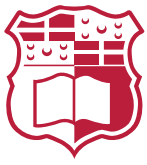
EndoAI Diagnostics Revolutionizing Early Detection and Diagnostics of Endometriosis

Britney Vella

Supervisor: Prof. Matthew Montebello

March 2025

Submitted in partial fulfilment of the requirements for the degree of Master of Science in Artificial Intelligence [Taught and Research (Mainly by Research)].



L-Università ta' Malta
Faculty of Information & Communication Technology

Abstract

Endometriosis is a chronic and debilitating gynaecological disorder affecting approximately 10% of women worldwide. Characterised by the abnormal growth of endometrial-like tissue outside the uterus, the condition often leads to severe physical pain, emotional distress, and mental health challenges, significantly reducing patients' quality of life. Despite its high prevalence, diagnosing endometriosis remains a major clinical challenge due to the heterogeneous nature of its symptoms, frequent misdiagnoses, and reliance on invasive procedures to confirm the diagnosis. Consequently, the average diagnostic delay extends up to eight years.

This dissertation proposes a four-stage solution that addresses these challenges by investigating the potential of Artificial Intelligence (AI) techniques to facilitate the early and accurate diagnosis of endometriosis. Specifically, the study develops a multi-model AI-driven diagnostic framework that leverages both self-reported patient symptom data and laparoscopic medical images. Six Machine Learning (ML) algorithms were employed to predict the likelihood of endometriosis based on symptomatology, incorporating feature engineering techniques to optimise model performance. Additionally, eleven Deep Learning (DL) architectures underwent transfer learning to enhance the detection of endometrial lesions from laparoscopic images.

The effectiveness of the proposed models was evaluated through a comparative analysis using key performance metrics, such as accuracy, precision, and recall. The results demonstrated that AI-powered diagnostic tools significantly enhance the identification of endometriosis, with feature selection and hyperparameter tuning playing a crucial role in improving predictive accuracy. This study further identified high-performing ML and DL models with strong clinical applicability, as well as key symptom-based features essential for detecting the disease.

These findings highlight the transformative potential of AI in medical diagnostics, particularly in addressing the persistent diagnostic delays associated with endometriosis. By integrating AI-driven methodologies into clinical workflows, healthcare professionals can improve early detection rates, minimise misdiagnoses, and ultimately enhance patient outcomes. Furthermore, this study underscores the feasibility of a self-diagnostic tool capable of predicting the likelihood of endometriosis, thereby increasing public awareness of the condition and empowering individuals to seek timely medical consultations. This research contributes to the advancement of AI in gynaecological healthcare, offering a pathway toward more efficient, accessible, and reliable diagnostic solutions for endometriosis.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Professor Matthew Montebello, for his invaluable mentorship, expertise, and unwavering support throughout the course of this research. His insightful feedback, patience, and encouragement have been instrumental in shaping this dissertation.

I am also profoundly thankful to my family and friends for their unconditional love, encouragement, and support. Their belief in me, especially during the most challenging moments of this journey, has been a constant source of motivation and strength. Without their reassurance and understanding, this achievement would not have been possible.

A special mention goes to my two wonderful dogs, Mika and Koda, whose companionship and boundless affection provided the best emotional support throughout this study. Their comforting presence and unwavering enthusiasm never failed to lift my spirits, making even the most stressful days more bearable.

To everyone who has supported me in any way during this journey—thank you.

Contents

List of Figures

List of Tables

List of Abbreviations

AdaBoost Adaptive Boosting.

AI Artificial Intelligence.

ANN Artificial Neural Network.

AR Augmented Reality.

AUC Area Under the Curve.

AUC-ROC Area Under the Receiver Operating Characteristic.

BFS Backward Feature Selection.

BMI body mass index.

CDS Clinical Decision Support.

CNN Convolutional Neural Network.

COCO Common Objects in Context.

CV Computer Vision.

DBSCAN Density-Based Spatial Clustering of Applications with Noise.

DenseNet Densely Connect Network.

DIE Deep Infiltrating Endometriosis.

DL Deep Learning.

DT Decision Tree.

ENID Endometrial Implants Dataset.

FEMaLe Finding Endometriosis using Machine Learning.

FFS Forward Feature Selection.

FN False Negative.

FP False Positive.

GBC Gradient Boosting Classifier.

GEO Gene Expression Omnibus.

GLENDA Gynaecologic Laparoscopy Endometriosis Dataset.

IoU Intersection over Union.

KNN K-Nearest Neighbours.

LASSO Least Absolute Shrinkage and Selection Operator.

LDA Linear Discriminant Analysis.

LR Logistic Regression.

mAP Mean Average Precision.

MCC Matthews Correlation Coefficient.

ML Machine Learning.

MLP Multi-layer Perceptron.

MRI Magnetic Resonance Imaging.

NAS Neural Architecture Search.

NASNet Neural Architecture Search Network.

NLP Natural Language Processing.

NN Neural Network.

PCA Principal Component Analysis.

PLSDA Partial Least Squares Discriminant Analysis.

PR Precision-Recall.

R-NN Region-based Neural Network.

ReLU Rectified Linear Unit.

ResNet Residual Network.

RF Random Forest.

RFE Recursive Feature Elimination.

RNN Recurrent Neural Network.

ROC Receiver-Operating Characteristic Curve.

S3VM Semi-Supervised Support Vector Machine.

SHAP Shapely Additive Explanation.

SVM Support Vector Machine.

t-SNE t-Distributed Stochastic Neighbour Embedding.

TN True Negative.

TP True Positive.

USG ultrasound.

Xception Extreme Inception.

XGBoost eXtreme Gradient Boosting.

1 Introduction

Endometriosis is a chronic medical condition that affects approximately 1 in 10 women globally [1]. It is a disease in which tissue similar to the lining of the uterus grows outside of the appropriate area, which eventually leads to the inflammation and formation of scar tissue in the pelvic area [1]. The symptomology of this condition varies from person to person; however, it is commonly associated with extreme pain during menstrual cycles, bowel movements, and even urination [2]. Moreover, it is also known to cause severe pelvic pain, abdominal bloating, nausea, and, in some cases, result in the patient being diagnosed with depression, anxiety, or infertility [2]. Therefore, it is remarked as a non-preventive, life-impacting disease that decreases the quality of life of the affected women with no current working cure.

At present, diagnosing endometriosis is perceived to be a difficult process due to a multitude of reasons. Primarily, this is due to the heterogeneous nature of the recorded symptomology of the disease, which often leads to misdiagnoses, and the reliance on these symptoms as being the first indication that a woman may be struggling with endometriosis. After endometriosis is suspected, a series of tests are performed in an attempt to detect it further, such as ultrasound (USG) or Magnetic Resonance Imaging (MRI) [3]. Nonetheless, today, the only method that can assuredly diagnose the condition is by performing the invasive procedure known as a laparoscopy [3]. Laparoscopies, however, are only performed by medical professionals when they deem it highly necessary. Furthermore, the lack of awareness about the illness by both healthcare workers as well as the general public has led to women receiving delayed diagnoses of approximately 8 years [3].

This dissertation aims to investigate numerous AI techniques with the aim of diagnosing endometriosis at earlier stages. Specifically, this study aims to conduct a thorough analysis through the development of several ML and DL models and assess their potential in detecting endometriosis by utilising clinical and medical imaging datasets. During this research, statistical, binary classification machine learning models such as Logistic Regression (LR) and eXtreme Gradient Boosting (XGBoost) will be employed with the task of detecting endometriosis using patient self-reported symptoms. On top of that, deep learning models like ResNet50 and InceptionV3 will be implemented to detect endometrial lesions from the laparoscopic image dataset. Finally, a comparative analysis of the developed models will be conducted with the intention of understanding how different AI models compare in terms of accuracy, precision, recall, and more. Hence, the most effective and clinically applicable approach for early endometriosis detection may be identified.

1.1 Motivation

Endometriosis is a chronic, inflammatory health condition affecting approximately 10% of the female population. The severity of the disease may vary depending on the depth and number of endometrial lesions. Some women may be asymptomatic and go through most of their lives without being diagnosed; however, other individuals may find that it negatively impacts their quality of life and even may take a toll on their mental health [3], [4]. Many symptoms of endometriosis may be overlooked or misinterpreted as being caused by a different, more commonly known condition. In addition, the lack of awareness and understanding of the disease and its symptomology by both patients and medical professionals increases the challenge and improbability of diagnosing the disease at an earlier stage [4]. Notably, the average time for identifying endometriosis is approximately 8 years, with certain cases taking up to 11 years [3]. Moreover, non-invasive procedures such as clinical diagnoses and medical imaging have been proven to be inefficient and unreliable in diagnosing endometriosis. While they are tools that are used to aid in the identification of the condition, the current state-of-the-art and gold standard of diagnosing endometriosis is a laparoscopy, which is a costly, highly risky, and invasive surgery that medical professionals refuse to perform unless they deem it necessary [4].

With the integration of AI being applied in the medical field, a significant leap forward has been taken in disease diagnostics. Motivated by recent advancements in AI modelling, this research proposes to conduct a thorough investigation of several ML and Computer Vision (CV) algorithms that can successfully detect endometriosis using clinical and medical imagery data. A predictive model based on patient symptoms would enable healthcare providers to identify the condition earlier. Despite the heterogeneous manifestations of the symptoms, with sufficient data, an AI-powered algorithm will be able to make accurate predictions based on symptomology and highlight the patients at risk of the disease. Meanwhile, an image detection model would aid professionals in accurately detecting the endometrial lesions in medical images. Therefore, a comparative analysis of the implemented models will be executed to understand how different AI models compare in accuracy, efficiency, and clinical applicability. Furthermore, this study gives insight into techniques that may be employed in healthcare systems to aid healthcare providers in identifying the condition, thereby reducing diagnostic delays as well as human errors and ultimately improving patient outcomes.

1.2 Aims and Objectives

This dissertation aims to investigate the early detection and diagnostics of a complex gynaecological condition, known as endometriosis, through the development of an AI-powered algorithm that will ultimately enhance detection accuracy, efficiency, and overall patient outcomes. The focus is on conducting an in-depth evaluation of various ML models that can predict patients at risk of suffering from the disease based on symptomology as well as DL models that are capable of detecting, with high accuracy, the endometrial lesions from medical images.

Main Research Question

How can AI techniques be employed to effectively and efficiently detect and diagnose endometriosis based on clinical and imagery data at early stages?

Objectives

The following four objectives were identified with the intention of attaining and addressing the above-mentioned aim and main research question of this thesis:

Objective 1 - Research and Investigate Various AI Techniques Effective in Disease Diagnostics

This objective focuses on the review of the current literature on AI in Disease Diagnostics and the AI-powered techniques that have been developed to aid medical professionals in image detection and disease prediction based on patient information. With the evolution of AI in this area, researchers have been attempting to develop an efficient diagnostic model that allows for early detection of endometriosis based on different types of data ranging from medical reports to ultrasound imaging. Specifically, this objective is aimed at conducting an in-depth investigation of the current state-of-the-art research on diagnosing or predicting endometriosis to establish what type of clinical and imagery data are best suited for this project. Moreover, insight regarding the ML and DL algorithms developed as well as the evaluation methodologies is provided.

Objective 2 - Attain and Preprocess Clinical and Imagery Datasets

This objective aims to acquire and apply the appropriate preprocessing techniques to the clinical and imagery datasets required for this study. Due to the privacy and ethical concerns surrounding medical data, the necessary datasets are expected to be extremely difficult to obtain. In addition, the imagery data requires a medical expert to identify the endometrial lesions to ensure that the model is trained on the appropriate data. After the data is acquired, preprocessing steps could be

applied to each dataset. The image dataset is resized and adjusted to the input requirements of the implemented models. As for the clinical dataset, data cleaning, integration, reduction, and transformation will be executed as necessary.

Objective 3 - Implement Several Machine Learning and Deep Learning Endometriosis Diagnostic Models

The goal of this objective is to construct a number of clinical and image-based diagnostic models using AI-driven algorithms with the aim of detecting endometriosis. The preliminary list of ML and DL algorithms established in Objective 1 is implemented, and training on base models is performed. In addition, feature engineering techniques, which include several feature selection strategies, are developed and employed in the experimentation process of this research. Furthermore, the features selected from these models will be extracted and assessed in order to identify the most essential features required when detecting endometriosis. With regards to the DL models, further training and fine-tuning of pretrained Keras [5] ImageNet models will be conducted and utilised for the image-based algorithms.

Objective 4 - Evaluate the Effectiveness and Efficiency of the Developed AI Models and Perform a Comparative Analysis

The final objective focuses on the evaluation and comparative analysis conducted on the implemented models' performance metrics. Notably, diagnosing endometriosis is considered to be a binary classification problem, meaning the output will be one of two conclusions. Given the results of the testing phase, a confusion matrix is created using the correct and incorrect detections that each model makes. With this information, several metrics can be tabulated to provide further insight into the models' performance, which include accuracy, precision, and recall. Additionally, accuracy and loss over epochs curve plots are mapped during ML model training to illustrate the improvement or deterioration rate of the models. Finally, a comprehensive comparative assessment of the models is made based on the gathered results of both clinical and image-based models. This will ultimately lead to the conclusion of which methodology is the most efficient, effective, and clinically applicable approach for early detection and diagnosis of endometriosis.

1.3 Proposed Solution

In order to address the diagnostic problem associated with delayed detection of endometriosis, this dissertation proposes to implement and evaluate several AI techniques with the aim of accurately diagnosing the condition at earlier stages and

improving patient care. The proposed solution consists of four key components, each correlating to the research objectives outlined in the previous subsection. In specific, these components include a comprehensive investigation of AI applications in disease diagnostics, the obtainment of relevant datasets, the implementation of ML and DL models, and lastly, the rigorous evaluation and comparative analysis of the developed models.

Through the thorough preliminary investigation conducted during **Objective 1**, an extensive review of the current state-of-the-art research being made in the field of AI in disease diagnostics is provided. This inquiry offers valuable insight into the challenges and limitations that researchers encountered in previous studies, as well as their approaches to overcome them. Key obstacles such as data acquisition constraints, model training limitations and performance evaluation strategies are explored. In addition, as a result of this research, a selection of potential statistical and CV-based models suitable for the detection of endometriosis was established. Furthermore, prior studies that have applied ML and DL algorithms to diagnose or predict endometriosis were identified, thereby enabling a comparative assessment of the models developed in this research in **Objective 4**.

Given the ethical and privacy concerns surrounding medical data, the attainment of the data in **Objective 2** was the first significant challenge this study was expected to overcome. While this study required patient medical records of women with positive and negative cases of endometriosis for the ML algorithms, personally identifiable information such as name, age, or place of residence was not necessary. These models rely solely on structured data that consists of patient symptoms and their corresponding diagnostic outcomes, thus ensuring patient anonymity and mitigating privacy concerns. Meanwhile, the DL algorithms require medical images depicting the absence or presence of endometrial lesions. To ensure data quality, these images must be professionally annotated, with the lesions clearly identified by medical experts. This step is crucial to ensuring that the AI models are trained on accurately labelled data. Hence, for the purpose of this study, photographs taken during laparoscopic procedures that preserved the anonymity of the patients that underwent the procedure were obtained.

Aligned with **Objective 3**, the third phase of this project focuses on the development and implementation of several AI models. In the case of statistical-based diagnosis, six ML algorithms were selected and implemented using the scikit-learn [6] and XGBoost [7] Python libraries. In addition, a number of feature engineering techniques, which included three feature selection strategies, were applied to optimise model performance. Moreover, hyperparameter tuning was employed where necessary to further refine each model, leading to the most optimal versions of each ML algorithm developed. A total of six fine-tuned models of each ML architecture will be

assessed and subsequently compared in the evaluation phase of this study. For CV-based diagnostics, multiple pre-trained ImageNet models were leveraged from the Keras [5] library to facilitate the development of the DL detection models. These architectures were subjected to a transfer learning procedure, which repurposed the models to identify endometrial lesions in laparoscopic images. Finally, the results of both ML and DL models were systematically recorded for subsequent evaluation conducted in **Objective 4**.

Lastly, the final phase of this study corresponds to **Objective 4**, which involves an in-depth evaluation and comparative analysis of the implemented models. A robust assessment framework is established to effectively compare model performance across key metrics such as accuracy, sensitivity, and specificity. Classification reports and confusion matrices provide detailed insights into the performance of both ML and DL-based classification models explored in this research. Additionally, factors such as training time and standard deviation are recorded and also taken into account when evaluating the models' computational efficiency. Furthermore, the training history values of the DL algorithms are utilised to generate accuracy and loss curve graphs, which offer deeper insights into the model convergence and generalisation. This study also aims to compare its findings with previous research to contextualise its contributions within the field of AI-powered diagnostic tools. Ultimately, the most clinically applicable model can be determined by considering both diagnostic accuracy and computational efficiency, ensuring that the final model selection aligns with real-world healthcare requirements.

1.4 Contributions

This dissertation aims to contribute to the medical field of AI-driven disease diagnostics, specifically in the early and accurate diagnosis of endometriosis. Based on the results and outcomes concluded during this research, the key contributions are as follows:

- **Development of high-performing ML models:** The ML models trained on self-reported patient data demonstrated exceptional diagnostic performance, with most models attaining over 90% accuracy, further highlighting the potential of symptom-based AI diagnostics.
- **Development of high-performing DL models:** The DL algorithms were able to accurately identify endometrial lesions from laparoscopic images, achieving over 95% accuracy scores. This further demonstrates the viability of AI-assisted image detection of endometriosis.

- **Identification of key diagnostic features:** By analysing the performance of the developed models and feature combinations, this research provides insights into the most key diagnostic indicators of endometriosis, contributing to a deeper understanding of the disease and its characteristics.
- **Identification of the most clinically applicable AI models:** Through the comparative analysis, this study identifies the most effective and computationally efficient ML and DL models that have the potential to be translated into practical clinical tools. These models could assist healthcare professionals in diagnosing endometriosis more efficiently, enabling earlier intervention and improved patient outcomes.

1.5 Document Structure

This dissertation is structured into several chapters, each addressing a key aspect of this research. This subsection focuses on providing an overview of the content covered in the subsequent chapters.

Chapter 2 establishes and explores the fundamental concepts relevant to the research. It provides a comprehensive review of endometriosis, detailing its heterogeneous symptomology as well as the challenges associated with its diagnosis. Following that, an in-depth discussion regarding AI methodologies applied in disease diagnostics is presented, covering both ML and DL approaches. Finally, this chapter introduces a number of evaluation metrics used for assessing binary classification modelling algorithms. Chapter 3 focuses on examining prior research conducted on AI in disease diagnostics, with a specific focus on projects related to the prediction and diagnosis of endometriosis. It reviews various model architectures developed in these studies, highlighting the different data types utilised, such as clinical, gene and medical imaging data. The chapter also summarises notable research initiatives being made in this domain. Chapter 4 details the methodology adopted in this study in order to achieve the aims and objectives established in Chapter 1. It describes the data collection and preprocessing strategies applied to the acquired self-reported patient and imagery data. Additionally, it provides a comprehensive explanation of the ML and DL model implementations, including details on the utilised architectures. Furthermore, this chapter outlines the experimentation process adopted in this study, which includes feature engineering and hyperparameter tuning techniques. In Chapter 5, an in-depth evaluation of the developed models is made. It begins by describing the evaluation plan adopted in this study, followed by a thorough assessment of the model's performance and their results. Furthermore, a comparative analysis is conducted between the implemented model and the state-of-the-art models reviewed in the

Literature Review chapter. This critical analysis ensures a thorough understanding of model effectiveness and potential clinical applicability. The final chapter concludes the dissertation by revisiting the research aims and objectives and summarising how they were addressed. In addition, it discusses the key findings and limitations encountered during this research. Lastly, this chapter highlights areas for future research and concludes with final remarks on the implications made by this study.

2 Background

This chapter focuses on providing contextual information and foundational concepts relevant to the work conducted during this dissertation. It offers a detailed overview of the medical condition endometriosis, including its symptomatology, current diagnostic approaches, and associated challenges. Additionally, this section explores the role of AI in disease diagnostics, particularly with respect to the ML and DL methodologies applied in binary classification tasks, along with relevant evaluation strategies.

2.1 Endometriosis: A Complex Gynecological Condition

Endometriosis is a chronic, poorly understood female gynaecological condition characterised by the presence of endometrial-like tissue outside of the uterine cavity, as depicted in Figure 2.1 [1]. As of today, the underlying biological mechanisms that contribute to the development of the disease are still unclear, with the most widely accepted theory being that of retrograde menstruation. This hypothesis suggests that the menstrual blood, which contains viable endometrial cells, flows back into the peritoneal cavity instead of exiting, causing these cells to adhere to the surrounding tissues, multiply and eventually form into endometrial lesions [2]. Other theories, such as immune dysfunction and coelomic metaplasia, imply that genetic or environmental factors may cause the transformation of the peritoneal cells into endometrial tissue [2]. Additionally, researchers have observed that the disease's inflammation enhances immunological and cytokine dysregulation, resulting in lesion persistence and excruciating chronic pain for patients. Moreover, hormonal imbalances may increase the rate of multiplication of the ectopic endometrial tissue, thereby accelerating disease progression [2].

