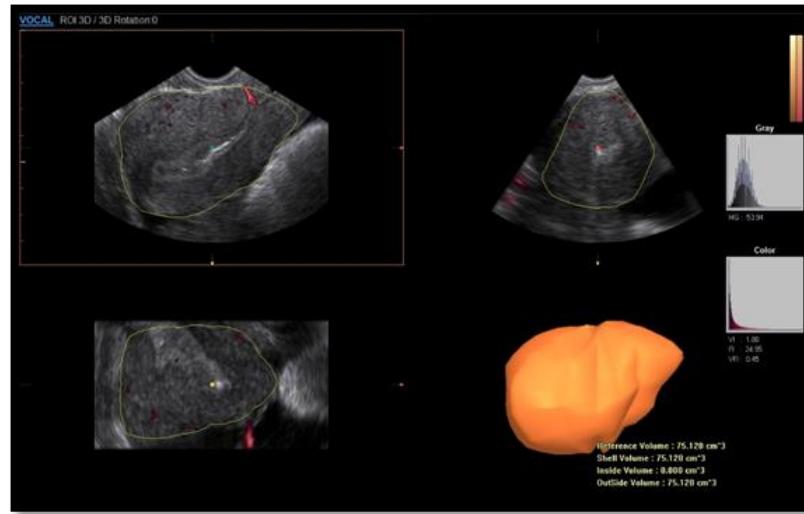


Figure A-5–8: Image represents the measurement of myometrium power Doppler indexes.

A.4 SEGMENTATION USING LEARNING BASED APPROACHES:

A.4.1 Segmentation using classical machine learning techniques:

Image segmentation can be seen as a classification task, where the goal is to categorize pixels or superpixels into distinct categories. As a result, it is common to utilize machine learning methods for image segmentation purposes [102]. As you can see” Table 1” in the Appendix, various medical image segmentation techniques have utilized learning-based approaches. Deep learning, a subset of machine learning, is regarded as a representation learning method capable of directly processing and autonomously acquiring mid-level and high-level abstract features from raw data, such as ultrasound (US) images. It has the capability to automate US image analysis tasks, including lesion/nodule classification, organ segmentation, and object detection [56]. Following the utilization of SVM, RF, and NN, along with Gray-Level Co-occurrence Matrix (GLCM) as feature extractors for endometrium segmentation, the obtained results were found to be unsatisfactory. Although the precision for endometrium ranges between 0.634 and 0.872, the F1 and Recall scores for endometrium classification were extremely low, ranging from 0.02 to 0.066. As a result, it becomes necessary to employ deep learning techniques for more accurate segmentation results.

A.4.2 Segmentation using deep Learning approaches:

Deep learning methods typically require a large database, but in this case, the available database is not extensive, making deep learning methods seemingly unsuitable at first glance. Although it is possible to use simulation models to generate artificial data for training, the results are likely to be subpar due to the high variability in ultrasound images. It could have been interesting to explore GAN (Generative Adversarial Network) generators, but time constraints prevented us from doing so. Nonetheless, we did experiment with a deep learning method, specifically U-Net, using a Field II generator. As expected, the results were not satisfactory.

The segmentation of organs and other substructures in medical images plays a crucial role in enabling quantitative analysis of clinical parameters such as volume and shape. This analysis is particularly important in fields like cardiac and brain analysis. Among the applications of deep learning in medical imaging, segmentation has received significant attention, resulting in a wide range of methodologies being proposed. One of the most prominent and well-known convolutional neural network (CNN) architectures for medical image analysis is U-net, which was introduced by Ronneberger [103]. U-net was specifically developed for biomedical image segmentation, with the aim of predicting the class for each pixel in the image. It is an extension of the Fully Convolutional Network architecture, modified to improve segmentation performance in medical imaging applications [89].