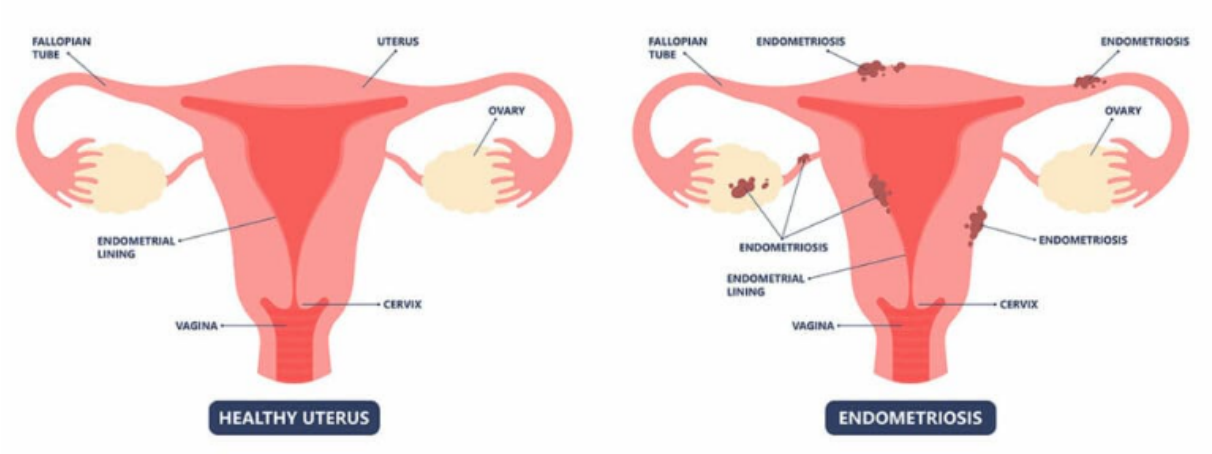


Figure 2.1 Endometriosis

An accurate measurement of the prevalence rate within the general population is difficult to ascertain due to the reliance on the invasive surgical procedure required for a definitive diagnosis. However, according to community-based prevalence studies, it is estimated that approximately 10% of the female population is affected by this disease [1]. To be precise, the endometriosis prevalence rates vary among population groups. According to estimates, endometriosis affects 2% to 11% of asymptomatic women, 5% to 21% of women currently hospitalised for pelvic pain, and 5% to 50% among those experiencing infertility [2]. Furthermore, adolescent females have a significantly high prevalence rate, ranging from 49% to 75%.

Endometriosis is a complex, heterogeneous disease with a broad spectrum of clinical manifestations that make the diagnostic process particularly challenging. Symptom severity has been noted to vary between individuals, often correlating with the extent and location of the endometrial lesions [3]. Commonly experienced symptoms include menstrual pain, chronic pelvic pain, dyspareunia, dyschezia, and fatigue [3]. Additionally, patients have also frequently reported gastrointestinal and urinary symptoms such as abdominal bloating, nausea, painful urination, and loss of bladder control [3]. Moreover, due to the inflammation, adhesions and scarring in the pelvic area, along with the hormonal abnormalities, endometriosis is also highly associated with infertility and has been known to cause reproductive health problems [3]. As a result, many women with endometriosis require fertility treatments and assisted reproductive technologies to achieve pregnancies [2]. Studies have also shown that endometriosis often coexists with other gynaecological conditions such as uterine fibroids and ovarian cancer, which may further increase the difficulty of diagnosis [2]. The disease has also been observed to have a detrimental effect on the patient's quality of life, which occasionally leads to mental health issues and, in certain cases, results in the patient developing anxiety and depression [2].

2.2 Current Diagnostic Approaches of Endometriosis

To this day, the diagnosis of endometriosis remains a challenge due to the heterogeneous nature of its symptomology as well as the lack of understanding of the condition and its pathophysiology. The current diagnostic approaches that have been utilised by medical professionals in an attempt to effectively and efficiently detect the disease at early stages include medical imaging tools, surgical interventions, and clinical assessments through symptom and medical history analysis.

Clinical assessment based on patient symptomatology and history is an integral part of diagnosing endometriosis. However, this approach presents significant challenges due to the heterogeneous and often overlapping nature of symptoms

associated with the disease. Additionally, this method heavily relies on the healthcare providers' knowledge of the condition and whether they have access to the necessary tools required to make a diagnosis [4]. Moreover, diagnosis through medical information such as biomarkers and genetic testing has also proven to be inconsistent in accurately detecting the condition, which further limits the tools available to medical professionals for diagnosis in a clinical setting. Although more healthcare systems are implementing clinical evaluation in an effort to reduce diagnostic delays of the disease, its highly subjective nature, susceptibility to misdiagnosis, and the lack of reliable clinical assessment tools make it an unreliable method for definitely diagnosing it [4]. Despite these limitations, this approach serves as a crucial first step in the diagnostic process by allowing for early identification of potential cases and recommending further conclusive diagnosis through more effective tools when necessary.

The three primary imaging procedures that have been employed by medical professionals to detect endometriosis more accurately include USG and MRI. These non-invasive techniques are commonly performed in the initial stages of the diagnostic process. These imaging methods have demonstrated high success rates for detecting endometriosis, however, they are often limited in their ability to detect smaller endometrial lesions or those located in less accessible anatomical regions [8]. Although these procedures are more cost-efficient than surgical interventions and more accurate than clinical assessments, they have proven to be insufficient for diagnosing early-stage endometriosis, as subtle lesions may remain undetected, thereby degrading patients' quality of life and lengthening the diagnostic process [8].

Laparoscopy is regarded as the gold standard technique for definitively diagnosing endometriosis due to its ability to provide direct observability of the lesions [4]. This minimally invasive surgical procedure allows for a more precise evaluation of lesion distribution and severity. Nevertheless, this approach is costly, carries a risk of surgical complications, and is highly dependent on the surgeons' knowledge of the condition to detect even the subtle early-stage lesions [4]. In cases where suspicious tissue is identified, a biopsy may be performed where a sample of the tissue is extracted for further assessment [8]. This sample is tested to confirm the presence of endometrial-like tissue through a histological examination [8].

Despite recent advancements in diagnostic techniques, misdiagnosis and delayed diagnosis continue to pose significant challenges when diagnosing endometriosis. The subjective and varied nature of symptom-based assessment, combined with the limitations of existing imaging and laboratory tests, have resulted in an average diagnosis time of approximately eight years [4]. This highlights the need for innovative diagnostic approaches. Therefore, the integration of AI-driven methodologies into clinical practice presents a potential avenue for enhancing diagnostic accuracy, minimising diagnsotic delays, and improving patient outcomes.

2.3 AI Methodologies in Disease Diagnostics

With the integration of AI in the healthcare industry, an unprecedented transformation has been made in several medical research domains, including disease diagnostics, disease progression monitoring, treatment planning, and patient care [9]. ML and DL methodologies have revolutionised medical data analysis by enabling automated interpretation of large volumes of clinical data such as test results, patient histories and medical images. These innovative AI-driven tools support medical professionals in the decision-making process by identifying patterns, trends and abnormalities with a high level of precision that often surpasses traditional diagnostic methods [10]. As a result, AI methodologies have significantly enhanced diagnostic accuracy, increased efficiency, personalised treatment strategies and improved early disease detection, thereby improving overall patient outcomes and reducing the risk of misdiagnosis [10].

2.3.1 Machine Learning

Machine Learning is a branch of AI focused on the development of self-learning algorithms that can generate predictions or diagnoses based on the extracted patterns [9]. These models are usually applied to structured, tabular datasets, as they rely on mathematical and statistical principles for pattern recognition and predictive analysis. In addition, as illustrated in Figure 2.2, ML techniques can be further categorised into four learning paradigms, including supervised, unsupervised, semi-supervised, and reinforcement learning [11].

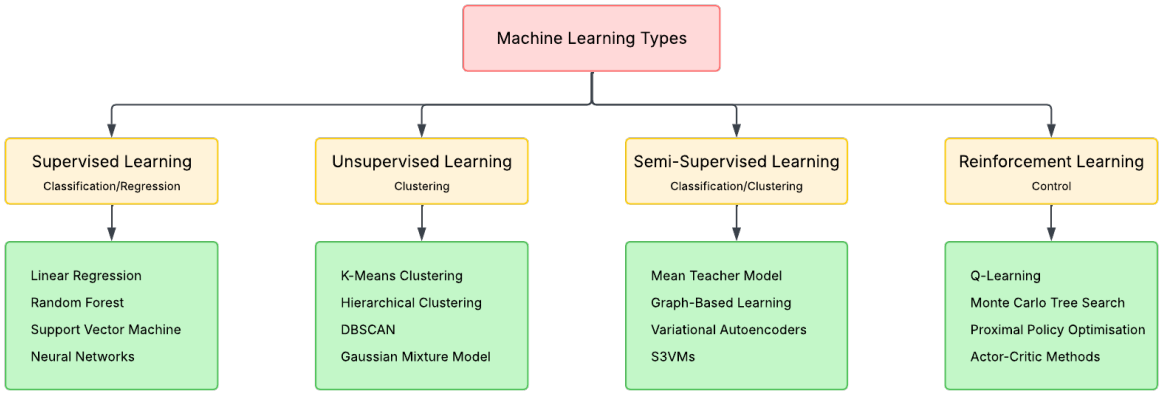


Figure 2.2 Machine Learning Types

Supervised learning is the most widely used ML paradigm in healthcare applications, requiring labelled training data in order to establish the relationship between the input features and desired output labels [11]. This learning technique can be further divided into two subcategories, namely classification and regression-based

models. Classification models generate predictions by assigning categorical labels to input data, whereas regression models predict continuous numerical values. Some common supervised learning models include Logistic Regression (LR) and Random Forest (RF).

In contrast, unsupervised learning is employed when dealing with unlabelled datasets to identify patterns and structures without the guidance of predefined output labels [11]. This type of learning algorithm is commonly applied to clustering problems, anomaly detection and pattern discovery [11]. K-Means clustering and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) are two popular unsupervised learning models.

Semi-supervised learning is a hybrid approach that combines elements of both supervised and unsupervised learning to generate predictions [11]. These algorithms leverage a small subset of labelled data in conjunction with a large set of unlabelled data in an attempt to improve model generalisation [12]. Such models can be applied to classification, clustering and anomaly detection tasks. Typical semi-supervised algorithms include Graph-Based learning and Semi-Supervised Support Vector Machine (S3VM) models.

Lastly, reinforcement learning involves training models to make sequential decisions by interacting with an environment through a reward-based trial-and-error process [11]. While this approach is commonly employed in robotics and autonomous systems, it has limited application in disease diagnostics. Some common reinforcement learning algorithms include Q-Learning and Monte Carlo Tree Search.

Overview of Selected Machine Learning Algorithms

Since diagnosing endometriosis is a binary classification problem that determines whether a patient has the disease or not, this study will focus on supervised learning techniques. In particular, the ML models explored in this research include Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Adaptive Boosting (AdaBoost), and Support Vector Machine (SVM).

LR is a linear, statistical algorithm that is widely applied in binary classification tasks. It utilises a logistic function, commonly referred to as the sigmoid function, to map and estimate the probability of an event occurring against the probability of the event not occurring [11], [12]. The model outputs a probability between 0 and 1 to signify the likelihood of the input belonging to a specific class.

DT models are linear, tree-structured algorithms employed in both classification and regression tasks. As illustrated in Figure 2.3, these algorithms are visually represented as tree-like structures with characteristics such as nodes, branches and leaves, where nodes represent the different decisions the model may select and the

leaves represent the final predictive class. These models make decisions by recursively splitting the data based on the feature values, forming branches that lead to different classification outcomes [11]. The final prediction is derived by traversing through the tree from the root to some specific leaf node, which signifies the classification [12].

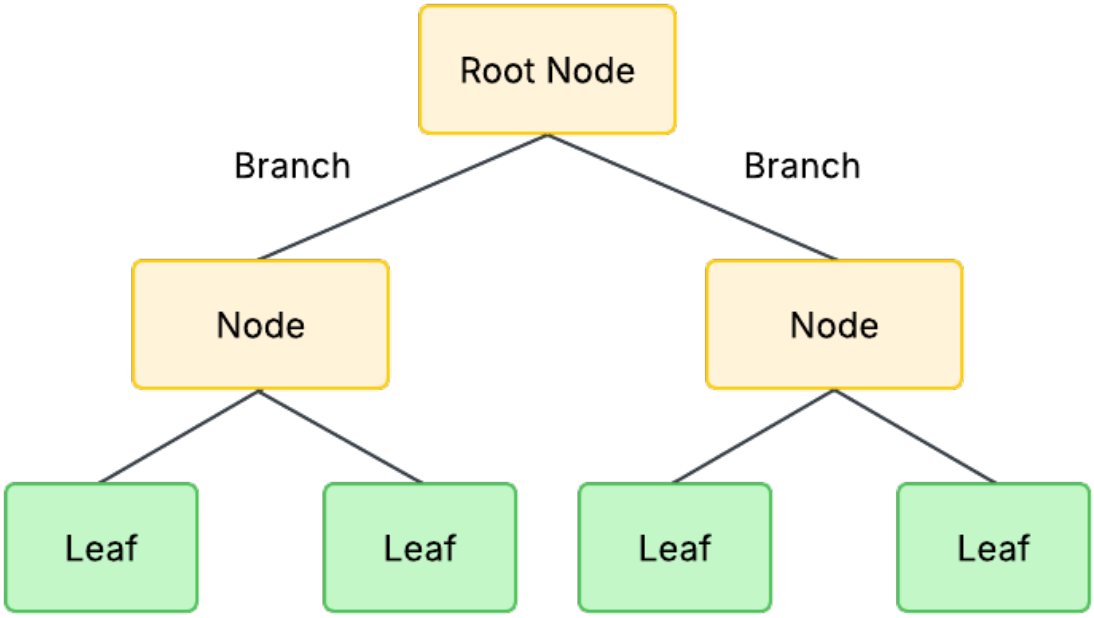


Figure 2.3 Decision Tree Architecture

The RF classifier is an ensemble algorithm that uses random feature selection and bootstrap aggregation, otherwise known as bagging, techniques to establish an effective predictive model [12]. This model constructs and trains several DTs in parallel using subsets of the dataset and then aggregates the outputs of all the individual trees to derive the final result [13]. In comparison to single DT models, this modelling algorithm is frequently used to enhance model robustness and minimise the risk of overfitting.

Similarly to RF, the gradient boosting technique, XGBoost, also employs weaker models, typically DTs, to build a more robust and optimised ML model [13]. However, while RF models construct trees independently, this algorithm iteratively refines model performance by minimising the calculated gradient loss function at each step until the most optimal model is derived [12].

AdaBoost is another sequential ensambling classifier that constantly makes adjustments based on prior errors in order to produce an effective model. This algorithm focuses on improving weak classifiers by assigning higher weights to misclassified instances in subsequent iterations. Interestingly, though, this model is vulnerable to outliers and noisy data in binary classification problems. [12]

Finally, SVM models are supervised learning algorithms that create an optimal hyperplane in high-dimensional space to separate and categorise data points into different groups [12]. Notably, this algorithm may behave differently depending on the kernel that is employed, as they modify the way the data points are separated. Some common SVM kernels include linear, polynomial, sigmoid and radial basis [12].

Feature Engineering

Feature engineering is a crucial step in ML model development as it significantly enhances the models' performance, improves interpretability and minimises the risk of overfitting. This process involves the transformation of raw data into meaningful features through feature selection and extraction techniques, ensuring that models learn from the most relevant and informative subset of the dataset [14]. Feature engineering can be categorised into feature selection and feature extraction approaches, with feature selection being further subdivided into filtering, wrapper and embedded methods, as illustrated in Figure 2.4. This section focuses on examining the various feature engineering techniques investigated in this research.

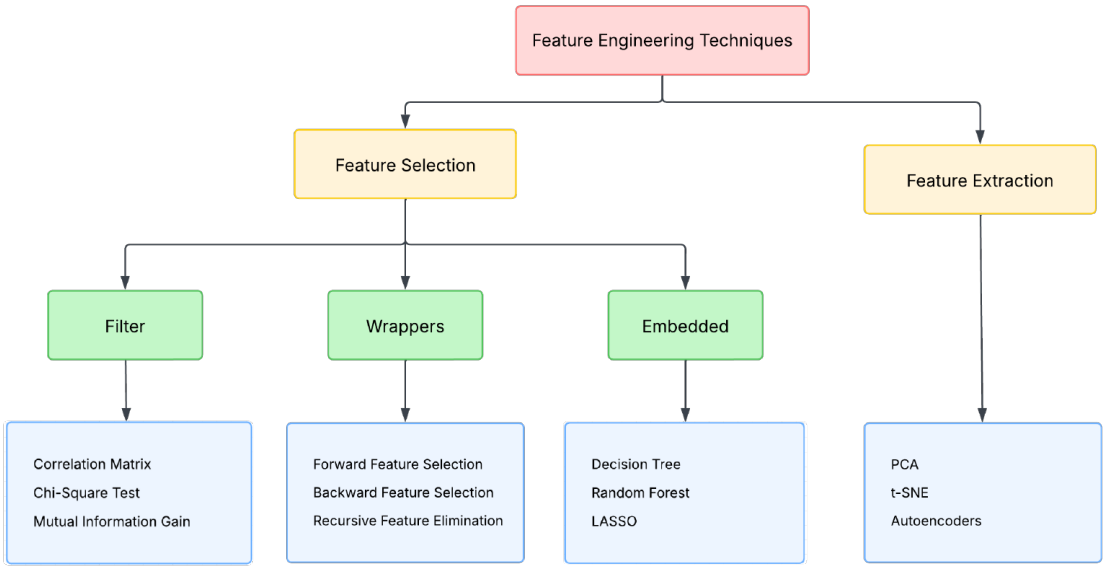


Figure 2.4 Feature Engineering Techniques

Feature Selection

Feature selection is the process of identifying an optimal subset of relevant features that contribute the most to generating accurate predictions while eliminating redundant, noisy, or irrelevant variables [15]. By reducing dimensionality, this process not only enhances computational efficiency but also improves model generalisability and interpretability. As depicted in Figure 2.4, feature selection techniques are divided into three primary categories, namely, filtering, wrapper and embedded methods.

Filtering methods employ feature ranking techniques in order to select the most relevant features from a dataset [14]. This process consists of ranking features within the dataset based on the statistical relationship with the target variable, where features that score below a predefined threshold are removed from the dataset [15]. Since these techniques measure the correlation between the features and target class independently of the ML classifiers, they ensure generalisation of the models and mitigate the risk of overfitting [15]. The correlation matrix, which calculates the linear relationship between features, is a commonly used filtering technique. Some common filtering methods include correlation matrices, chi-square test and mutual information gain [16]. These methods measure the linear relationship between features, the independence between the categorical features and target variables, and the importance a feature provides on the target variables, respectively.

Wrapper methods determine the most informative set of features by iteratively testing various feature subset combinations and assessing them according to the classifier's predictive performance [14]. In contrast to filter methods, these techniques are model-dependent, which implies that each ML model requires its own feature selection process in order to determine the most relevant feature combinations that produce the best results for that particular modelling algorithm [15]. Notably, while wrapper methods often yield higher predictive accuracy, they are computationally expensive when compared to the other feature selection methods due to their iterative nature [16]. One such wrapper technique is Forward Feature Selection (FFS), which iteratively adds features based on the impact on the model's performance. Another common wrapper technique includes Backward Feature Selection (BFS), which starts with all the entire feature dataset and iteratively removes the least important ones. Moreover, Recursive Feature Elimination (RFE) is a wrapper approach that evaluates feature importance by recursively training the model and removing the least impactful feature at each step.

Embedded methods integrate the feature selection process directly into the ML model training phase in order to balance the advantages of filter and wrapper methods [14]. Therefore, these techniques reduce the computational cost that arose in wrapper methods while still taking into account the classifier in the feature selection process [15]. In addition, they leverage regularisation and DT-based algorithms to identify

relevant features. Therefore, this means that tree-based models such as DTs, RF and gradient-boosting classifiers, which inherently rank feature importance based on their contribution to classification decisions, are prime examples of embedded ML approaches. Furthermore, Least Absolute Shrinkage and Selection Operator (LASSO) and Elastic Net are common embedded methods that apply regularisation strategies in order to determine the optimal feature subsets.

Feature Extraction

Feature extraction techniques focus on transforming high-dimensional raw data into lower-dimensional representation while preserving important information [16]. Unlike feature selection methods, which eliminate irrelevant variables, feature extraction creates new, compact feature representations that retain essential patterns [16]. These techniques can be categorised into linear and non-linear techniques, with some of the most widely used approaches including Principal Component Analysis (PCA), t-Distributed Stochastic Neighbour Embedding (t-SNE), and autoencoders. PCA is a linear algorithm that reduces the dimensionality of the dataset by finding an optimal hyperplane that best distinguishes between different classes. Meanwhile, t-SNE is a non-linear feature extraction technique that maps complex structures into lower-dimensional space while preserving local relationships [16]. Lastly, autoencoders are a type of Neural Network (NN) algorithms that learn compressed feature representations by encoding input data into a reduced-dimensional space and then reconstructing it. These methods are commonly used to extract features from unstructured data, like images [16].