U-net offers several advantages, including:

- Integration of Location and Context: U-net effectively combines location information and features obtained from the down-sampling path with contextual information from the up-sampling path. This integration enables the model to capture both localized details and contextual understanding, which are crucial for generating accurate segmentation maps.
- Flexibility in Input Image Sizes: Unlike architectures that rely on dense layers, U-net does not have this restriction. Consequently, images of various sizes can be used as input, as the learning parameters in the convolution layers are only the kernel weights, which are independent of the image size.
- Effective Data Augmentation: In domains like biomedical segmentation, where the availability of annotated samples is often limited, U-net leverages the use of extensive data augmentation techniques. By augmenting the existing data, the model can learn from a larger and more diverse dataset, enhancing its ability to generalize and perform well on unseen data.

In a study referenced as [90], an automatic segmentation technique achieved a dice similarity coefficient of 0.86 specifically for left ventricle segmentation on a large ultrasound dataset. Additionally, in the task of lymph

node segmentation, the U-net architecture demonstrated superior performance compared to other state-of-the-art deep learning methods, as mentioned in [91]. With the rise of deep learning, transfer learning has become a crucial component in various applications, particularly in medical imaging. The current standard practice involves utilizing existing architectures designed for natural image datasets like ImageNet, along with their pretrained weights, and fine-tuning the models using medical imaging data [92]. In this study, the authors explored the effectiveness of Convolutional Neural Networks, including U-Net, combined with VGG16 encoder for segmenting ultrasound scans of arteries [93] .

Table A-3: Accuracy and the Mean IoU between different models in segmentation

Model	Mean IoU	Loss
Unet	0.47327	0.0145
VGG16-Unet	0.55265	0.0029
ResNet50-Unet	0.68787	0.0023
SegNet	0.61169	0.0016

As it is shown in the figures 3 to 7, ResNet50_Unet which is a combination between U-net model and ResNet50 in the encoder part. This model gives the best segmentation of endometrium comparing to the mask and comparing to other models like simple U-net, the combination between VGG16 and Unet, and finally the SegNet model.



Figure A-9: Initial Image.

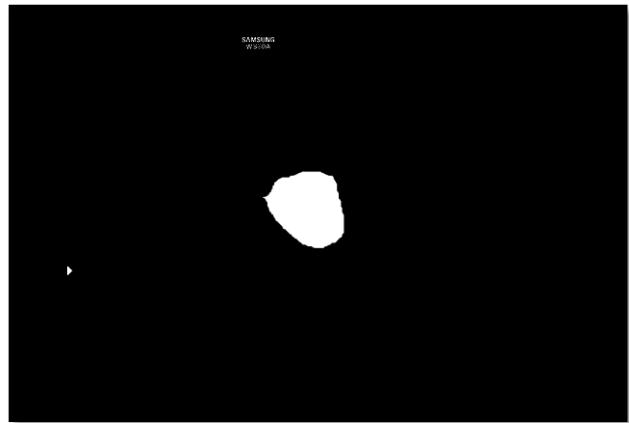


Figure A-10: Mask using GIMP software.



Figure A-11: Output of SeeNet.



Figure A-12: Output of vgg16_Unet.



Figure A-13: Output of Unet.

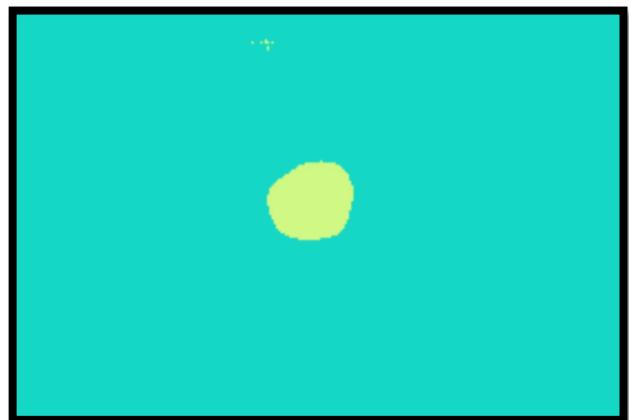


Figure A-14: Output of ResNet50-Unet.

APPENDIX B

GLOSSARY

Terms	Definition
Accuracy	Accuracy simply evaluates how often the classifier guesses accurately. Accuracy can be defined as the ratio of the number of correct predictions to the total number of predictions [104].
ANOVA	The ANOVA method is a sort of F-statistic known as an ANOVA f-test in this context. It is a univariate statistical test that compares each feature to the target feature to determine whether or not there is a statistically significant link between them [105]
Artificial Neural Network (ANN)	ANN is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. Artificial neural network is a computing system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs. The main characteristics of neural networks are that they have the ability to learn complex nonlinear input output relationships, use sequential training procedures, and adapt themselves to the data. Although the mathematics involved with neural networking is not simple, a user can rather easily gain at least an operational understanding of their structure and function. NNs can extract regularities and recognize patterns in sets of data. They can learn by example, and are able to capture the knowledge contained within a training data set and apply it to other data sets [106].
Adaptive Boosting (AdaBoost)	AdaBoost, also known as Adaptive Boosting, is a Machine Learning technique that is utilized as an Ensemble Method. AdaBoost's most commonly used estimator is decision trees with one level, which is decision trees with only one split. These trees are often referred to as Decision Stumps. This algorithm constructs a model and assigns equal weights to all data points. It then applies higher weights to incorrectly categorized points. In the following model, all points with greater weights are given more weight. It will continue to train models until a smaller error is received [107].

AUC-ROC	The Receiver Operator Characteristic (ROC) is a probability curve that plots the TPR (True Positive Rate) against the FPR (False Positive Rate) at various threshold values and separates the ‘signal’ from the ‘noise’. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes. From the graph, we simply say the area of the curve ABDE and the X and Y-axis [104].
Bayesian optimization	Bayesian optimization, which involves tuning hyperparameters with Bayesian logic, aids in reducing the time required to get an ideal parameter set. It increases test set generalization task performance. It operates by taking past hyperparameter combinations into account when deciding the next set of hyperparameters to assess [107].
Bagging Classifier (BC)	A BC is an ensemble meta-estimator that fits base classifiers on random subsets of the original dataset, then aggregates their individual predictions (through voting or averaging) to generate a final prediction. A meta-estimator of this type is often used to reduce the variance of a black-box estimator (for example, a decision tree) by introducing randomization into its building mechanism and then constructing an ensemble from it. Bagging, also known as Bootstrap aggregation, is an ensemble learning technique that helps machine learning algorithms increase their performance and accuracy. It is used to cope with bias-variance trade-offs and reduces a prediction model's variance. Bagging prevents data overfitting and is utilized in both regression and classification models, particularly decision tree techniques [108].
Chan-Vese Model	This model was initially proposed by Tony Chan and Luminita Vese in 1999. It is based on level sets and involves partitioning the image, where the interior region of the contour represents the object to be segmented, while the exterior region represents the background. These two regions are assumed to be homogeneous and distinguishable. They are described by statistical moments, parameters, or probability density. These features, known as region descriptors, are used to define an energy functional. The objective of this model is to minimize this energy functional in order to evolve the contour. Thus, we can express this energy functional as follows: $\delta\Gamma ((I - \mu_{int})^2 - (I - \mu_{ext})^2)N$ <p>Here, μ_{int} and μ_{ext} represent the mean values of the interior and exterior regions, respectively, and I represents the narrow band around the contour Γ.</p> <p>The figure below illustrates the working principle of the model.</p>

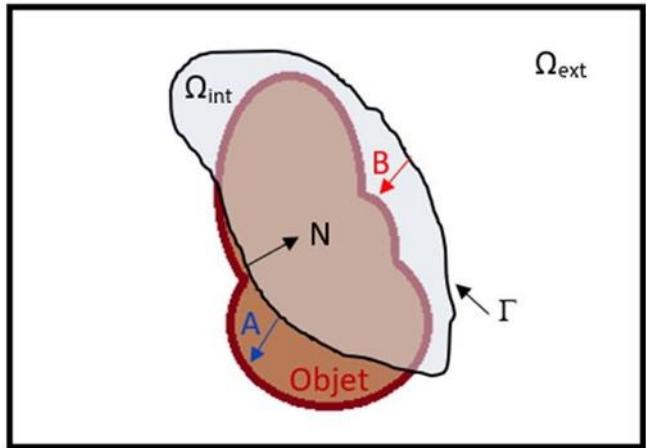


Figure represents a demonstration of the principle of evolution in the Chan-Vese model.