2.3.2 Deep Learning

Deep Learning is a subset of ML that utilises Artificial Neural Network (ANN) to model complex patterns within data. These architectures are inspired by the structure and function of the human brain, consisting of interconnected layers of artificial neurons designed to process and learn from vast amounts of input data [12]. The basic ANN architecture, represented in Figure 2.5, is made up of multiple nodes that comprise the input, hidden, and output layers as well as the connections between them. Unlike traditional ML approaches, which often rely on manual feature extraction techniques, DL models automatically learn hierarchical representations of data, thereby making them particularly effective in non-rule-based programming tasks such as image processing and recognition [11]. Several DL architectures have been developed, each serving distinct purposes within the broader field of AI-driven diagnostics, including Multi-layer Perceptron (MLP), Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN) models.

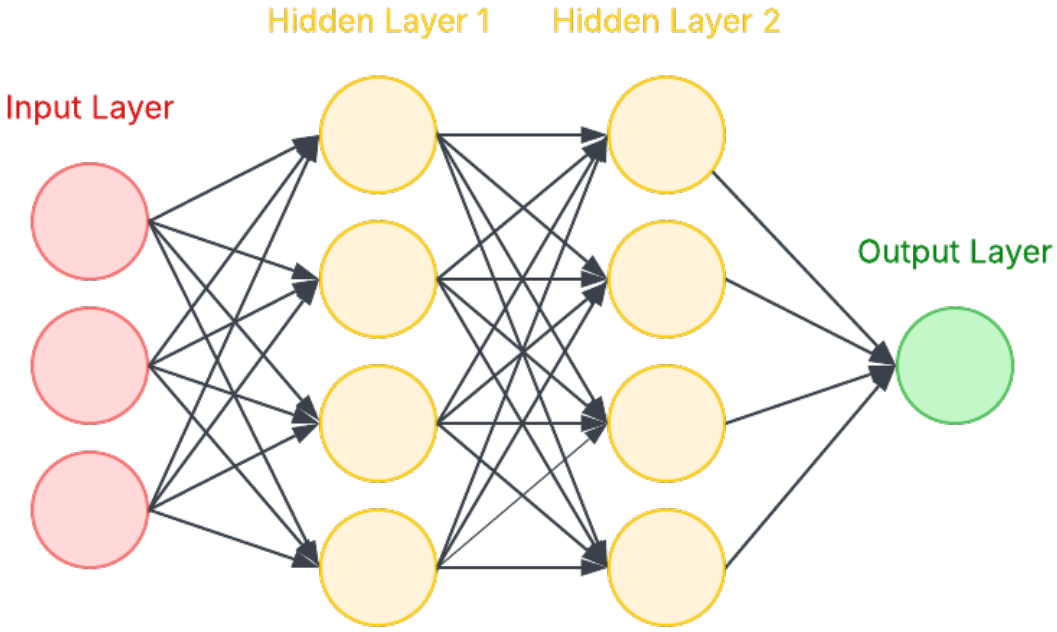


Figure 2.5 Artificial Neural Network Architecture

Although there are numerous DL methodologies available for disease diagnostics, CNN algorithms in particular excel in medical image analysis due to their ability to automatically extract spatial and hierarchical features as well as identify intricate patterns from visual data. As such, CNN-based models were selected for further investigation in this study to facilitate the detection of endometrial lesions in laparoscopic images.

As illustrated in Figure 2.6, CNNs process images through a series of convolutional and pooling layers before the classification process that occurs in the fully connected layers. The convolution layer is responsible for applying differing filtering strategies to extract important features, such as edges or textures, and creating a feature map that represents the presence of these elements. The pooling layer performs dimensionality reduction of the generated feature maps by applying downsampling operations, such as min-pooling or average-pooling, to enhance model generalisation as well as decrease computational complexity. Finally, classification occurs in the fully connected layer through the use of an activation function, like Softmax or the Rectified Linear Unit (ReLU), that determines the likelihood of a given class label. [11]

DL models require a substantial amount of labelled training data in order to achieve the desired high accuracy necessary in diagnostic tools. However, acquiring large-scale datasets in the medical field often presents a challenge due to data privacy concerns and the limited availability of annotated images. Therefore, to address this

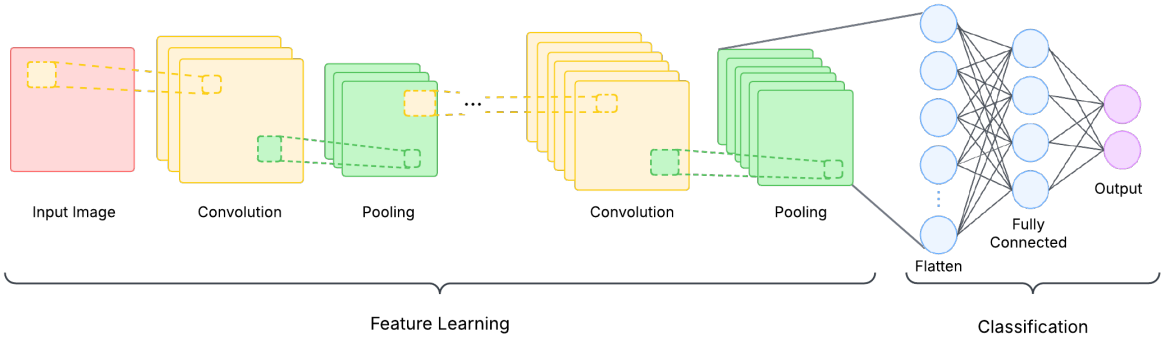


Figure 2.6 Convolutional Neural Network Architecture

issue, this study aims to investigate several pre-trained CNN-based DL models and apply transfer learning in order to retrain and fine-tune the models to detect endometrial lesions in laparoscopic images.

Overview of Selected Deep Learning Architectures

For the purpose of this study, eleven distinct DL architectures are selected based on their unique structural design, computational efficiency and suitability for image classification tasks. These architectures can be broadly categorised into traditional, optimised lightweight, deep and inception-based CNN algorithms. Namely, the models selected for investigation during this research include VGG16, ResNet50, ResNet50V2, DenseNet121, InceptionV3, Xception, InceptionResNetV2, NASNetMobile, EfficientNetV2 B0, as well as MobileNetV3 small and large. The following section provides an overview of each architecture, highlighting its structural composition and computational considerations.

VGG is one of the earliest CNN architectures developed by Simonyan and Zisserman [17]. Similarly to the CNN process illustrated in Figure 2.6, this model follows a straightforward sequential design, employing convolutional filters for feature extraction, max-pooling layers to reduce spatial dimensions, and fully connected layers to generate the final classification. Depending on the employed variant, VGG networks contain either 16 or 19 weight layers. Due to its deep yet uniform structure, this architecture is commonly used as a benchmark model for image classification tasks. Furthermore, despite their effectiveness, VGG frameworks are notably computationally and memory-expensive, making them significantly slower compared to other modern optimised architectures.

The Residual Network (ResNet) architectures were introduced by He et al. [18] in order to address the vanishing gradient problem in deep networks. These models incorporate residual connections, otherwise known as skip connections, that allow

gradients to propagate more efficiently by bypassing certain layers. The ResNet framework includes several variants with varying depths, such as 50, 101 or 152 layers. These architectures are highly popular in feature extraction and classification tasks as they balance depth, parameter efficiency, accuracy, as well as computational cost. An improved version of the ResNet framework, ResNet V2, was later introduced by He et al. [19]. These architectures refined the original design by batch normalisation and ReLU activation, which were applied before the convolution layers rather than after. This adjustment led to an optimised model with enhanced gradient flow, faster convergence and improved performance.

Huang et al. [20] introduced the Densely Connect Network (DenseNet), which connects each layer of the network to all subsequent layers through the innovative design of Dense Blocks. This connectivity promoted feature reuse across the network and improved gradient propagation, thereby mitigating the vanishing gradient problem. These architectures are available in variants with 121, 169 or 201 layers, incorporating convolution and average pooling layers for a compact parameter structure. Despite their depth, DenseNet models are parameter-efficient and memory-effective, resulting in high model performance rates when compared to deeper models.

The InceptionV3 architecture is a 48-layer CNN designed for efficient multi-scale feature extraction by Szegedy et al. [21]. It employs asymmetric and factorised convolutions as well as auxiliary classifiers to enhance gradient flow and computational efficiency. In addition, this architecture also applies batch normalisation and label smoothing in order to improve training stability and convergence speed. Due to its balance between model performance and computational efficiency, InceptionV3 is a widely adopted framework for image classification problems.

Chollet [22] developed an enhanced variation of the Inception architecture called Extreme Inception (Xception). This architecture replaces standard convolutions with depthwise separable convolutions, significantly reducing the number of parameters in the framework while maintaining high classification accuracy. With 71 layers, Xception is one of the most computationally efficient and high-performing CNN architectures.

The InceptionResNet V2 architecture, by Szegedy et al. [22], combines the Inception modules with residual learning from the ResNet framework in order to optimise the feature extraction process and gradient flow within the networks. This process enhances the training stability and classification accuracy of the model, making it one of the most high-performing modelling architectures for image classification problems. However, with a depth of 164 layers, this model is notably more computationally expensive than other Inception-based models.

The Neural Architecture Search Network (NASNet) architecture is an optimised lightweight framework developed by Zoph et al. [23]. Through the innovative design of

a new search space, the Neural Architecture Search (NAS), this architecture uses a reinforcement learning-based technique to optimise CNN building blocks and enhance transferability. Additionally, the NASNet framework employs separable convolutions and Scheduled Drop Path regularisation in order to improve computational efficiency while maintaining the model's high performance rate. These architectures have varying depths based on their configurations. NASNet Large consists of 529 layers, making it one of the deepest architectures specifically optimised for high-performance image classification tasks. In contrast, NASNet Mobile's 88-layer depth effectively balances model performance and computational costs, making it appropriate for environments with limited resources.

Tan and Le [24] introduce EfficientNetV2, which is a series of CNN algorithms that build upon its predecessor, the EfficientNet framework. These architectures incorporate a combination of training-aware neural architecture search and progressive learning strategies to enhance training speed and improve parameter efficiency. In addition, they dynamically adjust the regularisation techniques during the training process to ensure model accuracy remains optimal while optimising computational performance. This framework includes several variants, with B0 to B3 models ranging in depth from 82 to 110 layers and S, M and L models featuring depths of 137, 212 and 304 layers, respectively.

MobileNets are lightweight deep CNNs that utilise depthwise separable convolutions for optimisation. Howard et al. [25] introduce the MobileV3 architectures, which employ two hyperparameters that trade off between latency and accuracy in order to enhance the model's performance. These hyperparameters allow the model builder to determine the most optimal architecture based on specific application constraints. In addition, through hard-swish activation functions and a squeeze-and-excitation module, these frameworks enhance model efficiency and lower the computation cost while maintaining high accuracy scores. The MobileNetV3 Small and Large variants consist of 28 and 44 layers, respectively, making them highly efficient in applications with limited available resources.

By integrating these state-of-the-art DL architectures with transfer learning, this study aims to develop an accurate and efficient diagnostic tool for the detection of endometrial lesions in laparoscopic images. The abovementioned architectures offer a diverse range of depth, efficiency, and accuracy, allowing for a comprehensive performance assessment in Chapter 5 to identify the most effective model for this specific medical imaging task.

Transfer Learning

Transfer learning is a DL technique that involves leveraging pretrained models, originally trained for a specific task, and repurposing them as the foundation for solving a different, often unrelated, problem [26]. This approach is particularly valuable in medical imaging applications where obtaining large, annotated datasets often presents a challenge. By transferring knowledge previously acquired from a well-established dataset, this technique not only mitigates data scarcity issues but also significantly reduces computational demands and training durations while preserving high model performance [26].

The standard transfer learning procedure begins with a DL model being initialised using pretrained weights derived from prior training, usually on a general large-scale dataset such as ImageNet. Instead of learning from scratch, the model refines these predetermined weights using optimisation algorithms, allowing it to adapt to the specific patterns and features of the new dataset [26]. In this research, transfer learning is employed to fine-tune pretrained Keras models, initialised with ImageNet weights, for the task of detecting endometrial lesions in laparoscopic images. This process is illustrated in Figure 2.7.

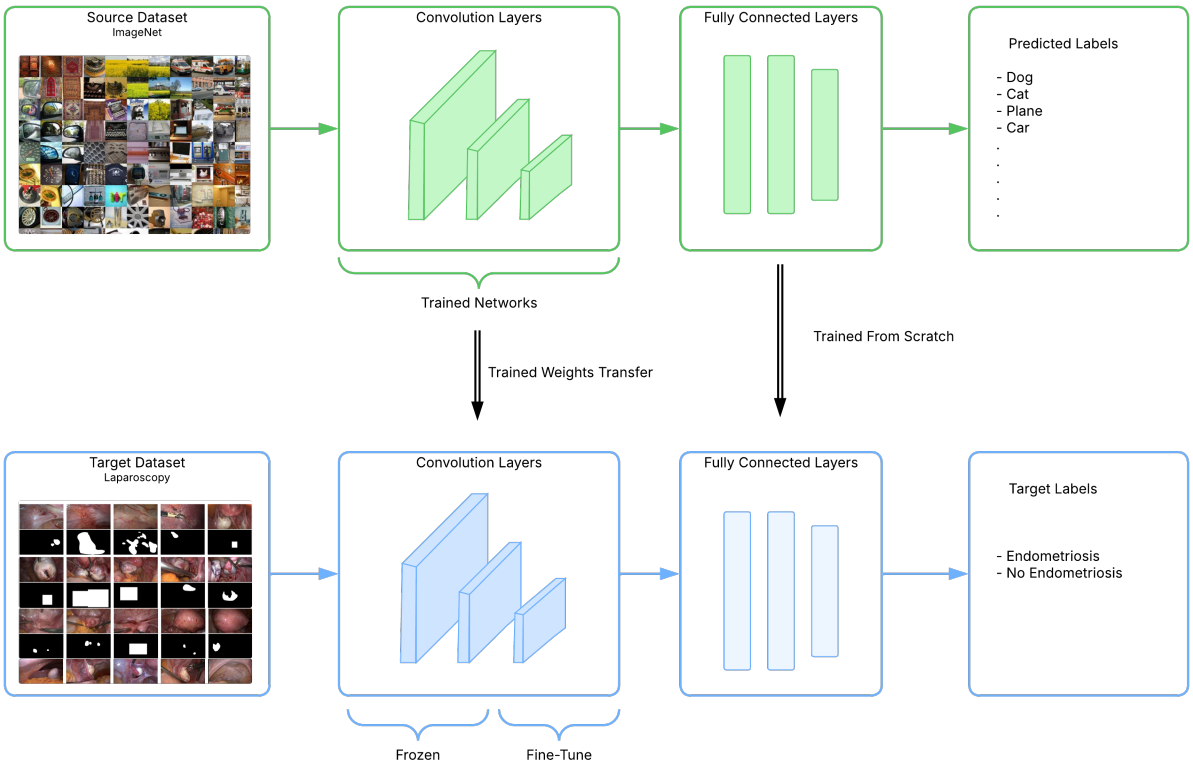


Figure 2.7 Transfer Learning Process

2.4 Evaluation Metrics for Binary Classification Tasks

The generated predictions in binary classification models fall into one of four categories, namely True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN). A true result implies that the model was able to correctly classify the instances. Conversely, a false result denoted that the model misclassified the instance. In specific, a TP occurs when the model correctly identifies a positive instance, while a TN represents a correctly classified negative instance. Meanwhile, a FP occurs when a negative instance is incorrectly classified by the model as positive, while a FN arises when a positive instance is incorrectly identified as a negative instance. These four outcomes are commonly represented in a confusion matrix, which provides a structured tabular summary of the model's classification performance. Additionally, from this matrix, several evaluation metrics can be derived to further assess the effectiveness and reliability of the developed model.

One of the most widely used evaluation metrics is accuracy, which quantifies the proposition of the correctly classified instances over the total number of predictions the model made during the testing process [27]. This metric is calculated using the mathematical formula denoted in Equation (2.1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

While accuracy provides an overall measure of the correctness, it may be misleading in cases where there is a class imbalance in the dataset since the model may achieve a high accuracy score simply by predicting the majority class. Therefore, the error rate, given by Equation (2.2), is a complimentary metric that is used to calculate the percentage of the misclassified instances [27].

$$ErrorRate = \frac{FP + FN}{TP + TN + FP + FN} \tag{2.2}$$

Another metric that is commonly used to evaluate the performance of classification models is precision, whose mathematical formula is represented in Equation (2.3). Also referred to as the positive predictive value, the precision score measures the proportion of correctly predicted positive instances relative to all positive predictions [28].

$$Precision = \frac{TP}{TP + FP} \tag{2.3}$$

A high precision score implies that when the model generates a positive prediction, it is highly likely to be correct. However, this metric does not account for the false negatives. Hence, this metric is commonly calculated in conjunction with the

recall score. Recall, also known as sensitivity or true positive rate, formulated by Equation (2.4), determines the model's ability to correctly identify all actual positive instances [28].

$$Recall = \frac{TP}{TP + FN} \tag{2.4}$$

In contrast to sensitivity, the specificity metric evaluates the model's ability to correctly classify all actual negative instances from the training process [27]. This metric, tabulated by Equation (2.5), is also referred to as the true negative rate.

$$Specificity = \frac{TN}{TN + FP} \tag{2.5}$$

A related metric is the false positive rate, which represents the model's probability of misidentifying a negative instance as positive [28]. This classification metric, defined by Equation (2.6), can be calculated using the false positive and true negative instances as well as the specificity score.

$$FPR = \frac{FP}{TN + FP} = 1 - Specificity \tag{2.6}$$

Another key performance metric commonly used to assess classification-based models is the F-Measure, which is commonly referred to as the F1-score. This metric provides a balanced evaluation by combining the harmonic mean of the precision and recall metrics formulated by Equation (2.7) [27]. This metric is particularly useful when the positive and negative instances are disproportionate to each other, as it ensures that both precision and recall measurements are taken into consideration.

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{2.7}$$

The Area Under the Receiver Operating Characteristic (AUC-ROC) Curve is a widely used visual evaluation metric that illustrates the performance of the model by plotting the abovementioned true positive rate against the false positive rate metrics [28]. This curve provides a comprehensive measure of the model's ability to differentiate between classes, where a higher Area Under the Curve (AUC) value indicates better overall discrimination.

Finally, the Precision-Recall (PR) Curve is another visual evaluation metric that plots the precision and recall metrics at different thresholds in order to assess the model performance [28]. This is particularly useful in imbalanced datasets where the number of negative samples in the dataset significantly outweighs the positives.

These evaluation metrics allow for a thorough assessment of a binary classification model by analysing various aspects of its performance, including its predictive accuracy and class discrimination. This comprehensive evaluation strategy is

particularly crucial in applications where misclassification costs vary, such as medical diagnostics, enabling informed decisions on model selection and optimisation.

2.5 Conclusion

This chapter provided a comprehensive background on the key concepts, methodologies and techniques relevant to this research. It began with a discussion on the gynaecological disease endometriosis, outlining its characteristics, prevalence rates and symptomology. The chapter then examined the current diagnostic approaches, such as clinical assessments and imaging techniques, highlighting their limitations and the urgent need for more effective detection models. Afterwards, the role of AI in the medical field of disease diagnostics was explored, followed by an overview of ML and DL methodologies critical to this study. Several widely used ML algorithms were introduced, along with a brief discussion on feature engineering and its significance in optimising model performance. Additionally, an overview of a selection of DL architectures is presented, focusing on their structures and advantages in medical imaging problems. The discussion extended to transfer learning, a DL technique that leverages pretrained models to enhance performance while reducing computational cost and training time. Finally, this chapter outlines the main evaluation metrics commonly employed in binary classification tasks, detailing their mathematical formulations and the aspects of the model performance they assess.

By integrating these foundational concepts, this chapter established the necessary theoretical framework for developing an AI-driven diagnostic tool for the diagnosis of endometriosis through ML and DL methodologies. The following chapter builds upon this background by providing an in-depth literature review, examining prior research on AI applications in disease diagnostics, existing solutions for endometriosis classification, and relevant research initiatives focused on the detection of this condition.

3 Literature Review

This section provides a comprehensive examination of the current state-of-the-art research and contributions made in the field of AI in healthcare. It explores various AI-driven methodologies, including ML and DL techniques, that have been implemented in disease diagnostics, specifically focusing on the diagnosis of endometriosis. Additionally, the evaluation strategies employed in these studies are analysed in detail, and the key findings from relevant research contributions are outlined.