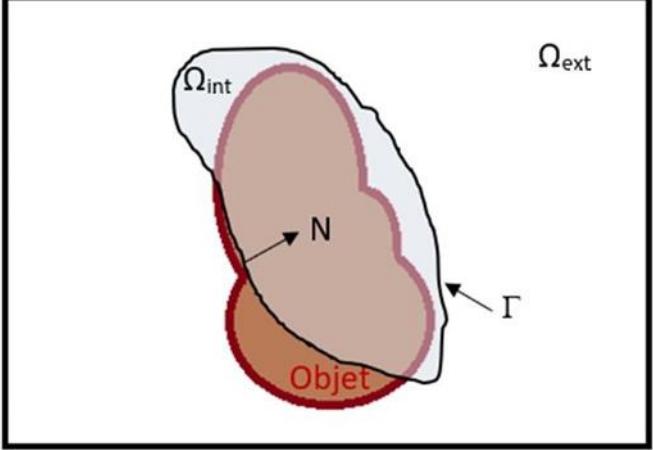
In the case of the Chan-Vese model, we focus on the force applied at each point of the contour. Taking point, A as an example, it has an intensity similar to that of μ_{int} and is opposite to the normal vector N . Conversely, at point B , we observe the opposite behavior. As a result, in the next iteration, we will see an expansion of the curve around point A and a contraction around point B .

This concept is based on the principle that the curve seeks to position the contour point in the region that is closest to it in terms of mean values.

The model calculates the means at each iteration and evolves the curve until it no longer changes. The only drawback of this model is that it falls under variational models, which can be relatively slow in terms of computations. Thus, we have explored whether there are optimization methods available for these models.

Confusion Matrix	<p>Confusion Matrix is a performance metric for machine learning classification tasks with two or more classes as output. It is a table that contains a mix of predicted and actual data.</p> <p>The confusion matrix, as shown in the Table below, is defined as a table that is frequently used to explain the performance of a classification model on a set of test data for which the true values are known [104].</p> <p>Table Shows a confusion matrix and their components [104]</p> <table border="1"> <thead> <tr> <th colspan="2"></th><th colspan="2">Actual Values</th></tr> <tr> <th colspan="2"></th><th>Positive (1)</th><th>Negative (0)</th></tr> <tr> <th rowspan="2">Predicted Values</th><th>Positive (1)</th><td>TP</td><td>FP</td></tr> </thead> <tbody> <tr> <th>Negative (0)</th><td>FN</td><td>TN</td></tr> </tbody> </table> <p>True Positive: We predicted positive and it's true. In the image, we predicted that a woman is pregnant and she actually is.</p> <p>True Negative: We predicted negative and it's true. In the image, we predicted that a woman is not pregnant and she's actually not pregnant.</p> <p>False Positive (Type 1 Error) - We predicted positive and it's false. In the image, we predicted that a woman is pregnant but she's actually not pregnant.</p> <p>False Negative (Type 2 Error) - We predicted negative and it's false. In the image, we predicted that a woman is not pregnant but she actually is pregnant [104].</p>			Actual Values				Positive (1)	Negative (0)	Predicted Values	Positive (1)	TP	FP	Negative (0)	FN	TN
		Actual Values														
		Positive (1)	Negative (0)													
Predicted Values	Positive (1)	TP	FP													
	Negative (0)	FN	TN													
Exhaustive Feature Selection (EFS)	It is the most straightforward approach for feature selection is an exhaustive search: one can go over all possible feature															

	combinations and pick up the model with the highest accuracy [109].
Extra Tree Importance	The Extra Tree Classifier is an ensemble technique that seeds many tree models built at random from the training dataset and sorts out the most voted-for features. Rather than a bootstrap replica, it fits each decision tree to the entire dataset and chooses a split point at random to separate the nodes. The constituent tree models determine the split that results in the most homogeneous sub-child based on all factors available in the dataset. This reduces variance and makes the model less susceptible to overfitting [110].
Extreme Gradient Boosting (XGBoost)	XGBoost is a more regularized version of Gradient Boosting. XGBoost employs sophisticated regularization (L1 & L2) to increase model generalization. When compared to Gradient Boosting, XGBoost provides superior performance. It has a very rapid training time and can be parallelized across clusters.
F1 Score	It gives a combined idea about Precision and Recall metrics. It is maximum when Precision is equal to Recall. F1 Score is the harmonic mean of precision and recall [104].
Gradient boosting Algorithm (GB)	GB is a machine learning boosting method in which a prediction model is generated by combining weaker prediction models. The gradient boosting algorithm is made up of three parts. The loss function varies depending on the situation at hand, as do weak learners used for prediction and the additive model, which combines trees via a gradient descent process. By merging the next model with the prior ones, the approach predicts the best possible model while minimizing error [111].
Grid Search	Grid search is the most basic hyperparameter tuning approach. Essentially, we partition the hyperparameter domain into a discrete grid. Then, using cross-validation, we try every possible combination of values from this grid, determining various performance measures. The best combination of values for the hyperparameters is the grid point that maximizes the average value in cross-validation. Grid search is an exhaustive method that searches all possible combinations to locate the best point in the domain. The main disadvantage is that it is quite slow. Checking every combination of the space takes a lot of time, which isn't always available. Remember that every point in the grid requires k-fold cross-validation, which necessitates k training steps. As a result, tuning a model's hyperparameters in this manner can be quite complex and costly. Grid search, on the other hand, is a great way to find the optimal combination of hyperparameter values [112].
Incremental Component Analysis (ICA)	ICA is a linear dimensionality reduction method that takes a mixture of independent components as input data and attempts

	<p>to accurately identify each of them (removing any extraneous noise). Two input features are said to be independent if their linear and nonlinear dependences are both equal to zero [113].</p>
Intersection over Union (IoU)	<p>The Intersection-Over-Union (IoU), also known as the Jaccard Index, is one of the most commonly used metrics in semantic segmentation and for good reason. The IoU is a very straightforward metric that's extremely effective. This metric ranges from 0–1 (0–100%) with 0 signifying no overlap and 1 signifying perfectly overlapping segmentation [114].</p>
Level-set algorithms	<p>This algorithm, developed in the late 1980s, is based on an intrinsic and Eulerian representation of the evolving curve. The evolution of this closed curve Γ is governed by a partial differential equation.</p> $\frac{\delta\Gamma(s)}{\delta t} = F(s).N$ <p>It relies on N, which represents the inward normal to the closed curve, and F, a velocity function based on geometric quantities that are independent of the curve's parameterization.</p>  <p>Demonstration of the initialization of the level-set algorithm. The curve Γ is regarded as the level set 0 line of a function, typically referred to as the signed distance function φ. This function effectively assigns negative values to the region inside the curve Ω_{int} and positive values to the region outside the curve Ω_{ext}.</p> <p>In this case, the geometric quantities in the evolution equation can be expressed in terms of the signed distance function. Furthermore, by differentiating the relation $\varphi(\Gamma) = 0$, we obtain the following evolution equation:</p> $\delta\varphi = F \nabla\varphi $ <p>The objective of the level-set algorithm is to construct the signed map based on the initial curve corresponding to level 0 and evolve it using the above equation until convergence, ultimately extracting the level 0 curve.</p>