3.1 AI In Disease Diagnostics

The integration of AI in disease diagnostics has significantly transformed and revolutionised modern medical practices, enabling enhanced accuracy, efficiency and early detection of various conditions. Recent advancements in AI-driven healthcare solutions have been extensively reviewed in this literature, emphasising their role in clinical decision-making, patient care and treatment planning. Alowais et al. [29] provide a comprehensive overview of the current state of AI applications in healthcare, particularly in the domains of disease classification, predictive analytics and diagnostic automation. The study outlines the potential of AI algorithms in identifying complex patterns within medical data through ML, DL, and data mining techniques. Moreover, this research references multiple studies where AI tools have been developed to aid medical professionals in the diagnostic process of several conditions, including breast cancer detection, skin cancer classification, and acute appendicitis prediction. Furthermore, by highlighting the improved early diagnostics and patient management these systems have made, the potential benefits of an AI-driven tool in advancing the detection of diseases such as endometriosis are reinforced. However, a critical assessment reveals that many promising diagnostic models suffer from a lack of generalisability. Performance metrics are often reported based on highly curated, single-centre datasets, which introduces a significant bias and limits real-world applicability across diverse patient populations and clinical settings. Furthermore, while Alowais et al. focus primarily on the general performance benefits of AI models, they provide limited comparative evaluation of algorithmic approaches across different diagnostic contexts. This makes it difficult to definitively identify the most robust and clinically reliable diagnostic architecture for a given disease.

A systematic review conducted by Umapathy et al. [30] explores the present and prospective applications of AI in medical diagnostics, focusing on its capability to analyse extensive datasets and facilitate more rapid and precise diagnoses. This study

identified various ML and DL methodologies that have been successfully implemented in healthcare to aid medical professionals in diagnosing several conditions, such as acute appendicitis and Alzheimer's disease. It acknowledges the efficiency of AI-driven models in processing and interpreting large volumes of medical information, resulting in early and accurate diagnoses as well as reduced clinical workload. However, the study also addresses the ethical implications and potential challenges associated with AI in healthcare, including concerns related to data privacy, algorithmic bias, and the risk of over-reliance on automated systems. Finally, the authors emphasise that while AI has the potential to enhance medical decision-making, it should function as a supportive tool rather than a substitute for professional clinical expertise, ensuring that human oversight remains an integral component of patient care. Compared with Alowais et al. [30], Umapathy et al. adopt a more critical perspective by integrating ethical and operational considerations alongside technical evaluations. However, both reviews tend to generalise the "efficiency" of AI without dissecting key trade-offs, such as model interpretability versus accuracy, or dataset scalability versus clinical generalisability. For example, DL models may outperform traditional algorithms on benchmark datasets, yet remain difficult to deploy in real-world hospital settings due to their opaque decision-making processes and dependence on high-quality, annotated data. These contrasts underscore a recurring tension between technical performance and clinical practicality within AI diagnostic research.

In a related study, the research presented in [31] examines recent advancements in AI-based diagnostic models and evaluates the feasibility of implementing AI-driven solutions for endometriosis detection. Although no prototype was developed, the study identifies several ML algorithms with strong pattern recognition capabilities that could be considered for implementation in endometriosis diagnosis. AI methodologies such as supervised learning, unsupervised learning, and reinforcement learning are explored, with models including Random Forest, K-means clustering, and Q-learning proposed as potential candidates. A comparative critique of these proposed methods is essential. While Random Forest is a robust classifier, it may not effectively integrate the heterogeneous, multi-modal data required for a complex condition like endometriosis, which often necessitates combining subjective patient history with objective imaging and lab results. Furthermore, K-means is an unsupervised clustering approach prone to noise and relies heavily on the quality and pre-processing of features, which is a significant drawback given the variability in gynaecological data collection. The complexity of Reinforcement Learning methods often makes it computationally prohibitive and difficult to validate in a diagnostic context where real-time clinical action is critical, meaning these models are often compared unfavourably against simpler, more established supervised methods like SVM or LR which offer a better balance of performance and regulatory tractability.

Additionally, the study discusses the diverse range of medical data that could be leveraged to train these models, including biomarkers, medical imaging, and clinical patient records. Moreover, this study references prior research on AI-based diagnostic tools for endometriosis, highlighting the successes and limitations of existing models and offering valuable insights into the challenges associated with the implementation of AI tools in gynaecological diagnostics. Ultimately, this review concludes with the prospect of AI-powered diagnostic systems in reducing diagnostic delays, lowering healthcare costs, and enhancing the overall quality of life for patients suffering from endometriosis.

Nonetheless, unlike the broader reviews by Alowais et al. [31] and Umapathy et al. [30], this study is more exploratory and remains conceptual, as it lacks empirical validation or benchmarking against existing diagnostic AI systems. The absence of a prototype or performance metrics limits the ability to evaluate the proposed models' clinical viability. Moreover, while Random Forest and K-means have demonstrated utility in structured data analysis, their diagnostic accuracy may fall short compared with deep learning models trained on multimodal datasets. This reveals an ongoing need for comparative experimentation to determine which AI architectures are most suited to nuanced conditions like endometriosis, where symptoms are heterogeneous and often under-documented.

3.2 Review of AI Projects for Endometriosis Detection

Sivajohon et al. [32] conducted a comprehensive review of AI technologies applied to the diagnosis and prediction of endometriosis using diverse datasets. The study aggregates the methodologies and findings of 36 unique research projects published between January 2000 and March 2022, providing an in-depth analysis of each approach and presenting the findings through structured tabular representations. This review identified several ML algorithms that were explored in these studies, including LR, DT, SVM, and Natural Language Processing (NLP). Additionally, it categorised the types of data utilised in these systems, such as biomarkers, protein spectra, clinical variables and symptoms, genetic variables, mixed variables, and imaging. Beyond providing a summarisation of the AI methodologies that were employed in these studies, this paper critically analysed the strengths and limitations of each developed model, offering insights into their predictive performance while also identifying the most effective algorithms. Finally, the review concluded with a discussion on future research directions in AI-driven endometriosis diagnostics, emphasising the potential advancements in predictive accuracy, early intervention and improved patient care. Given its extensive coverage of the current state-of-the-art research in diagnosing

endometriosis, this study serves as a foundational reference for the present dissertation, providing valuable insights into AI methodologies, data types, and the diagnostic efficiencies of different models.

3.3 AI in Disease Diagnosis of Endometriosis

Numerous researchers have explored the development of AI-based systems for non-invasive diagnosis and prediction of endometriosis. These studies have implemented a range of ML and DL algorithms, leveraging diverse data sources to enhance predictive accuracy and clinical applicability.

3.3.1 Clinical Data

A study presented in the medical paper [33] proposed a non-invasive diagnostic model based on symptoms and laboratory data collected through data sampling from a research group. Multiple ML algorithms were explored in this research, including LASSO regression, DTs, Naive Bayes and K-means clustering and evaluated using classification metrics such as accuracy, sensitivity, specificity, positive predictive value, and negative predictive value. Notably, the LASSO regression model attained the highest predictive performance and identified seven critical features vital for detecting endometriosis. Namely, the relevant features selected by this algorithm included body mass index (BMI), age of menarche, cycle length, dysmenorrhea severity, contraceptive use, CA125 concentration, and VEGF1 levels. Further experimentation using the LR model narrowed the most significant predictors to dysmenorrhea severity, BMI, and CA125 concentration.

Kleczyk, Yadav, and Amirtharaj [9] developed an AI-powered diagnostic model that utilises historical medical data to predict the likelihood of endometriosis. The study employed LR and XGBoost models trained on 26 months of medical records obtained from the United States healthcare database. The performance evaluation was conducted through classification metrics such as accuracy, precision, sensitivity, specificity, F1-score, and AUC-ROC curve. Additionally, a confusion matrix was constructed to further analyse the models predictive results. Furthermore, this study also established key clinical data features instrumental in the accurate diagnosis of endometriosis, including infertility, hormone imbalances and family history.

Tore et al. [34] investigated the predictive capacity of ML models based on patient age and comorbidities, which are defined as the simultaneous presence of multiple medical conditions. Classification modelling algorithms such as LR, DTs, RF, AdaBoost and XGBoost were developed to detect endometriosis, while Shapely Additive Explanation (SHAP) was employed to determine the most influential

predictive data features. The performance of these models was assessed through the accuracy, precision, sensitivity, and specificity, and AUC evaluation metrics. Notably, this study concluded that the XGBoost model attained the best classification results and outperformed the other ML classifiers. Furthermore, the SHAP analysis determined that the top five predictive factors for diagnosing endometriosis based on age and comorbidities were age, infertility, uterine fibroids, anxiety, and allergic rhinitis features.

In the research published by Zhao et al. [35], various ML approaches were examined in diagnosing endometriosis using serologic data, such as white blood cell count and mean platelet volume. This study evaluated the predictive performance of the RF, DT, SVM, LR, LogitBoost, Naive Bayes, and K-Nearest Neighbours (KNN) algorithms through key classification metrics including, accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and AUC. The results indicated that the RF model achieved the highest classification performance, demonstrating superior diagnostic accuracy compared to alternative algorithms.

3.3.2 Genomic Data

In the scientific study by Pei et al. [36], a diagnostic model was constructed based on genomic data accessed from the Gene Expression Omnibus (GEO) database [37]. This study explored multiple ML classifiers with the aim of detecting endometriosis through biomarkers, including univariate filter LR, LASSO regression, and SVM classification. Additionally, RFE was employed to determine the most relevant genetic features from the dataset. Moreover, the assessment of the developed model consisted of evaluating the AUC-ROC curve and p-values. The main findings of this study show that the models identified three diagnostic markers with strong predictive potential, thereby setting the foundational groundwork for future genetic-based diagnostic models.

Chen et al. [38] further explored ML-based prediction of endometriosis using genomic data obtained from the GEO database [37]. This study investigated three ML algorithms, including LASSO, RF, and LR and identified five glycolysis-related hub genes that were key in detecting the disease. Each model was evaluated using the AUC-ROC scores, calibration plots, and decision curve analysis to assess the diagnostic performance. The findings of this study suggest potential novel approaches in detecting endometriosis using this type of data. However, it acknowledges that further experiments are required to confirm the findings and better understand the mechanisms of glycolysis-related gene regulation of immune cell infiltration.

In the study by Zhand et al. [39], a ML algorithm capable of diagnosing endometriosis through biomarker identification is proposed. Four datasets were derived from the GEO database [37], and an abundance of AI algorithms were explored

for the purpose of this project. Initially, eleven distinct ML algorithms were developed with the aim of identifying the most effective modelling techniques. The mentioned algorithms included LASSO, StepGLM, glmBoost, SVM, Ridge regression, Elastic Net, plsRglm, RF, XGBoost, Linear Discriminant Analysis (LDA), and Naïve Bayes. The study employed model stacking techniques, constructing various predictive model combinations to determine the optimal classifier based on AUC scores. Following the identification of the best-performing models, nine additional ML algorithms were used to evaluate diagnostic gene significance, including RF, DT, SVM, LASSO, XGBoost, NN, KNN, Gradient Boosting Machine, and generalised linear models. Furthermore, the Receiver-Operating Characteristic Curve (ROC) and AUC scores were used to evaluate the effectiveness and predictive accuracy of the model.

Akter et al. [40] conducted a systematic performance assessment of multiple ML classifiers with the aim of detecting endometriosis through biological patterns, specifically transcriptomic and methylomics data. For this research, the dataset was sourced from three independent institutes where laparoscopic procedures were performed. This study applied supervised ML techniques to diagnose the disease, including DT, SVM, RF and Partial Least Squares Discriminant Analysis (PLSDA) models. Furthermore, the evaluation plan consisted of calculating the accuracy, sensitivity, specificity, precision, and F1-score metrics. Additionally, this study also assessed the Matthews Correlation Coefficient (MCC), AUC-ROC curve, and leave-one-out cross-validation approach to ensure model robustness.

3.3.3 Self-Reported Data

In the research by Bendifallah et al. [13], the authors investigated the application of ML algorithms for diagnosing and screening endometriosis based on 16 clinical and patient-based symptom features. These features, selected with the assistance of endometriosis experts, encompassed demographic characteristics, quality of life indicators, and specific endometriosis phenotypes such as dysmenorrhea. The dataset utilised in this study was obtained from the French online health platform Ziwig [41], which is a research program committed to helping the diagnostic problems associated with endometriosis and includes records of patients with symptoms suggestive of the condition. The study's methodology, illustrated in Figure 3.1, involved implementing various ML models, including LR, DT, RF, XGBoost, as well as hard and soft Voting Classifiers. Additionally, the Chi-Square test was used for feature selection to optimise the dataset to contain the most informative symptoms that lead to accurate predictive results. Model performance was evaluated using key metrics such as sensitivity, specificity, F1-score, and AUC. Notably, the models were validated using data collected from a cohort study.

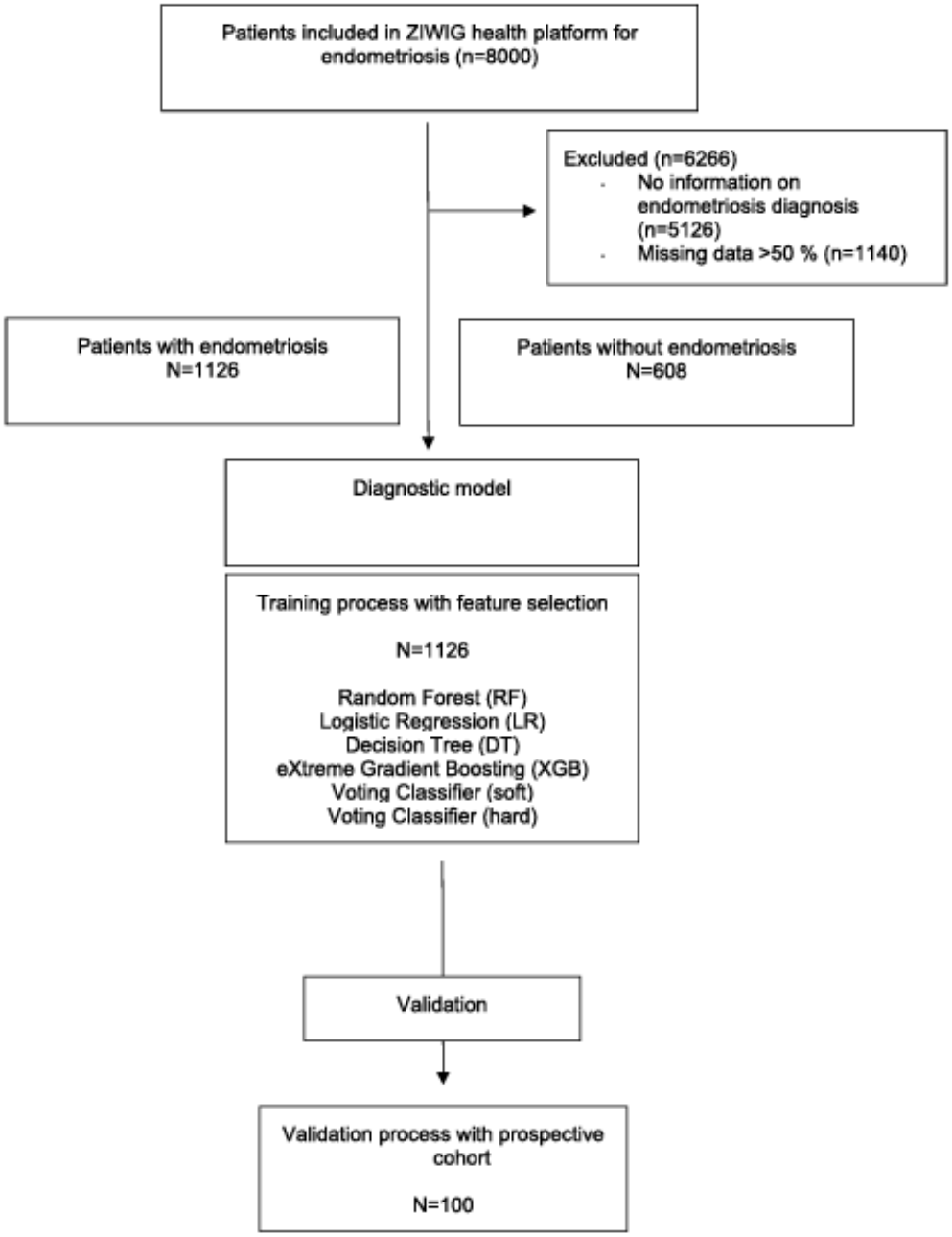


Figure 3.1 Study [13] Methodology Flowchart

While the ML algorithms exhibited mixed results, as depicted in Figure 3.2 and Figure 3.3, the key findings of this study [13] highlight the potential effectiveness of AI-driven tools to facilitate early detection of endometriosis. The study also underlined the feasibility of patient-driven self-assessment tools that could promote awareness and encourage earlier consultations with healthcare providers. However, it acknowledged a limitation in its applicability to asymptomatic individuals, who may remain undiagnosed despite the tool's predictive capabilities.

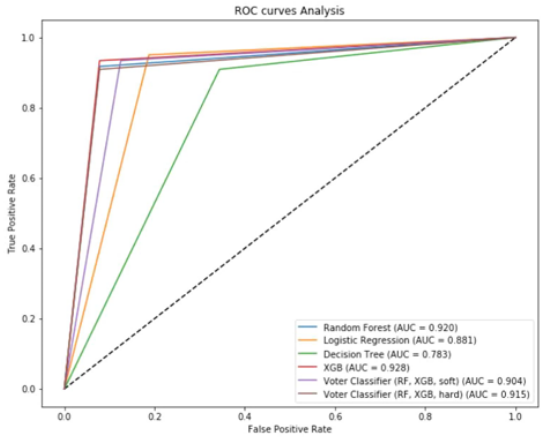


Figure 3.2 Training Set ROC Analsyis

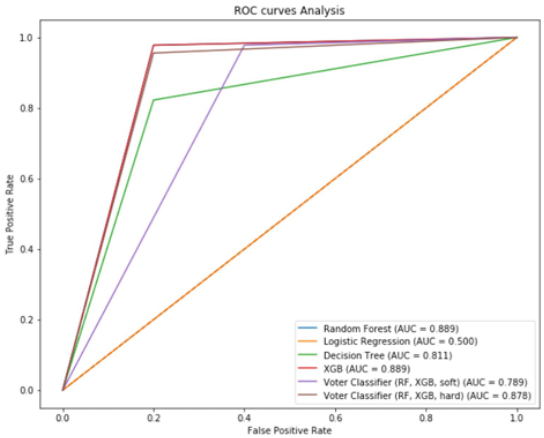


Figure 3.3 Validation Set ROC Analsyis

The two researchers in the scientific study [42] explored the development of a self-diagnostic tool that predicts the likelihood of endometriosis based on patient-reported symptoms, with the aim of reducing diagnostic delays. The dataset was collected via an online survey distributed on Facebook, wherein participants self-reported their experiences with 56 distinct symptoms. Several ML algorithms were explored during this study, including DT, RF, AdaBoost, and Gradient Boosting Classifier (GBC). The performance of these models was assessed using accuracy, sensitivity, specificity, precision, F1-score and AUC-ROC metrics. Additionally, a ten-fold cross-validation procedure was implemented to ensure the robustness and consistency of the models' predictive results. Figure 3.4 illustrates the performance of the developed models according to the different number of features utilised in the training process. This study also conducted feature importance analysis to determine the contribution of each symptom towards the predictive accuracy in detecting endometriosis. This was done by employing the Jaccard Index on the dataset to measure the similarity and identify potential redundancy among the symptoms. A subset of 24 features were determined to exhibit the highest predictive accuracy for endometriosis, including symptoms such as dysmenorrhea, fertility issues and dyspareunia. The AdaBoost model was trained utilising this symptom subset and demonstrated strong predictive capabilities, resulting in a high AUC and F1-scores. This research significantly contributed to the early stage detection of endometriosis by identifying the most influential symptoms that lead to accurate predictions and establishing a structured approach for patient-driven diagnostics. Furthermore, the authors propose that the developed model can be integrated into a self-diagnostic tool, accessible via a website, to provide women with an initial assessment of their likelihood of having endometriosis. This tool has the potential to reduce the time-to-diagnosis by prompting women with a high likelihood to seek further medical evaluation.