Linear Discriminant Analysis (LDA)	LDA aims to maximize the distance between the mean of each class and minimize the spreading within the class itself. LDA uses therefore within classes and between classes as measures. This is a good choice because maximizing the distance between the means of each class when projecting the data in a lower-dimensional space can lead to better classification results [115]
Logistic regression (LR)	It analyzes the relationship between multiple independent variables and a categorical dependent variable, and estimates the probability of occurrence of an event by fitting data to a logistic curve [116].
Mean Squared Error (MSE)	MSE quantifies the overall differences between the original image and the denoised image. It is widely employed to measure the total amount of variation between the two images. Higher MSE values indicate larger differences, while lower values suggest smaller discrepancies between the original and filtered images. Identical images yield an MSE of zero [74].
Multi-Layer Perceptron (MLP)	A fully connected multi-layer neural network is called MLP.
Naïve Bayes Classifier (NB)	A Naive Bayes Classifier is an algorithm that uses Bayes' theorem to classify objects. Naive Bayes classifiers assume strong, or naive, independence between attributes of data points. These classifiers are widely used for machine learning because they are simple to implement [117].
Peak Signal-to-Noise Ratio (PSNR)	PSNR measures the performance of speckle noise reduction by comparing the maximum possible power of the signal to the noise in the image. Higher PSNR values indicate better filtering performance and improved image quality as a general guideline, a higher PSNR value (e.g., above 30 dB) is often considered indicative of good image quality [74].
Principle Component Analysis (PCA)	<p>PCA is one of the most used linear dimensionality reduction techniques. When using PCA, we take as input our original data and try to find a combination of the input features which can best summarize the original data distribution to reduce its original dimensions. PCA is able to do this by maximizing variances and minimizing the reconstruction error by looking at pair wised distances. In PCA, our original data is projected into a set of orthogonal axes and each of the axes gets ranked in order of importance.</p> <p>PCA is an unsupervised learning algorithm, therefore it doesn't care about the data labels but only about variation. This can lead in some cases to misclassification of data. The problem with the PCA method is that it does not preserve the original features which is sometimes necessary. In order to preserve features numerous methods based on swarm intelligence algorithms were proposed as it was described in the previous section. Besides selecting the appropriate feature set, classification accuracy depends on the classifier. Each classification method has some parameters that affect the accuracy of the created model [118].</p>

Precision	Precision explains how many of the correctly predicted cases actually turned out to be positive. Precision is useful in the cases where False Positive is a higher concern than False Negatives. Precision for a label is defined as the number of true positives divided by the number of predicted positives [104].
Random Forest Importance	Random forests are made up of 4 to 12 hundred decision trees, each of which is constructed using a random extraction of observations from the dataset and a random extraction of features. Because not every tree sees all of the features or all of the data, the trees are de-correlated and hence less prone to over-fitting. Each tree is also a series of yes-no questions based on one or more attributes. The three separates the dataset into two buckets at each node, each of which contains observations that are more similar among themselves and dissimilar from those in the other bucket. As a result, the significance of each attribute is determined by how "pure" each of the buckets is [119].
Random Forest (RF)	<p>Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result.</p> <p>Put simply: random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. One big advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems [120].</p>
Random search	Random search is similar to grid search in that it tests a randomly selected fraction of the grid's points rather than all of them. The faster but less precise the optimization, the smaller this subgroup. The richer the dataset, the more precise the optimization, but the closer it is to a grid search. When you have numerous hyperparameters with a fine-grained grid of values, random search is a great solution. We can generate a relatively excellent set of hyperparameter values by using a subset of 5-100 randomly picked points. It is unlikely to be the optimal point, but it may be a decent group of values that leads to a good model [112].
Recall (Sensitivity)	Recall explains how many of the actual positive cases we were able to predict correctly with our model. It is a useful metric in cases where False Negative is of higher concern than False Positive. Recall for a label is defined as the number of true positives divided by the total number of actual positives [104].
Recursive Feature Elimination Cross-Validation (RFECV)	The algorithm employed at this stage is Recursive Feature Elimination with Cross-Validation (RFECV). The RFECV algorithm will eliminate redundant and insignificant features that have minimal impact on classification outcomes. Along with retaining strong independent attributes to increase model

	performance. To rank qualities, this technique uses an iterative procedure. This algorithm begins by creating a model of all the attributes in the dataset. It ranks the attributes according to their impact on the classification. In addition, the RFECV algorithm will delete the attribute with the least correlation to the classification results and repeat this step to recalculate the attribute ranking. The RFECV algorithm's main goal is to pick the best number of characteristics via automatic cross-validation. This algorithm is an enhancement to the RFE algorithm [121].
Signal-to-Noise Ratio (SNR):	SNR is commonly used to evaluate speckle reduction in the presence of multiplicative noise. It calculates the ratio between the original image and the denoised image. Higher SNR values indicate a stronger filtering effect and higher quality of the filtered image. A higher SNR value (e.g., above 20 dB) is typically desired for high-quality images [74].
Support vector Machine (SVM)	SVM is most powerful classification algorithms in terms of predictive accuracy. They are based on strong mathematical foundations and statistical learning theory. They can classify both linear and nonlinear data. SVMs were initially designed for two-class problems but later used for multi-class problem also. The basic principle of SVM is to find an optimal hyper plane with a maximum distance to the closest point of the two classes. A set of tuples that is closest to the optimal hyperplane is called a support vector. SVM uses these support vectors to find the optimal hyperplane. Finding the optimal hyperplane provides a linear classifier, whereas to classify nonlinear data, the original training data is transformed into higher dimension using nonlinear kernel functions such as polynomial, radial, Gaussian, sigmoid etc. SVM works on the principal that data points are classified using a hyper plane which maximizes the separation between data points and the hyper plane is constructed with the help of support vectors. SVMs can be applied for numeric prediction as well as classification [122].

Titre : Prédire le résultat de l'injection intracytoplasmique de spermatozoïdes chez les femmes en utilisant l'apprentissage automatique combinant les données Doppler et cliniques

Mots clés : FIV, ICSI, apprentissage automatique, paramètres Power Doppler, Chan-Vese, Split-Bregman.

Résumé : La fécondation in vitro (FIV) est désormais largement utilisée dans le traitement de l'infertilité et son succès est la principale préoccupation des patients. Il est difficile pour les praticiens d'identifier les éléments qui pourraient conduire à une grossesse réussie par FIV en l'absence de technologies automatisées. Le développement d'une technique fiable pour estimer les chances de succès des couples reste un défi ouvert que nous allons tenter de relever en proposant notre nouvel outils SPIRL (System for Predicting the success Rate of IVF using machine Learning combining clinical and Doppler data). De nombreux facteurs contribuant à une grossesse réussie ont fait l'objet de recherches approfondies. Basées sur l'expérience du centre Al Hadi de Beyrouth, l'hypothèse de ce travail repose sur la prise en compte des paramètres écho-Doppler de l'endomètre et du myomètre. Un premier sous-objectif vise alors à explorer l'impact de ces paramètres Doppler sur le succès de la FIV. Pour ce faire, nous avons évalué les capacités prédictives de dix modèles d'apprentissage automatique différents. Ces modèles ont été entraînés à l'aide d'une combinaison de paramètres Doppler et cliniques. Parmi les techniques explorées la méthode de Bagging combinée à la sélection des caractéristiques d'importance Extra Tree a obtenu les meilleures performances (sensibilité de 100 %, spécificité de 80 % et précision de 89,4 %) pour 94 patients. Ensuite, pour 572 patients, le