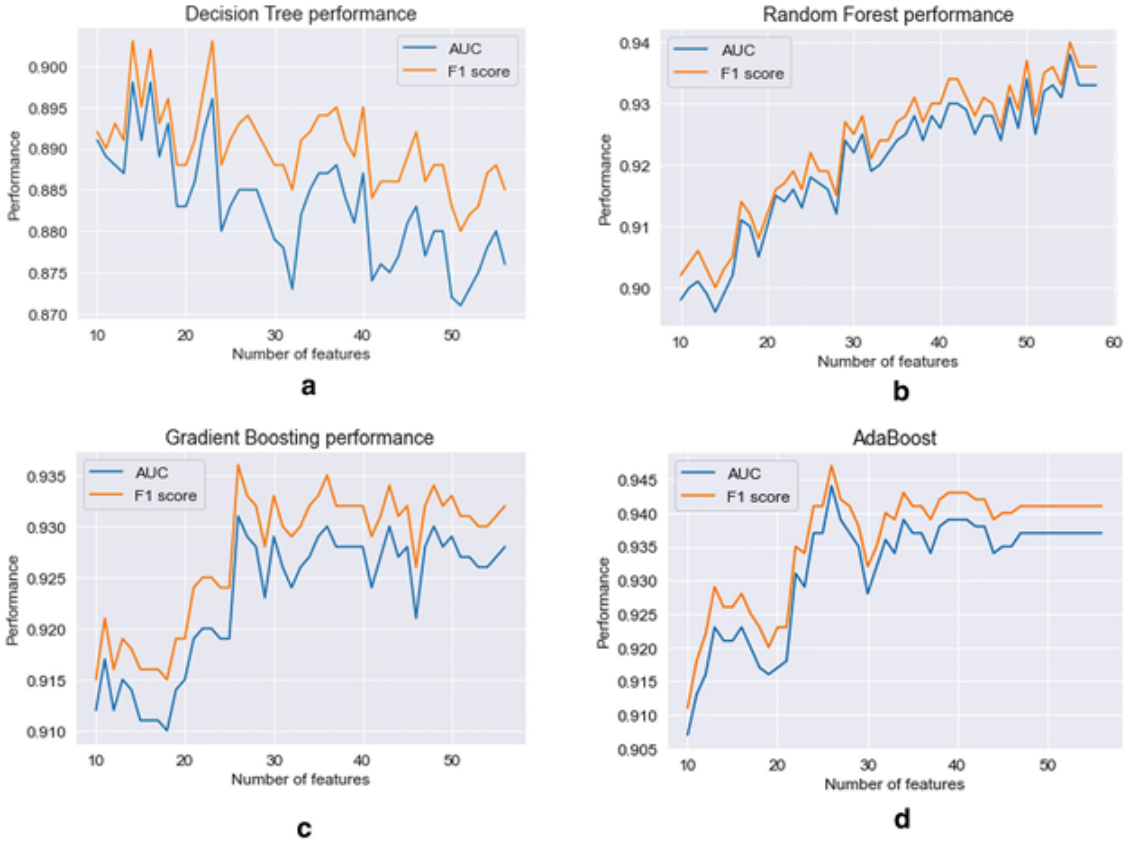


Figure 3.4 ML Model Performances Across Different Number of Features

Zielinski et al. [3] further advanced ML-based symptom prediction by focusing on implementing feature selection techniques to identify the most critical predictors of endometriosis. A LightGBM model served as the benchmark model during this study, and three feature selection methods were applied to optimise the dataset, including the Boruta algorithm, RFE and manually selected symptoms by endometriosis experts. The dataset, comprising approximately 14,000 self-reported questionnaire responses, was sourced from the Invicta Fertility Clinics database [16]. From 258 initial features, 20, 165 and 67 were selected by the Boruta algorithm, RFE and endometriosis experts, respectively. Notably, each technique selected varied features that were identified to be vital in the prediction of endometriosis. Assessment of the models incorporated the calculation of the accuracy, precision, recall, specificity, AUC, and MCC metrics. The final findings of this study uncovered the 20 most predictive symptoms that greatly improve the model's diagnostic accuracy. These symptoms included ovarian cysts, hernias, menstrual pains and infertility.

3.3.4 Medical Imagery Data

Visalaxi and Muthu [43] explored the use of DL methodologies for diagnosing endometriosis through medical imaging. For the purpose of this study, laparoscopic images were obtained from the Gynaecologic Laparoscopy Endometriosis Dataset (GLENDA) [44] to train several pretrained CNN architectures to classify the disease. This was achieved by applying a transfer learning process to the five DL frameworks implemented during this project, including VGG16, ResNet50, InceptionV3, Xception, and InceptionResNetV2. The methodology workflow followed in this study is depicted in Figure 3.5.

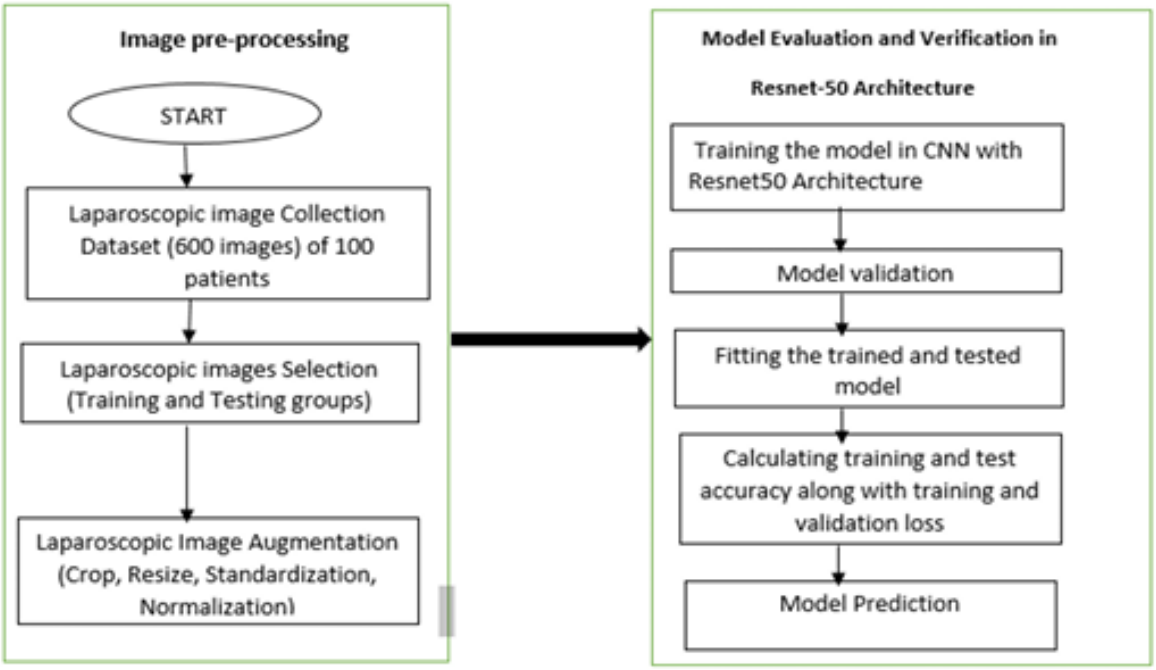


Figure 3.5 Study [43] Workflow

The dataset was split into training, testing and validation subsets, with the training dataset containing 60% of the images. The reminder of the dataset was equally divided between the testing and validation datasets. Additionally, model evaluation consisted of tabulating several classification metrics, including the accuracy, specificity and sensitivity scores of the training and validation datasets. Notably, however, the highest-performing model was further evaluated using metrics such as precision, F1-score, and AUC. Although all models demonstrated high performance, ResNet50 yielded the best results among them, as illustrated in Figure 3.6. Based on these findings, the authors concluded the potential of DL models to assist surgeons in identifying endometriosis through laparoscopic image data, further enhancing diagnostic accuracy.

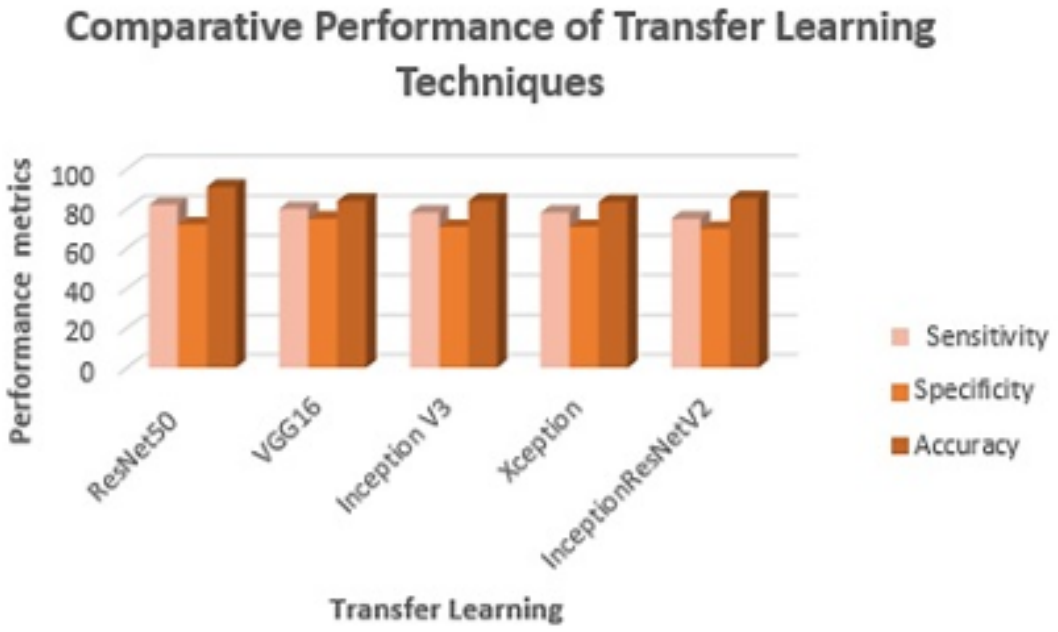


Figure 3.6 DL Model Performance Scores

Another study [45] focused on developing a binary classification mode for diagnosing endometriosis using CNNs applied to laparoscopic video data. This research utilised three datasets for training the DL models, two of which are currently available on the ITEC datasets, including GLENDA [44] and Endometrial Implants Dataset (ENID) [44]. These datasets were preprocessed and augmented to ensure data diversity and model generalisability. Faster and mask variants of Region-based Neural Network (R-NN) architectures with a ResNet50 and ResNet101 backbone were employed to identify the endometrial lesions. Similarly, to the previous study, the training dataset consisted of 60% of the data, while the testing and validation datasets contained 20% each. Moreover, the models were also initialised with pretrained trained weights from the Common Objects in Context (COCO) dataset [46], and transfer learning was conducted to identify patterns for endometriosis classification. Model performance was assessed using Mean Average Precision (mAP) for both bounding box detection and pixel mask segmentation at varying Intersection over Union (IoU) thresholds. The study demonstrated the potential of deep learning techniques in automating the identification of endometriosis lesions from laparoscopic procedures, thus aiding surgical decision-making.

3.4 Endometriosis Detection Research Initiatives

The Finding Endometriosis using Machine Learning (FEMaLe) project [47], funded by the European Union's Horizon 2020 Research and Innovation Programme, represents a

significant multidisciplinary initiative aimed at improving endometriosis detection. The project called for the collaboration of 17 endometriosis experts to design and develop a novel AI-driven diagnostic model to enhance predictive accuracy and minimise diagnostic delays. Through state-of-the-art AI and Big Data technologies, these researchers proposed several innovative tools aimed to aid both patients as well as medical healthcare professionals. One such tool included a sophisticated menstrual cycle monitoring application that was altered to identify individuals at risk of endometriosis. This tool increases awareness of the condition and encourages users to seek timely medical consultations. Additionally, a Clinical Decision Support (CDS) tool was proposed to assist healthcare providers in diagnosing and managing endometriosis more effectively, thereby improving patient care. Furthermore, this research also explored the implementation of an Augmented Reality (AR) surgical application to facilitate surgeons during laparoscopic procedures using DL and ML techniques, further advancing surgical precision and patient outcomes.

Another pioneering initiative is the IMAGENDO Study [48], which is an Australian research program led by Professor Louise Hull at the University of Adelaide in collaboration with the Australian Institute of Machine Learning. This study focuses on developing non-invasive diagnostic tools leveraging DL and ML technology for medical imaging analysis. The main aim of this research is to reduce diagnostic delays by applying AI techniques to detect endometrial lesions in MRI and USG scans, thereby eliminating the need for laparoscopic surgery. By reducing reliance on invasive diagnostic procedures, the IMAGENDO Study presents a promising step towards more accessible and efficient endometriosis detection.

3.5 Conclusion

This literature review examined the role of AI in disease diagnostics, highlighting its potential for early disease intervention and improved patient care. With a particular focus on the early detection and diagnosis of endometriosis, this chapter reviewed existing AI-driven research projects that leveraged various ML and DL techniques that utilise diverse data sources, including clinical records, genetic information, self-reported symptoms, and medical images. Additionally, current research initiatives aimed at advancing AI-based diagnostic tools for endometriosis detection were explored, underscoring ongoing efforts to enhance accuracy and accessibility. The insights from this review provide a foundation for the methodology discussed in the following chapter.

4 Methodology

This chapter outlines the methodological approach employed in this research, detailing the steps taken in order to develop the proposed solution. The first section describes the acquisition and preprocessing of the necessary, extremely sensitive clinical and medical imaging datasets. Following this, a comprehensive explanation of the ML and DL techniques utilised in this study is provided, including justifications for the selected architectures, methodologies and hyperparameter tuning strategies applied.

Software and Libraries

The work for this dissertation was conducted within a customised Anaconda environment, using Jupyter Notebooks and the Python programming language. Various Python libraries were integrated to facilitate the model development, performance evaluation and data visualisation of the implemented AI algorithms. Pandas [49] is one such data analysis library used for the traversal and manipulation of the structured datasets employed in this project. Visualisation libraries such as Matplotlib [50] and Seaborn [51] were utilised to aid in the interpretation of dataset characteristics and model performance through graphical representation. The scikit-learn [6] library was a core component in the development process of this dissertation, as it was utilised in multiple phases of the ML pipeline, including model initialisation, feature engineering and computation of the selected evaluation metrics. Notably, however, the XGBoost [7] library was specifically used for the initialisation and training of the XGBoost classifier. In addition, NumPy [52] and OpenCV [53] were instrumental in the preprocessing of the image datasets. Furthermore, TensorFlow [54] and Keras [5] were the primary DL frameworks used for constructing and training the CNN architectures involved in this research. Further details regarding the software tools, libraries, and dependencies utilised in this study are provided in Appendix A. In addition, the work conducted during this research, including the dataset and code, are available on the EndoAI-Diagnostics GitHub repository[1].

4.1 Data Collection and Preprocessing

The attainment of the datasets required for this research presented unique challenges, primarily due to the ethical considerations and privacy concerns often associated with handling medical data. Although various types of tabular data, such as laboratory test results, genetic variables, and symptomatology, have been employed in ML models to predict the likelihood of endometriosis, medical imaging techniques, including MRI,

[1] https://github.com/britneyv/EndoAI-Diagnostics

USG, and laparoscopy, have also been widely used in DL models for its detection. This dissertation adopts a comprehensive approach by integrating both clinical data and medical imagery to develop ML and DL models for endometriosis prediction and classification. Specifically, publicly available self-reported patient symptom data is utilised for ML-based analysis and predictive modelling, while laparoscopic medical images serve as input for DL classifiers tasked with detecting endometrial lesions. This dual-modality approach enhances the early detection of the disease by providing tools for both patients and medical professionals to facilitate the diagnostic process. The following subsections provide a detailed overview of data collection procedures, preprocessing techniques, and feature engineering methodologies applied in this study.

Despite the methodological rigor applied during data collection and preprocessing, both datasets employed in this research present several inherent limitations and potential biases that may influence the generalisability and robustness of the developed machine and deep learning models. These constraints primarily stem from the data acquisition process, sample composition, and dataset design, all of which can introduce systematic errors or restrict the models' ability to generalise beyond the scope of the original data sources.

4.1.1 Self-Reported Symptom Patient Dataset

The clinical data obtained for this dissertation was originally collected by Goldstein and Cohen for their research in self-reported symptom-based endometriosis prediction using ML algorithms [42]. The data collection process involved the design and distribution of a survey to women diagnosed and not diagnosed with endometriosis via the social media platform Facebook. Participants were asked to respond to a multitude of true or false statements regarding symptoms commonly correlated with the disease. In order to comply with ethical standards and protect patient privacy, demographic and personally identifiable information was not gathered to ensure patient anonymity. In specific, a total of 886 women participated in the survey, comprising of 474 individuals with a confirmed diagnosis of endometriosis and 412 who do not.

Following the data collection process, a thorough analysis was carried out on the acquired dataset to gain a deeper understanding of its quality and to determine the necessary preprocessing steps required in order to prepare it for ML model training. The dataset is composed of 59 columns, with 58 representing symptom-based input features and one indicating the presence or absence of an endometriosis diagnosis. The features of the dataset encompass various symptomology domains pertaining to mental, menstrual, and physical health. Notably, the dataset contains no missing values and consists solely of binary values. Figure 4.1 presents a heatmap illustrating the distribution of the 58 symptoms in the dataset across participants with positive and

negative pathology of the disease.

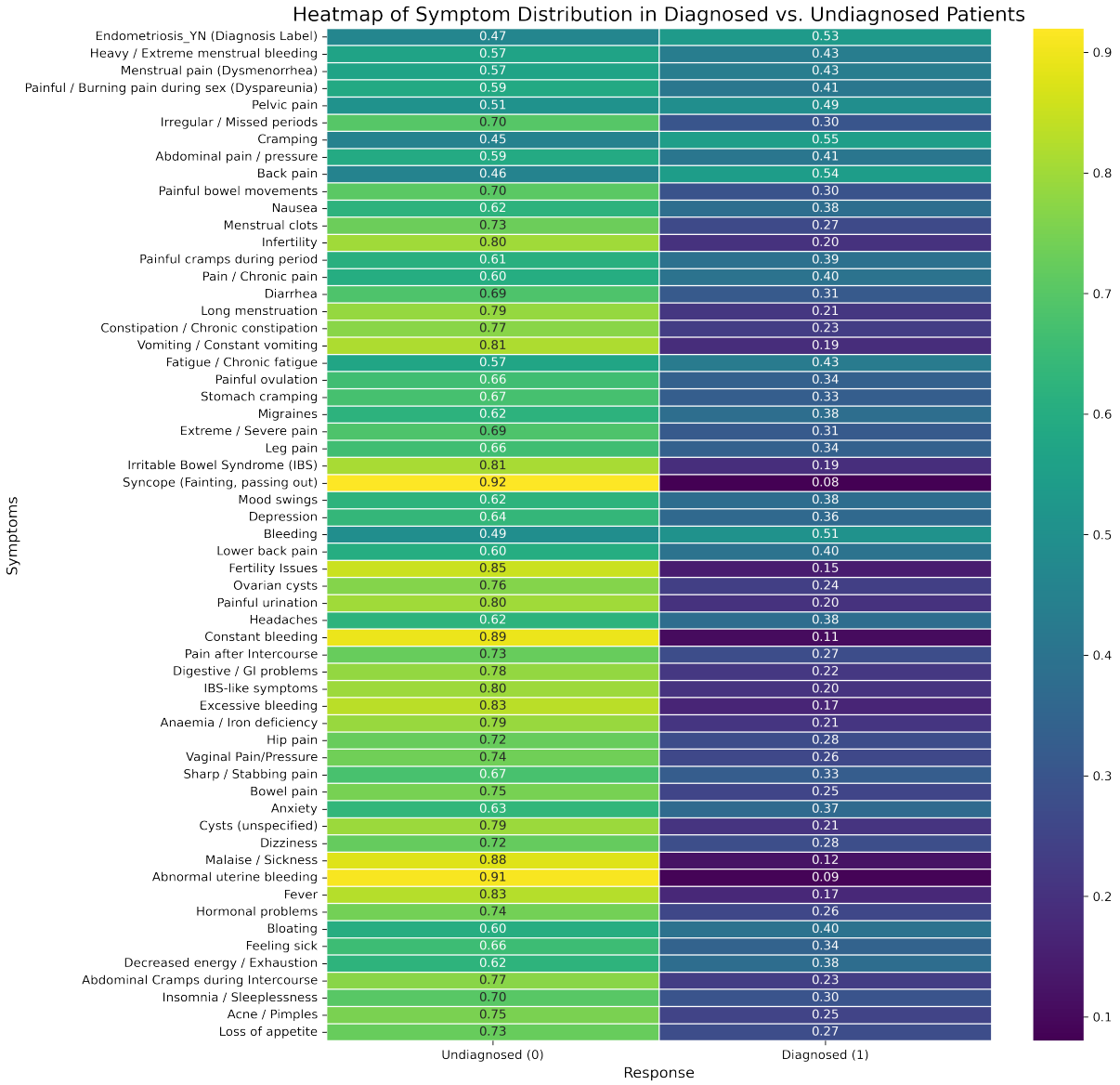


Figure 4.1 Symptom Distribution Heatmap

Despite the dataset's high initial quality and utility for initial predictive modelling, several inherent limitations and potential biases stemming from the collection methodology pose significant threats to the generalisability of any trained ML model. Firstly, the data relies entirely on self-reported responses, which introduces the risk of recall bias and response bias. Participants may underreport or overstate the severity or presence of symptoms, particularly for sensitive topics related to reproductive health, pain perception, or psychological well-being. This subjectivity can affect the reliability of the input variables and introduce noise into the dataset. Additionally, another concern is that the control group may include individuals who have the disease but are undiagnosed, thus contaminating the integrity of the dataset. Secondly, the dataset was collected through social media recruitment which is highly

susceptible to sampling bias. This method inherently restricts the diversity of participants and may not accurately represent the general population. Such an approach may skew the sample toward younger, more technologically literate women, potentially excluding older demographics or individuals without access to social media or the specific group the survey was distributed in. This may ultimately affect the external validity of the ML models and their applicability to broader, clinically diverse populations. Moreover, although the dataset includes a near-balanced distribution of diagnosed and non-diagnosed participants, there remains a potential class imbalance when the data is stratified by specific symptom categories or subgroups. This imbalance can lead to model overfitting toward the majority class, particularly in algorithms sensitive to uneven label distributions. Furthermore, ethical considerations also limit the dataset's comprehensiveness. While anonymisation ensures patient confidentiality, the lack of demographic and clinical covariates, such as age, ethnicity, hormonal status, or medical history, prevents more granular subgroup analyses. These omitted variables could have served as confounders or moderating factors in endometriosis symptom presentation and, therefore, their exclusion may reduce the interpretability and generalisability of the results. Finally, as the dataset comprises binary features representing the presence or absence of symptoms, subtle variations in symptom severity, frequency, or duration are not recorded. This simplification may obscure clinically meaningful patterns that could have enhanced predictive performance or model interpretability.

Data preprocessing involves the transformation of raw, unorganised data into a structured format suitable for ML modelling. This process mainly consists of four key steps, which include data cleaning, integration, transformation and reduction. The first step, data cleaning, is the process concerned with the correction of any errors or inconsistencies in the dataset as well as handling duplicate or missing values. Secondly, the data integration step entails the merging of several datasets to produce a more insightful dataset. Data transformation involves several data conversion techniques, such as normalisation, standardisation and aggregation, to convert the data into an appropriate format suitable for model training. Finally, the data reduction step is focused on simplifying the dataset by applying feature engineering techniques such as feature selection, data dimension reduction or data sampling methods. Notably, given the high quality of the dataset, minimal preprocessing was required. As evident from the data analysis process, no corrective measures were necessary in the data cleaning step as the dataset contained no missing, inconsistent or repeated data. In addition, since only a single dataset was utilised in this study, the data integration step was also not required. Moreover, given that the dataset was already formatted as binary variables, where 0 and 1 represented negative and positive instances, respectively, data transformation was also unnecessary. Notably, however, the column headers

were modified to a standard format more suitable for readability. Lastly, provided that the dataset consisted of a large number of distinct features, several feature engineering techniques were applied to improve model interpretability and performance.

Six feature engineering techniques were implemented with the aim of further understanding the data features, their correlation with one another, and their significance in diagnosing endometriosis. These approaches provide valuable insights into vital symptoms that indicate the presence of endometriosis and will aid in optimising the predictive performance of the developed ML algorithms. The feature engineering techniques employed in this research, implemented using scikit-learn methods [6] and visualised through Matplotlib [50] and Seaborn [51], include the construction of a correlation matrix, chi-square test and feature importance graph, as well as the application of FFS, BFS and PCA. Particularly, the correlation matrix, chi-square test and feature importance algorithms were used to gather general insight into the features with the most significance when predicting endometriosis. Meanwhile, FFS, BFS and PCA were applied to each ML model for the extraction of relevant model-specific features that produce the best predictive performance.

The first feature engineering technique utilised in this study was the correlation matrix. This technique measures the relationship between all possible pairs of features in the dataset and visually represents them in a matrix format. Figure 4.2 depicts the correlation matrix of the entire self-reported symptom dataset. Due to the high dimensionality of the dataset, a filtered version of this approach was constructed to highlight the feature pairs with correlations above the 45% threshold. This matrix is illustrated in Figure 4.3.

To further aid in understanding the results of this approach, annotations were added to display the correlation measurement of each feature pair, and the feature pairs below the 45 percentile that remained due to the other pairings have also been redacted from the graph, thereby producing the highly informative confusion matrix presented in Figure 5.1 in the Chapter 5.

The second feature engineering technique implemented was a filtering feature selection method known as the Chi-Square test. This statistical method determines the relationship between the feature and target variables by calculating their corresponding probability value, p-value. Afterwards, the results are sorted in descending order, and a horizontal bar chart is created depicting the features on the y-axis and the p-values on the x-axis. Annotations to the features with p-values of above 0.0001 were added to better understand the probability measurements.

Another filtering feature selection algorithm applied to the dataset was the feature importance method. Three distinct interpretations of the method were implemented to calculate the general as well as model-specific feature importance. The general feature importance graph is designed to estimate the statistical

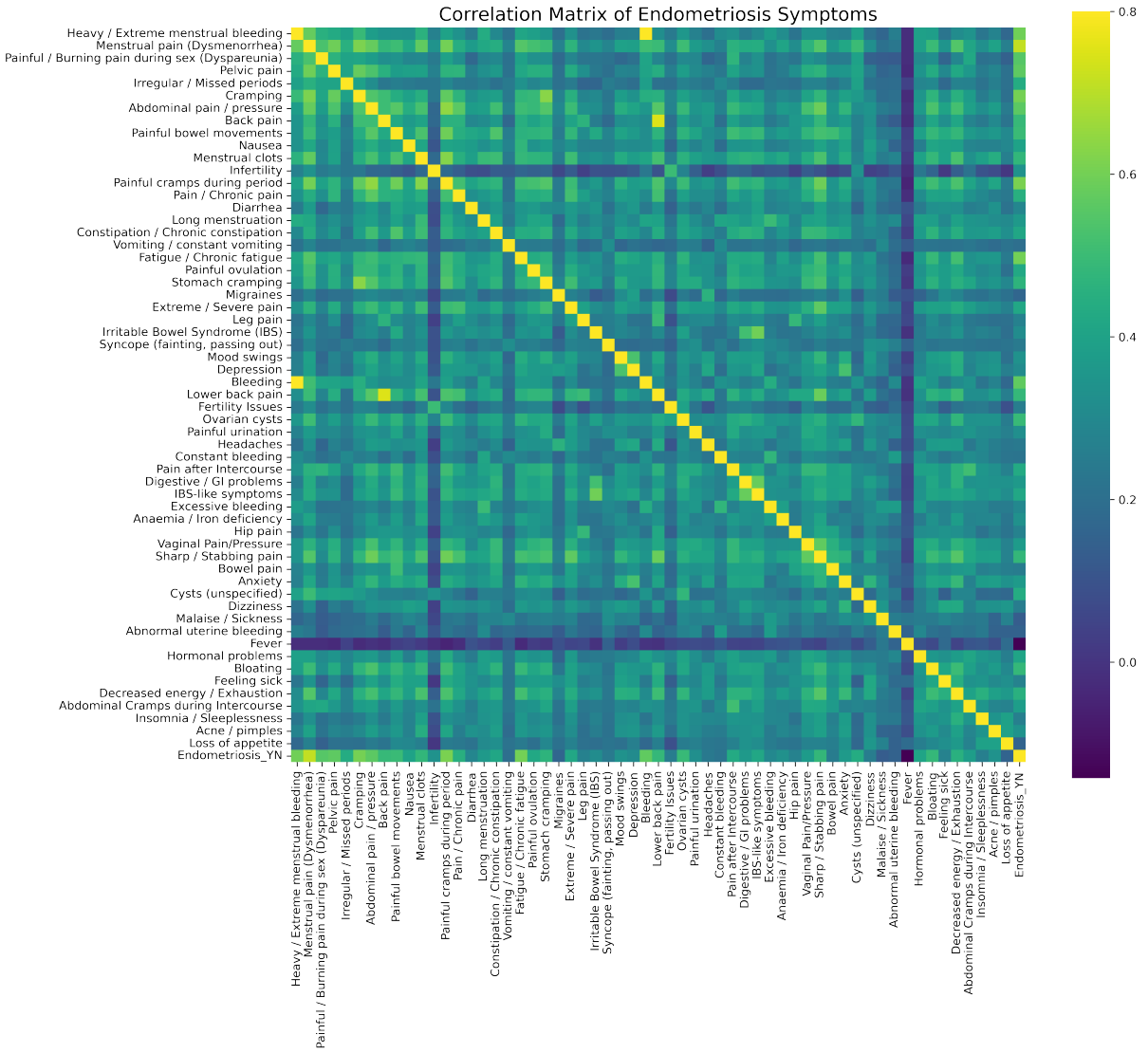


Figure 4.2 Symptom Correlation Matrix

dependency between the features for linear and non-linear models. This is measured using scikit-learn's [6] Mutual Information Classification method on the training features and target variable. Afterwards, the feature importance calculations were sorted in ascending order and presented on an annotated horizontal bar graph with the features on the y-axis and the importance score on the x-axis. A linear-based variation of the feature importance method was implemented to measure the most informative features for the LR and SVM models. The contribution of each feature towards the target prediction is based on the coefficients of the specified models. Hence, the model coefficients are extracted, sorted and plotted on a bar chart similarly to the one previously described. The non-linear-based version of this method, applied to the RF, DT, XGBoost, and AdaBoost models, is identical to the linear version with the exception that instead of the model coefficients, it extracts the model feature importance variables.

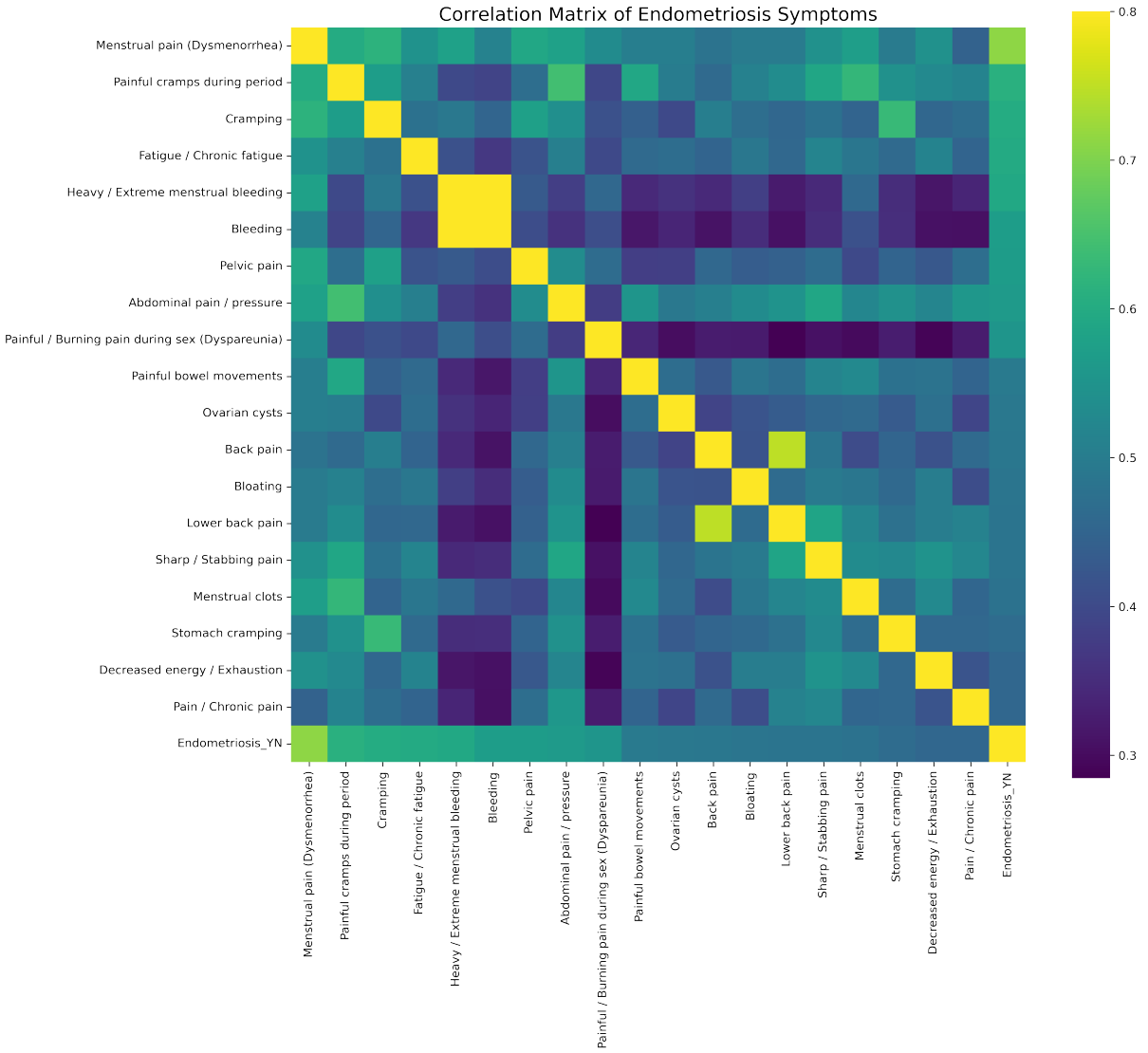


Figure 4.3 Simplified Correlation Matrix

The next feature selection algorithms developed in this project consisted of the FFS and BFS strategies. Often referred to as wrapper methods or greedy algorithms, these techniques traverse through multiple feature combinations to determine a subset of the most important features for endometriosis prediction. FFS begins with an empty set of features and iteratively adds features that increase the model's performance until there is no improvement. Alternatively, BFS, also commonly called backward feature elimination, begins with the complete feature set and repeatedly eliminates the least important feature until no performance gain is observed. Two distinct functions were developed for these techniques that utilise the scikit-learn [6] sequential feature selector method. This method takes the ML models as input, and the backward or forward approach can be adjusted through the direction parameter. Moreover, the number of selected features is set to auto, allowing the method to determine the optimal amount of features in the subsets. Following this, the sequential

feature selector is used to transform the training and testing datasets, and the most important features are selected and returned for model training.

Finally, the dimensionality reduction algorithm, PCA, was employed as the last feature engineering technique in this study. This algorithm aims to reduce the number of variables in a dataset while maintaining vital informative features. The scikit-learn [6] PCA method was used to initialise the algorithm, and a similar approach to the previous function was taken to transform the data and establish a subset of principal component variables. The number of components was adjustable through the function parameters, where it was changed during the experimentation phase of the training process. In addition, the most prominent features in each component were identified and outputted for further investigation.

This concluded the data collection and preprocessing phase for the self-reported symptom dataset. A detailed discussion regarding the feature engineering method's respective results and impact on model performance is presented in Chapter 5 of this document. Additionally, the application of the feature selection and dimensionality reduction techniques in ML model development is elaborated further in section 4.2.2.

4.1.2 Medical Imagery Dataset

The GLENDA dataset, retrieved from the ITEC Datasets repository [44], was the medical image dataset utilised for the DL modelling algorithms in this study. This extensive collection of medical images was curated in collaboration with medical experts to support scientific research on the binary classification and detection of endometriosis. In particular, this dissertation acquired the 1.5 version of the dataset since it had undergone revisions that excluded any unnecessary, unannotated frames from the pathology files and updated the file system structure.

The dataset is composed of over 350 pathological images depicting endometrial lesions captured during 100 laparoscopic procedures, alongside more than 13,000 non-pathological images collected from over 20 surgeries. In addition, the dataset included pathology images classified into four distinct types of endometriosis, further enhancing the model's ability to diagnose the condition in various anatomical regions. Specifically, these endometrial classes included Deep Infiltrating Endometriosis (DIE), which is present in areas such as the rectum, rectovaginal space, or uterine ligaments, peritoneum endometriosis, which affects the lining of the abdominal cavity, ovaria endometriosis, and uterine endometriosis. The pie chart in Figure 4.4 visualises the distribution of cases within the dataset according to these classifications.

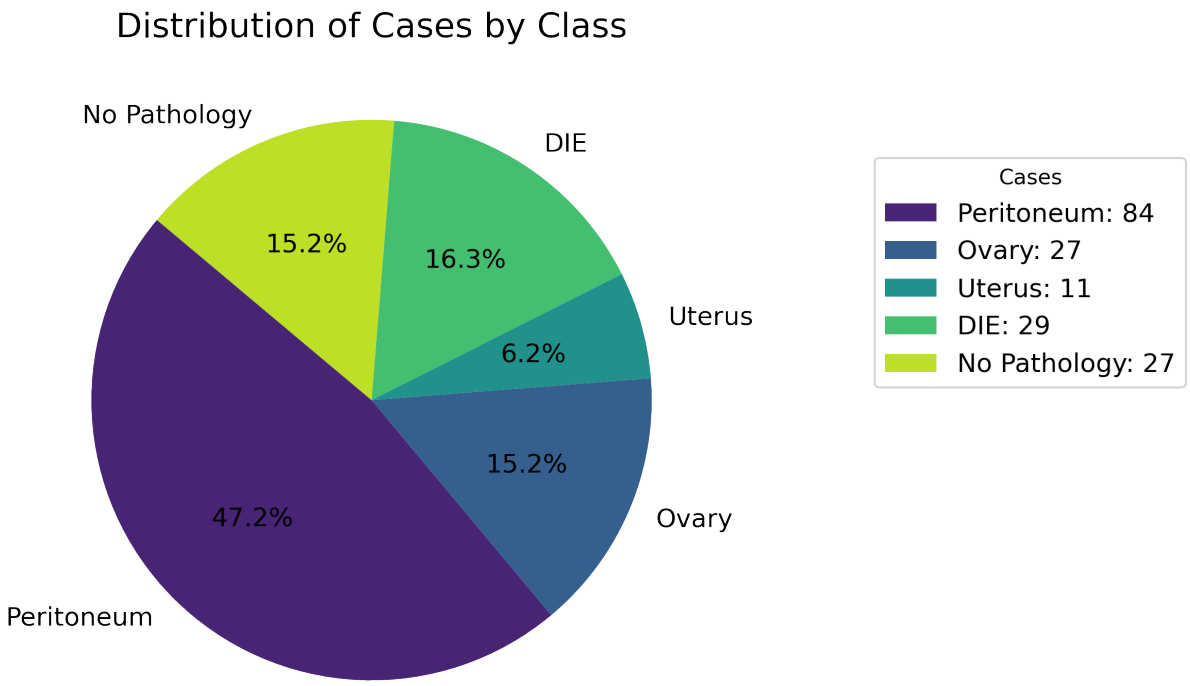


Figure 4.4 GLENDA Distribution Pie Chart

An overview of the pathology dataset is provided in Figure 4.5, illustrating the number of observed cases for each endometrial class, the quantity of frames extracted from these cases, and the amount of annotations identified by medical experts. It is important to note that multiple endometrial lesions, whether from the same class or different classes, may be present in a single frame. In instances where multiple classes are annotated within the same image, the dataset assigns the frame to the class with the largest coverage area.

While this dataset was extensive and clinically curated, it carries several constraints that may impact the generalisability of the DL model outcomes. The first limitation concerns the data collection setting and environment. All images were collected under controlled surgical settings using specific laparoscopic equipment, lighting conditions and imaging protocols. Consequently, the models trained on this dataset may not perform consistently when applied to images captured in different hospitals, with different camera systems, or under variable illumination and visual noise conditions. This is, however, addressed through data augmentation. Additionally, although the dataset includes over 13,000 non-pathological and approximately 350 pathological images, there exists a substantial class imbalance between normal and diseased cases. This disparity may bias the models toward predicting the majority class, thereby reducing sensitivity to pathological findings. To mitigate this, class rebalancing strategies, such as data augmentation and stratified sampling, were

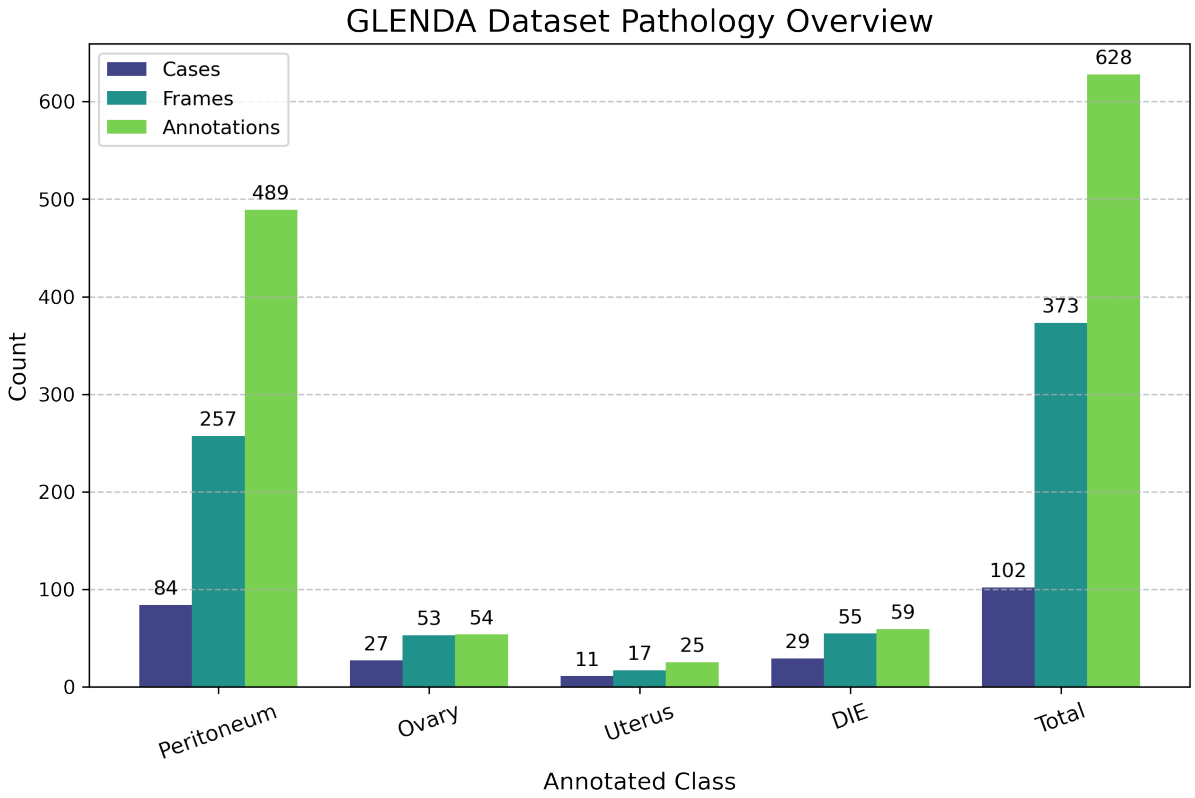


Figure 4.5 GLENDA Dataset Pathology Overview

implemented; however, these can only partially alleviate imbalance effects. Moreover, although the dataset's size is considerable, it still represents a limited number of patients and surgical cases of approximately 120 in total. This relatively small and homogeneous cohort may not encompass the full diversity of endometriosis manifestations, particularly across different ethnicities, age groups, or clinical severities. Therefore, while the dataset provides a robust foundation for model development, further validation using larger and more demographically diverse cohorts would be necessary to confirm clinical applicability. Lastly, another important limitation concerns the annotation process. Another important limitation concerns the annotation process. While the pathology masks and lesion classifications were produced by medical experts, such labels remain subject to inter-observer variability. Differences in clinical judgement or annotation criteria can lead to inconsistencies in ground truth labels, introducing potential label noise that affects the supervised learning process. Moreover, the GLENDA dataset assigns each image to a single class even when multiple lesion types co-exist, which simplifies a complex multi-pathology scenario and may constrain the model's diagnostic precision in real-world conditions.