classifieur Bagging combiné au paramètre d'importance de la forêt aléatoire a présenté la plus grande AUC de 78,2 %. Les résultats obtenus par l'outils SPIRL soulignent l'importance des paramètres écho-Doppler dans le taux de réussite de la FIV. D'autre part, les mesures échographiques de l'endomètre à partir de l'outils VOCAL étant très chronophage et opérateur-dépendant, le second objectif visé est la segmentation automatiquement l'endomètre. En combinant les techniques d'apprentissage automatique et de segmentation automatique et semi-automatique (techniques de Chan-Vese et de Split-Bregman) nous obtenons un indice de Jaccard supérieur ou égal à 0,5 pour plus de 60% des images traitées. Enfin, le troisième objectif était la version entièrement automatisée de SPIRL utilisant notre technique de segmentation entièrement automatisée, qui mesure la surface de l'endomètre, la moyenne de gris de l'endomètre et l'épaisseur de l'endomètre, combinée à d'autres caractéristiques cliniques. Ensuite, il classe et prédit le résultat de l'ICSI. Les résultats sont étroitement alignés avec la prédiction du résultat de l'ICSI en utilisant les mêmes caractéristiques mesurées par des cliniciens à l'aide d'échographies.

Title: Predicting the outcome of intracytoplasmic sperm injection in women using machine learning combining Doppler and clinical data

Keywords: IVF, ICSI, Machine learning, Power Doppler parameters, Chan-Vese, Split-Bregman.

Abstract: In-Vitro Fertilization (IVF) and Intra-Cytoplasmic Sperm Injection (ICSI) are now widely used in the treatment of infertility. Patients' primary concern is the success of an IVF treatment, which is dependent on a variety of influencing factors. It is difficult for practitioners to detect any influencing trend of the characteristics and elements that may lead to a successful IVF pregnancy in the absence of automated technology. The goal was to develop a reliable technique for estimating the likelihood of success for couples seeking to conceive: the SPIRL tool (System for Predicting the success Rate of IVF using machine Learning combining clinical and Doppler data). Several factors that contribute to an accomplished pregnancy have been extensively researched. Based on the experience in Al Hadi Centre, the assumption of this work leads on that Power Doppler parameters related to the endometrium and myometrium play a crucial role in the treatment. The first aimed to explore the impact of Doppler parameters, specifically those related to endometrial and myometrium ultrasonography measures, on the success of IVF. The goal was to develop a reliable technique for estimating the likelihood of success for couples seeking to conceive. To achieve this, we evaluated the predictive capabilities of ten different machine learning models, including Logistic Regression, K-Nearest Neighbor, Multilayer Perceptron, Support Vector Machines, Naive Bayes, Bagging Classifier, Gradient Boosting, Extreme Gradient Boosting, AdaBoost Classifier, and Random Forest. These models were trained using a combination of Doppler and clinical parameters. Initially, among the techniques explored for predicting the outcomes of IVF using machine learning, the Bagging method combined with Extra Tree Importance Feature selection demonstrated superior performance for 94 patients. It achieved a sensitivity of 100%, specificity of 80%, and an accuracy of 89.4%. Subsequently, for 572 patients, the Bagging classifier combined with RF importance exhibited

the highest AUC of 78.2%. Importantly, several variables, such as myometrium VFI, volume of the endometrium, myometrium mean grey, myometrium FI, myometrium VI, endometrium mean grey, and endometrium FI, were identified as significant factors in the prediction process. These findings underscore the importance of ultrasound measurement parameters, particularly Doppler parameters, in influencing the outcomes of IVF and ICSI. The developed technique provides valuable insights for estimating the success rates of these assisted reproduction procedures. However, these parameters of endometrium and myometrium are currently measured manually using VOCAL software, which is a time-consuming process as it requires calculations based on six slices of the endometrium. The second objective of the work aimed to automatically segment the endometrium to calculate its surface, thickness and endometrial mean grey. Then, we predict the outcome of ICSI using machine learning techniques with the new value. First, we employed automatic and semi-automatic segmentation techniques, specifically the Chan-Vese and Split-Bregman techniques. Out of the 98 ultrasound images, 60 achieved a Jaccard index greater than or equal to 0.5, indicating high-quality segmentation. This accounted for 61% of our initial database. Finally, the third goal was the fully automated version of SPIRL taking our fully automated segmentation technique, which measures the endometrium surface, endometrium mean grey, and endometrium thickness, combined with other clinical features, then classify and predict the outcome of ICSI. the results are closely aligning with the prediction of ICSI outcome using the same features measured by clinicians through ultrasound machines.