A further methodological consideration concerns the decision to employ laparoscopic imagery as the input data for DL classification rather than non-invasive imaging modalities such as ultrasound or MRI. This choice warrants careful justification

given the clinical realities of endometriosis diagnosis. In contemporary medical practice, laparoscopic procedures are used sparingly owing to their invasive, costly, and resource-intensive nature, while imaging modalities such as MRI and high-resolution ultrasound have achieved specificity values frequently exceeding 90% in detecting various manifestations of endometriosis. From a purely clinical perspective, the diagnostic value of an automated classifier on laparoscopic imagery is limited, as endometrial lesions are typically readily apparent to trained gynaecological surgeons during the procedure. Moreover, cases of DIE that are not visually discernible to the human eye would equally evade detection by an image-based DL model, thereby constraining its clinical applicability.

Nonetheless, the use of laparoscopic images in this study was motivated by pragmatic and exploratory research considerations rather than by immediate clinical deployment potential. First, laparoscopic images provide a clear and direct visual representation of confirmed pathology, offering an ideal benchmark for testing and validating the feasibility of convolutional neural networks in recognising endometrial tissue characteristics under controlled conditions. This enables a rigorous assessment of model architectures and transfer-learning strategies before extending them to more challenging and diagnostically valuable imaging modalities such as MRI or ultrasound, which contain greater noise, variability, and artefacts. Second, publicly available and ethically cleared laparoscopic datasets such as GLENDA are more accessible to researchers than large-scale MRI or ultrasound datasets, which are often protected by patient confidentiality and hospital governance policies. Consequently, this study positions laparoscopic image classification as a proof-of-concept stage within a broader research trajectory aimed at ultimately supporting non-invasive diagnostic approaches.

Importantly, this methodological decision is not intended to suggest that laparoscopy-based AI systems should replace clinician expertise in intra-operative diagnosis. Rather, the study demonstrates how DL techniques can extract and model visual patterns associated with endometriosis, thereby contributing to the future development of pre-operative, non-invasive diagnostic systems. Future research should, therefore, prioritise the adaptation of these models to ultrasound or MRI datasets, where successful classification could meaningfully aid early detection and triage, aligning more directly with the ethical and clinical motivation of reducing the need for invasive diagnostic procedures.

The dataset's original file structure organises the pathology frames and corresponding annotated mask images into multiple subfolders, with each folder containing a singular image. Meanwhile, the non-pathology folder is structured so that the frames are categorised into subdirectories according to their respective case numbers. Specifically, the non-pathology dataset contains a total of 27 subfolders, while the pathology file system consists of over 300 subfolders, storing the

corresponding annotation mask and frame of each instance in separate subdirectories. Therefore, in order to prepare the datasets for model training, a data restructuring process was implemented. The images were extracted from their corresponding subfolders, consolidating them into two primary directories containing the pathological and non-pathological images, respectively. Subsequently, OpenCV [53] and NumPy [52] methods were employed to convert each image into an array format and resize them to conform to the DL model input requirements of 224 by 224. In addition, the dataset was also shuffled to prevent potential bias in the training process.

As part of an additional experiment conducted on the DL models using this dataset, data augmentation was applied to modify the database further and assess its impact on model generalisation. Although the original dataset was already extensive and provided sufficient information for accurate predictions, data augmentation was introduced to enhance the robustness of the models by artificially increasing variability within the training data. Given that real-world medical imaging conditions may vary due to differences in lighting, camera angles and surgical environments, this technique enables the models to be more resilient to such variations. The image data generator function from the Keras [5] library was utilised to facilitate the implementation of the image augmentations. This function was defined with a range of transformation parameters specifically chosen to introduce realistic variations without significantly distorting the original images. The augmentations included rotations up to 30 degrees to simulate variations in imaging angles, 20% horizontal and vertical shifts to account for slight positional changes, as well as 20% shear transformations to slightly alter image perspectives. In addition, horizontal flipping was enabled to increase diversity of image orientations, and a 20% zooming parameter was defined to create variations in scale and focus. Moreover, brightness adjustments within a range of 0.8 to 1.2 were added to simulate different lighting conditions that may be experienced during the laparoscopic surgeries. Furthermore, a nearest-neighbour fill mode is also initialised to replace missing pixels resulting from these transformations to preserve image integrity. Figure 4.6 depicts the different types of augmentations that may be applied to the images.

The augmentation method was integrated into the preprocessing pipeline to ensure that the images were resized before the transformations were applied. Initially, the function was configured to generate multiple augmented images per original instance. However, this approach consistently led to memory allocation errors due to computational limitations. Despite several attempts to optimise the augmentation function and reduce image size, the available GPU memory could not support multiple augmentations per sample without exceeding hardware constraints. As a result, the experiment was restricted to producing only one randomly augmented image per original sample to maintain computational feasibility. Once augmentation was

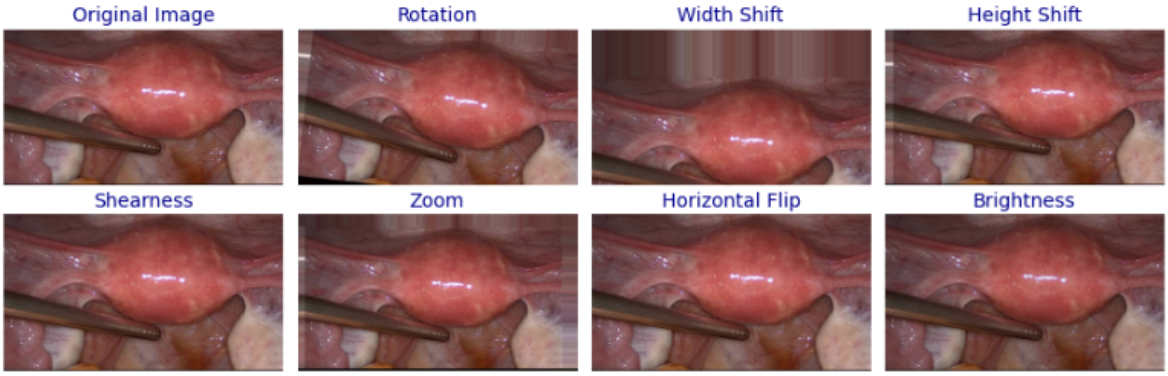


Figure 4.6 Data Augmentation Transformations

successfully applied, the dataset was shuffled, and training commenced. Further details regarding the experimental setup and evaluation of this approach are presented in the Section 4.3.2 and Section 5.3.3, respectively.

4.2 Machine Learning Modelling

For the purpose of this study, six distinct ML models were selected based on a comprehensive review conducted of AI algorithms in disease diagnostics, with a specific focus on endometriosis detection. These models implemented include LR, RF, DT, SVM, XGBoost, and AdaBoost. These models were specifically chosen to ensure a balance between interpretability, computational efficiency and predictive performance. The scikit-learn [6] and XGBoost [7] Python libraries facilitated the development process of this research, providing the necessary tools to initialise, train, and assess each model.

Following the implementation of the base models, an extensive experimentation phase was conducted in an attempt to refine their predictive capabilities. This process involved applying the feature engineering techniques outlined in the Section 4.1.1. As a result, five variations of each classifier were developed, each incorporating different feature selection and dimensionality reduction strategies to optimise the model's ability of identifying patients at risk of endometriosis. In addition, hyperparameter tuning was also implemented to refine and enhance the models accuracy and robustness. Therefore, a total of sixty ML models were trained and assessed with the aim of determining the most effective classifiers for the classification task proposed in this research. The subsequent sections describe the model architectures and the methodology adopted for their respective implementation, experimentation and fine-tuning processes.

4.2.1 The Architectures

The ML model architecture developed during this research can be categorised into linear or tree-based modelling algorithms. The linear models, which include LR and SVM, were selected for their simplicity, interpretability, computational efficiency and relatively short training durations. These models assume that there is a linear relationship between the input features and the target variable that can be represented by a straight line or a hyperplane in higher-dimensional spaces, thereby making them particularly useful as baseline models.

LR is a generalised linear modelling algorithm that operates by utilising a logistic function, sometimes referred to as a sigmoid function, to map predicted values between the range of 0 and 1 on an S-shaped curve. While it is a widely used model for binary classification tasks, its sensitivity to outliers can significantly impact the decision boundary, potentially affecting the model's predictive performance. Meanwhile, SVM classifiers are able to handle both linear and non-linear classification problems through different kernel functions. This model operates by determining the optimal hyperplane that maximises class separation by mapping data points, commonly referred to as support vectors, into a high-dimensional space. The use of a linear kernel in this study ensures that the SVM model functions as a linear classifier that provides a strong decision boundary while minimising the risk of overfitting. Despite these advantages, SVM models can be computationally expensive, particularly when handling large datasets.

In contrast to linear models, the tree-based models explored in this study specialise in capturing complex, non-linear relationships within the data through a hierarchical decision-making structure based on an 'if-then' programming approach. DTs are the simplest of these architectures, generating predictions using a singular decision-based tree structure. Although these models are easy to interpret and visualise, they are prone to overfitting due to the reliance of the prediction being solely based on a single tree model that is highly sensitive to variations in the dataset. To mitigate this limitation, ensambling algorithms were employed. These models utilise multiple tree structures in order to generate stronger and more reliable results by combining the outputs of several weak learners. RF is one such algorithm that uses bootstrap aggregation, or bagging, and feature randomness to train multiple independent DTs in parallel using random subsets of the dataset's features, then aggregates the output through a majority voting protocol. This approach enhances the model's predictive performance and generalisation capabilities while reducing overfitting. However, it's complexity and computational demands make it less interpretable than simpler tree-based models.

Further exploring ensemble learning, this research also incorporated boosting

techniques, which iteratively enhance weak learners to improve classification performance. XGBoost is a gradient boosting algorithm that constructs trees sequentially, utilising a gradient-based optimisation strategy where each new tree learns from the errors made on the previous iteration with the aim of minimising the loss function. Known for its efficiency and superior predictive performance, XGBoost is highly optimised for structured data problems and is resistant to overfitting due to its built-in regularisation techniques. However, this algorithm requires careful hyperparameter tuning to achieve optimal results, making it more complex to configure compared to traditional ML models. AdaBoost is another boosting-based modelling approach and the final ML algorithm developed in this study to predict the likelihood of endometriosis in patients based on self-reported symptoms. This algorithm creates an optimised classifier by iteratively enhancing weak learners based on weighted errors. By focusing on instances that are difficult to classify, AdaBoost improves generalisation but remains highly sensitive to noisy data and outliers, which may negatively impact performance.

The implementation of these models was conducted using established Python ML libraries. Specifically, the LR, SVM, DT, RF, and AdaBoost classifiers were initialised through their respective methods from the scikit-learn [6] library with default parameters, thereby establishing a baseline for model comparison. Notably, the LR model was configured with a maximum iteration count of 10,000 to ensure convergence, while the SVM model was specified to use a linear kernel. Meanwhile, the XGBoost classifier was implemented using the corresponding method from the XGBoost [7] library. The following sections detail the specific training and hyperparameter tuning procedures undertaken to refine and optimise these models.

4.2.2 Model Implementation

The self-reported symptom dataset was divided into training and testing data subsets of a 3:1 ratio, meaning that 75% of the data was used for training and the remaining 25% for testing. Alternative split ratios, such as a 4:1 ratio, were also investigated but showed no impact on the final results. Hence, the 3:1 split ratio was selected for the final implementation, resulting in 664 training and 222 testing data samples. Furthermore, in order to guarantee the integrity and consistency of the model comparison, all models were trained and evaluated using the same datasets within a singular Jupyter Notebook environment.

For each classifier, 5 variations of the ML models were developed. The initial implementation consisted of the base models with no feature engineering or fine-tuning techniques. This variation was vital during this study as it provided the baseline results that will be used for model comparison. The second and third

variations included applying FFS and BFS strategies to the datasets to determine an optimised and informative feature subset of the data with the aim of enhancing the model's predictive performance. Finally, two interpretations of the PCA algorithm with different numbers of components were investigated to thoroughly assess the impact of dimensionality reduction on the dataset.

After initialising the models through their respective classifier methods, the base models were iteratively trained and assessed using a systematic approach. This process leveraged the scikit-learn [6] fit and predict methods to train and generate the model's predictions, respectively. Since no feature selection technique was applied during this process, the models were trained and tested using 58 distinct features. Each classifier was then assessed using a custom function that evaluates the model's predictive performance. Notably, the methodology of the evaluation function and the corresponding model results will be discussed in great detail in Chapter 5.

The second and third variations of the ML models incorporated the feature selection techniques whose implementation details were discussed in Section 4.1. The datasets were transformed using the FFS and BFS algorithms to identify and retain only the most critical features that would provide the best predictive performance of the models. Afterwards, applying the same procedure as that of the base models, the newly processed datasets were used for training and assessment.

The final two model variations developed for this study employed the PCA algorithm that is detailed in Section 4.1 of this document. The training and testing datasets were transformed based on the defined number of components. Multiple iterations were conducted in order to identify the optimal number of components for best model performance. Notably, for the purpose of this research, two PCA models with component values of 58 and 29 will be investigated, where 58 was automatically selected by the model as the most optimised component number and 29 was manually selected based on the number of features determined by the feature selection strategies.

4.2.3 Hyperparameter Tuning

To enhance the predictive performance of the ML models, hyperparameter tuning was employed to identify the optimal parameter configurations for each classifier. This process aimed to refine the models by systematically selecting the most effective hyperparameter values that yield the best classification outcomes. Scikit-learn's [6] GridSearchCV method was utilised for this purpose, enabling an exhaustive search across a predefined set of hyperparameters for each model. This method performed an iterative evaluation, training multiple versions of each model with different hyperparameter combinations and selecting the configuration that achieved the

highest performance score. To ensure robust and reliable tuning, a 5-fold cross-validation strategy was applied, which involved partitioning the dataset into five subsets, iteratively training the model on four subsets, and validating it on the remaining subset. This approach mitigated the risk of overfitting and provided a more generalised assessment of model performance.

For the LR classifier, the hyperparameter grid included variations in the penalty term, which included L1, L2 and Elastic Net regularisation methods, the regularisation strength, five different solvers, as well as adjustments to the maximum number of iterations to ensure model convergence. Additionally, options for fitting the intercept, selecting the number of jobs for parallel computation, and setting different random state values were explored to enhance stability.

The hyperparameter tuning of the RF model focused on optimising the number of estimators, the maximum depth of individual trees, and the minimum number of samples required for node splitting and leaf formation. Furthermore, different feature selection strategies were evaluated through the maximum features parameter, while alternative splitting criteria were examined to determine their impact on classification accuracy. The Boolean bootstrap sampling technique was also considered to assess its contribution to model robustness.

XGBoost's tuning process included the number of boosting rounds, tree depth, learning rate, and subsampling ratios to control the proportion of data used for each tree. Additional parameters such as L1 and L2 regularisation terms were adjusted to prevent overfitting, while variations in the gamma parameter were explored to determine the minimum loss reduction required for further partitioning.

The DT classifier underwent tuning on key structural parameters, including the maximum depth of the tree, the minimum number of samples required for a split, and the splitting criteria. Different splitting strategies were evaluated through the splitter parameter, which compared the best split against the random split method in order to identify the most effective approach.

For SVM, the hyperparameter grid covered different kernel types, including linear, polynomial, radial basis function, and sigmoid, each influencing how the model maps data into higher dimensions. This was done to ensure that the linear-based approach implemented during this study was the optimal kernel function based on the input data. In addition, variations of the regularisation strength were examined. Moreover, the different coefficients were investigated for the gamma parameter, which is ignored for the linear kernel. Furthermore, the degree of the polynomial function was also explored in the case that the polynomial kernel was selected.

Finally, the AdaBoost classifier was tuned by adjusting the number of weak learners, the learning rate, and the choice of the boosting algorithm, where the SAMME or SAMME.R variants were explored. These parameters were optimised to enhance

model adaptability while mitigating sensitivity to noise and misclassified instances.

Once the GridSearchCV method exhausted all possible parameter combinations, the best-performing configuration for each classifier and corresponding variant was selected based on its cross-validation performance. The optimised models were then retrained using the best hyperparameter settings and subsequently evaluated using the same performance assessment framework as the preceding modelling algorithms.

4.3 Deep Learning Modelling

This study implements a variety of DL algorithms with the aim of accurately identifying endometriosis in medical images. Specifically, a total of eleven pretrained CNN models were investigated and modified to address the classification problem of this dissertation. These models were selected based on their demonstrated effectiveness in image classification tasks, particularly in medical imaging applications. The selected architectures include VGG16, ResNet50, ResNet50V2, DenseNet121, InceptionV3, Xception, InceptionResNetV2, MobileNetV3 Small, MobileNetV3 Large, NASNetMobile and EfficientNetV2 B0. These architectures vary in terms of complexity, computational efficiency, and performance. The implementation of these models, along with the methods used to apply transfer learning and refine them, was carried out using the TensorFlow [54] and Keras [5] libraries. Notably, these models will be assessed using the same evaluation methodology as the ML models, which will be covered in more detail in Chapter 5.

The following sections provide a comprehensive overview of the network architectures developed during this research, followed by a description of the implementation process and details regarding the fine-tuning procedure applied to the developed models.

4.3.1 The Network Architectures

This section provides an overview of the DL architectures investigated in this study. The selected models consist of classical, deep residual networks, inception-based, and lightweight optimised CNN models. Each architecture presents unique advantages in terms of computational efficiency, feature extraction capability, and classification accuracy. Table 4.1 presents the top-1 and top-5 accuracy scores of the implemented models alongside their respective size, number of parameters, and depth, as reported by Keras[5].

VGG16 is a traditional CNN architecture compromised of 16 weight layers, employing small 3×3 convolutional filters, max pooling layers, and fully connected layers for image processing. It is known for its simplicity and effectiveness in image

Table 4.1 Keras Reported Network Architecture Details

Architecture	Reported Performance				
	Size (MB)	Parameters	Depth	Top-1 Accuracy	Top-5 Accuracy
VGG16	528	138.4M	16	71.3%	90.1%
ResNet50	98	25.6M	107	74.9%	92.1%
ResNet50 V2	98	25.6M	103	76.0%	93.0%
DenseNet121	33	8.1M	242	75.0%	92.3%
Inception V3	92	23.9M	189	77.9%	93.7%
Xception	88	22.9M	81	79.0%	94.5%
InceptionResNet V2	215	55.9M	449	80.3%	95.3%
MobileNet V3 Small	11	2.5M	88	68.1%	-
MobileNet V3 Large	22	5.5M	110	75.6%	-
NASNetMobile	23	5.3M	389	74.4%	91.9%
EfficientNet V2 B0	29	7.2M	151	78.7%	94.3%

classification problems. However, as seen in Table 4.1, it is computationally expensive and memory-intensive due to its high number of parameters. Its inclusion establishes a baseline for the remaining models and allows for a comparison between traditional deep networks and modern optimised architectures.

ResNet50 is a deep CNN architecture that employs residual connections to bypass one or more layers while mitigating the vanishing gradient problem. With 50 layers, it is built using convolution layers to reduce dimensionality and feature extraction. As shown in the Table 4.1, this architecture provides a balance between model depth, parameter efficiency, and classification performance, making it a strong candidate for medical-based image classification tasks. Meanwhile, ResNet50V2 is an enhanced version of the ResNet50 architecture that applies batch normalisation and ReLU activation before the convolution layers rather than after. This architectural adjustment enhances optimisation, leading to improved convergence and performance. Table 4.1 further highlights the models' slight decrease in depth and increase in accuracy. DenseNet121 is another deep CNN architecture that utilises Dense Blocks to connect each layer with all its subsequent layers to enhance information flow and gradient propagation. Consisting of 121 layers, it is more parameter-efficient than deeper architectures while maintaining competitive performance rates, as seen in Table 4.1.

The InceptionV3 model is an inception-based CNN architecture consisting of 48 layers, widely used due to its balance between accuracy and computational efficiency. This network utilises Inception models that apply multiple convolution sizes in parallel to each in order to identify features in images. It improves upon the original Inception model by employing asymmetric kernels, auxiliary classifiers, batch normalisation, factorised convolutions and label smoothing to enhance training stability and performance. Xception is another inception-based model that replaces standard

convolutions with depthwise separable convolutions to reduce the number of parameters while maintaining high performance. With 71 layers, this is one of the most high-performing and computationally efficient architectures implemented in this study, with the second-best reported accuracy score from Table 4.1. InceptionResNetV2 combines the Inception structure with the residual connection method to improve model performance and stability. This architecture utilises the extensive feature extraction capability of InceptionV3 while employing the ResNet residual learning techniques for further optimisation of the model. Although this model achieves high accuracy, as seen in Table 4.1, it is computationally heavier than all the architectures investigated in this study, with the exception of VGG16.

The MobileNetV3 architecture is an optimised, lightweight CNN model commonly utilised for mobile and edge devices. These architectures employ squeeze-and-excitation modules, hard-swish activations and a combination of depthwise separable convolutions to increase efficiency while maintaining high performance rates. Two variations of this architecture were implemented for study. The first being the MobileNetV3 Small, which is specifically tailored for lower computational power applications where efficiency takes priority. While the second model consisted of the MobileNetV3 Large variation, which is designed to handle more complex tasks and offers higher accuracy rates while preserving efficiency. The model performance rates and computational cost are further depicted in Table 4.1. NASNetMobile is another efficient CNN architecture that balances performances and computation cost. This is achieved through the use of the NAS method, where blocks of the CNN models are searched through Reinforcement Learning. Through the development of separable convolutions, learned architectural patters and a Scheduled Drop Path regularisation technique, this model aims to maximise performance with fewer parameters. As demonstrated in Table 4.1, this architecture has the second-lowest number of parameters and third-lowest size, making it a highly efficient model suitable for resource-limited environments. EfficientNetV2B0 is the final lightweight DL architecture explored in this study. This version of the EfficientNet architecture combines depthwise convolutions, squeeze-and-excitation black and a progressive learning strategy with the aim of optimising training speed and parameter efficiency. In addition, it utilises regularisation techniques to compensate for the performance of the model. Therefore, it is considered as one of the fastest and most accurate DL architectures.

The following sections detail the methodology applied to implement, train, and refine the abovementioned DL architectures to accurately detect endometrial lesions from laparoscopic images.

4.3.2 Model Implementation

Following the data preprocessing steps outlined in the Section 4.1.2, the image dataset without data augmentation comprised of precisely 25,682 samples. Of these data samples, 13,438 images correspond to the negative pathology cases, while the remaining 12,244 instances represent the positive pathology cases. This dataset was partitioned into training and testing subsets using a 4:1 ratio, resulting in 20,545 allocated for training and 5,137 for testing. With respect to the dataset derived after data augmentation, a total of 51,364 data samples will be utilised for model development. The augmentation process effectively doubled the original dataset, yielding to precisely 24,488 images classified as positive for endometriosis and 26,876 classified as negative for the disease. Using the same split ratio of 4:1, 14,091 images will be utilised for training the models, while 10,273 cases will be used for testing. This consistent partitioning approach ensured comparability between the original and augmented datasets during model evaluations.

To maintain reproducibility and prevent bias from data partitioning, the random state parameter of the split method was set to 42. Henceforth, both the original and augmented datasets underwent identical training procedures to ensure consistency. Moreover, in alignment with the ML implementation process, the DL models were trained and tested using the same dataset splits within a single Jupyter Notebook environment, thereby standardising the experimental framework and enhancing result reliability.

The implementation of the DL modelling algorithms consisted of the construction of a custom function to streamline the training and experimentation process of the selected architectures. This function accepted input parameters, which included the model architecture, loss function and activation function, applied transfer learning effectively and returned a trained classification model that detects endometrial lesions in laparoscopic images. Specifically, this function automated model definition, compilation, training, evaluation, and visualisation, thus ensuring procedural uniformity across architectures. As the models were created and trained in succession, the function began by clearing the Keras session to remove any residual computational graphs from previous training runs, thereby mitigating memory leaks and preventing performance degradation. The model was then initialised through the model name parameter with the appropriate input shape, fixed at (224, 224, 3), and loaded with the pretrained ImageNet weights. The three fully connected layers at the top of the network's base architecture were removed, and a global average pooling feature extraction technique was applied to extract high-level feature representations, producing a 2D tensor output. In accordance with Keras' [5] standard transfer learning procedure, the base model was frozen by setting its Boolean trainable attribute to

False, ensuring that only the newly added layers would undergo training. Afterwards, a sequential model was constructed atop the frozen feature extractor, incorporating three additional layers, which included a fully connected dense layer with 128 units and ReLU activation, a dropout layer with a rate of 0.3 to mitigate overfitting, and a final dense output layer of size two for binary classification, where the activation function is configurable through the function parameters. This design ensured that the feature extraction layers remained static while only the classification head adapted to the specific diagnostic task.

Upon model construction, a summary of the updated architecture was generated, followed by compilation. The constructed compiler was developed using the Adam optimiser, with a fixed learning rate of 0.001, an accuracy evaluation metric, and a parameter-configurable loss function which was set to categorical crossentropy. To prevent overfitting, an early stopping mechanism was implemented, monitoring validation loss and halting training if no improvement was observed over three consecutive epochs. Additionally, the best-performing model weights were restored post-training to ensure optimal predictive performance. The training process followed a batch-wise approach, running for a maximum of 50 epochs with 30% of the training data allocated for validation. The training history was stored programmatically, and the elapsed training duration was automatically computed and printed in hours, minutes, and seconds for reproducibility. Accuracy and loss curves were plotted for each model using a secondary custom plotting function to visually assess model convergence and stability.

Each of the architectures detailed in the preceding subsection was iteratively initialised and trained using the described custom function. Although various loss and activation functions, such as binary cross-entropy and ReLU, were explored during the experimentation phase of this project, categorical cross-entropy and softmax were ultimately selected as the final functions. This decision was made due to the minimal performance improvements observed when alternative functions were employed. Following the completion of the training process, all trained models were saved in the .keras format and evaluated according to the evaluation plan outlined in Chapter 5.

Notably, the training process for the DL models utilising the data augmentation-based dataset introduced significant computational challenges compared to the models trained on the original dataset. Due to the substantial increase in data volume, the computational demands in terms of memory allocation, processing power, and training duration were significantly higher. This increase led to frequent occurrences of system instability, including Python kernel crashes, memory overflow errors, and execution timeouts. These issues were primarily attributed to the excessive GPU and RAM utilisation required to process the enlarged dataset during model training. Despite implementing various optimisation strategies, such as reducing

batch sizes and leveraging early stopping mechanisms, these computational limitations persisted, restricting the number of models that could be successfully trained using this dataset. Out of the eleven DL architectures initially considered, only eight were able to complete the training process. These models included EfficientNetV2B0, DenseNet121, ResNet50V2, InceptionV3, Xception, InceptionResNetV2 and both MobileNetV3 Small and Large.

4.3.3 Hyperparameter Tuning

To optimise the performance of the DL models, a systematic hyperparameter tuning process was conducted using the Keras [5] RandomSearch Tuner method. This approach was employed to automate the exploration of various hyperparameter configurations and determine the optimal combination that yields the highest validation accuracy. The tuning procedure was applied to all DL models implemented during this study, including the models trained on the augmented dataset, thereby ensuring a comprehensive evaluation of the developed architectures.

Hyperparameter tuning required adjustments to the previously defined custom function that initialised and trained the DL models. This new function was implemented to construct each model dynamically according to the hyperparameter search space defined by the Keras-tuner module. The pretrained base model is loaded with ImageNet weights, and its layers were frozen to preserve the learned feature representations during the training process. The model architecture was then extended by adding a fully connected dense layer where the number of neurons was set as a tunable parameter with a predefined range of 64 to 256, incremented in steps of 64. Afterwards, the dropout layer's dropout rate was also tunable, with parameters ranging from 0.2 to 0.5 in increments of 0.1. The final output layer comprised a softmax-activated Dense layer with a neuron count equal to the number of target classes. In addition to the network topology, the optimisation parameters were also tuned. Both the optimiser type and the learning rate were treated as hyperparameters, where the optimiser was selected from either Adam or RMSprop, and the learning rate was varied across three discrete values of 0.0001, 0.001, and 0.01. Each configuration was compiled using the categorical cross-entropy loss function and accuracy as the performance metric.

The RandomSearch tuner method was configured to explore the hyperparameter space with a maximum of five test runs, where each trial tested a different combination of hyperparameters. While this implies that not all parameter configurations will be evaluated, this method balances computational costs and training duration while attempting to optimise the predictive performance of the models. During the search phase, the models were trained for 10 epochs on 70% of

the training data, with 30% allocated for validation, where an early stopping callback monitored the validation loss and halted training after three consecutive epochs if no improvement was observed, then restores the best model weights. Once the search was complete, the best hyperparameter configuration was extracted and a new model was rebuilt with these optimal parameters. The final optimised model was then trained for 20 epochs using the same validation split and early stopping mechanism employed during tuning. Following training, predictions were generated on the test dataset, and the model performance was evaluated. This process was repeated for all the DL architectures as well as the models trained on the augmented dataset.

The use of automated hyperparameter tuning enhanced both the objectivity and reproducibility of model optimisation. By restricting the number of trials and search epochs, the approach achieved an effective balance between computational feasibility and model performance, ensuring the fine-tuned networks were robust and generalisable.

4.4 Conclusion

This chapter provided a comprehensive overview of the methodology employed in developing AI models for diagnosing endometriosis using clinical and medical imagery data. The data collection and preprocessing stages were outlined, highlighting the feature engineering and data augmentation techniques applied to the self-reported patient symptoms and laparoscopic image datasets, respectively. Additionally, the ML modelling process was described, providing an overview of the selected algorithms, their implementation, and the hyperparameter tuning strategies used to enhance predictive performance. Similarly, the DL methodology was detailed, discussing the CNN network architectures employed in this study, their respective implementation through transfer learning, and the optimisation techniques employed to refine model accuracy.

In addition to technical performance, the ethical implications and practical deployment considerations of integrating these ML and DL systems into clinical settings must be explicitly recognised and addressed. Great care must be taken when handling patient data to ensure patient anonymity and privacy, as it is sensitive in nature. Additionally, while the pursuit of optimal accuracy is paramount, the ultimate clinical utility rests on transparency and safety. The different model types employed present a trade-off where traditional ML models offer greater explainability, which is vital for building physician trust and meeting regulatory requirements for diagnostic devices, whereas the high-performing DL models pose a challenge due to their inherent opaqueness. For deployment, the models must be designed as

decision-support tools, not autonomous diagnostic agents, to ensure human oversight remains the final safeguard against error and over-reliance. Additionally, deployment should also include interpretability tools presented in clinician-facing interfaces, along with clear documentation of what the model can and cannot detect reliably. Future evaluation must rigorously assess not only technical metrics but also the ethical fairness and safety profile of both the ML and DL models before considering any translational pathway into clinical practice.

The following chapter will present the evaluation of these models, assessing their diagnostic effectiveness and clinical applicability.

5 Evaluation

This chapter is focused on presenting a rigorous and comprehensive evaluation of the ML and DL implemented in this study, aligning with Objective 4. It begins by outlining the evaluation plan adopted to thoroughly assess the model's performance, reliability and computational efficiency. Following this, the evaluation of the ML models is presented, where the predictive performance of the developed algorithms is assessed. Finally, DL model analysis is performed to examine and assess the performance of the implemented architectures, focusing on their predictive capabilities and computational feasibility.

A key objective of this evaluation is to systematically determine the most effective approach for detecting endometriosis while balancing computational efficiency, predictive performance, and generalisability. Beyond analysing model accuracy and robustness, this evaluation also incorporates a comparative analysis against state-of-the-art benchmark models discussed in the Chapter 3. Therefore, through this rigorous assessment, the most efficient, reliable, and clinically applicable AI-powered techniques for the early detection and diagnosis of endometriosis is determined.

5.1 Evaluation Plan

To ensure an exhaustive and systematic assessment of the implemented models, this study employs a wide range of evaluation metrics to effectively measure the classification performance, reliability and computational efficiency. The evaluation process was structured into two sections, focusing separately on the ML and DL models, respectively. In addition to the comprehensive analysis of the model performances, a comparative assessment against state-of-the-art models discussed in the Chapter 3 is conducted, thereby contextualising the findings of this study within the field of endometriosis detection.

The performance of both ML and DL models was evaluated using several classification metrics, previously introduced in the Section 2.4, to capture various aspects of their predictive performance. The accuracy metric was utilised as a general indicator of the model's ability to make correct diagnostic classifications. However, given the potential class imbalance and the critical nature of medical diagnosis, additional evaluation metrics were incorporated to ensure a rigorous and unbiased assessment. Precision was employed to evaluate the proportion of correctly classified positive cases, whereas recall was examined to measure the model's ability to correctly identify actual positive cases, thereby assessing the false positive and false negative

rates of the models. Since both metrics are crucial in medical applications, the F1-score was calculated to provide a balanced representation of precision and recall. Additionally, the AUC-ROC was computed to assess the model's ability to discriminate between positive and negative cases, thereby providing an aggregate measure of performance across varying classification thresholds. Moreover, to provide further insights into the performance of the implemented models, a confusion matrix was generated to visually represent classification outcomes. Furthermore, a classification report was produced to summarise the precision, recall and F1-score performance across the different classes.

Beyond the abovementioned classification metrics, further evaluation criteria were employed to enhance the robustness of the ML model performance analysis. The standard deviation was recorded to assess model stability and consistency across multiple iterations. Moreover, the ROC curves and PR curves were plotted to provide visual insight into the trade-off between sensitivity and specificity as well as precision and recall, respectively. These visualisations were particularly useful in examining how different models performed under varying decision thresholds.

With respect to the DL architectures, additional evaluation metrics were introduced to further assess the learning dynamics and computational efficiency of the developed models. The training time of each model was recorded to ensure that high-performing models were also computationally feasible, as efficiency is a key consideration for real-world medical applications. The progression of learning was also analysed through the plotted accuracy and loss curves that illustrated the improvement or deterioration over training epochs. In particular, the accuracy curve provided insights into the model's progression in predictive accuracy, while the loss curve indicated convergence rates and optimisation stability. Furthermore, the model summary was extracted and examined to analyse the trainable and non-trainable parameters, offering a deeper understanding of model complexity. Given that this research employed transfer learning, the trainable parameters represented the layers that would be updated and fine-tuned during training, signifying the model's adaptability, while the non-trainable parameters remained fixed to show the knowledge transfer process from the pretrained networks.

To ensure a systematic and structured comparison of the developed ML and DL models, a standard evaluation framework was implemented using scikit-Learn [6] methods for metric computation, while Matplotlib [50] and Seaborn [51] were employed to generate the visualisation analysis techniques. The results of these assessments were also compared against state-of-the-art benchmark models identified in the Chapter 3, ensuring an objective and data-driven selection of the most effective algorithm. By adhering to this structured evaluation strategy, this study aimed to identify the most accurate and computationally efficient diagnostic model, contributing

valuable insights into the feasibility of AI-driven diagnostic tools for endometriosis detection in clinical applications.

5.2 Machine Learning Assessment

The evaluation of the ML models in this study encompasses a comprehensive analysis of their predictive performance, interpretability, and computational efficiency. Feature engineering techniques played a crucial role in optimising the model performance as they directly influenced the quality of the input data and, consequently, the accuracy of predictions. This section outlines the experimentation process conducted on the ML algorithms employing several feature engineering methods and hyperparameter tuning. Additionally, it presents a detailed discussion of the results obtained from each model, highlighting the impact of the feature engineering techniques on the final predictive outcomes. Furthermore, this evaluation includes a comparative analysis against state-of-the-art ML models designed with the premise of diagnosing endometriosis through self-reported data, as established in the Chapter 3.

5.2.1 Feature Engineering Filter Methods Assessment

A wide range of feature engineering techniques were implemented during this study, beginning with filter-based methods such as the correlation matrix analysis, chi-square test and feature importance rankings. These techniques were applied during the data preprocessing stage to assess the relevance of individual features in relation to the target variable. Additionally, feature selection and extraction strategies were employed to refine the feature space, including FFS, BFS, and PCA. Notably, the selected features from the FFS and BFS algorithms were subsequently compared with those identified by the filter-based methods to evaluate their consistency and effectiveness.

The first filter-based technique utilised was the correlation matrix, which was employed to analyse relationships between features and identify the most highly correlated variables. As described in the Chapter 4, due to the large number of features in the dataset, low-correlation features were removed to enhance readability. Hence the correlation heatmap, presented in Figure 5.1, reveals the strongest positive correlations within the dataset.

Out of 58 initial features, the correlation matrix identified 19 features with high correlation values. Notably, the menstrual pain feature exhibited the highest correlation with endometriosis diagnosis, exceeding the 70% threshold. This is followed by painful cramps during period and cramping symptom features, with correlations of above 60%. Other significant features included chronic fatigue, bleeding, heavy menstrual bleeding, pelvic pain, abdominal pain and dysmenorrhea, all

of which displayed correlations exceeding 55%. Features such as back pain, bloating, and decreased energy levels showed moderate correlations, ranging from 45% to 50%. Interestingly, fever was the only feature to exhibit a negative correlation with the target variable.

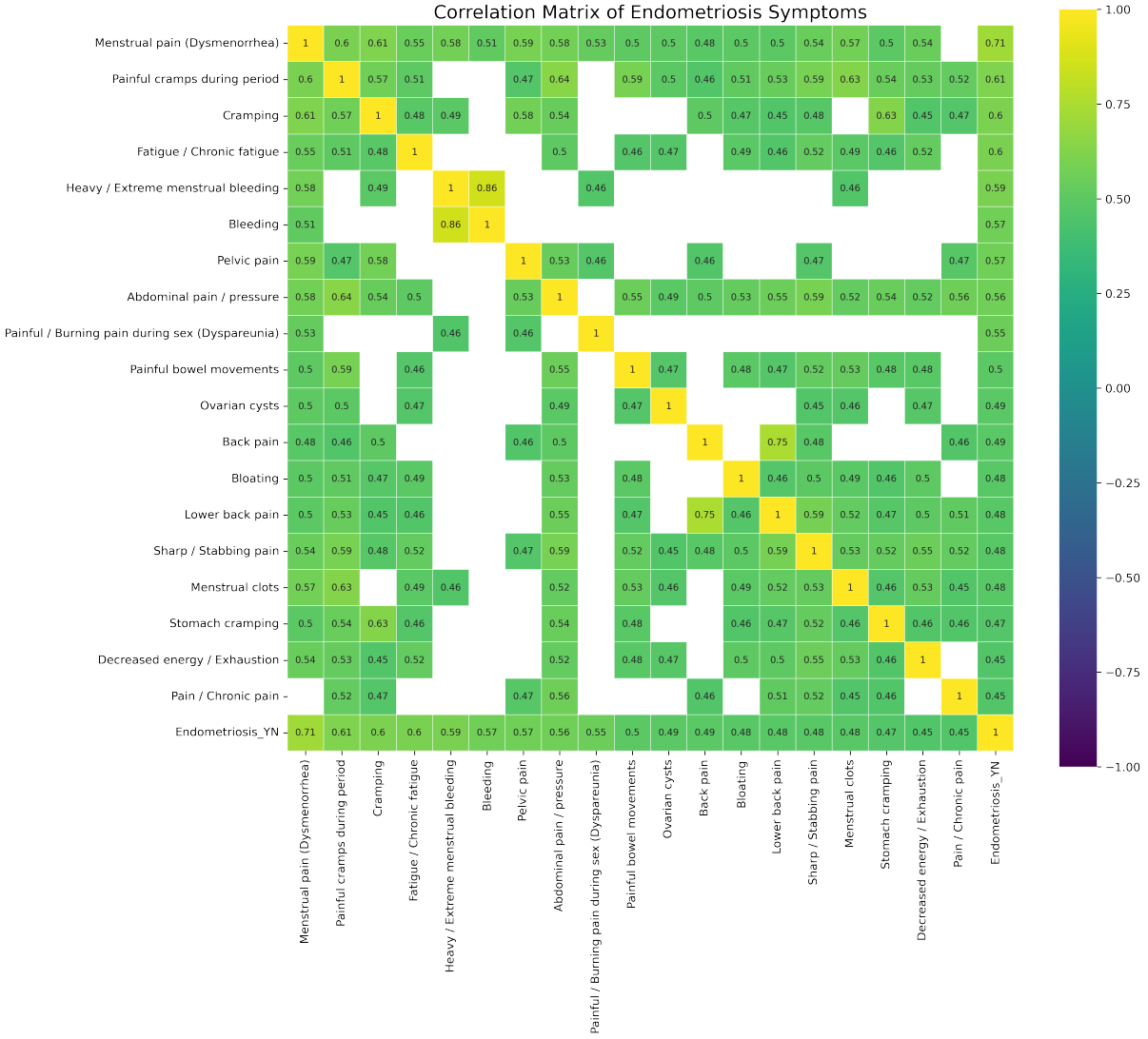


Figure 5.1 Optimised Correlation Matrix

The second filtering method applied to the dataset was the chi-square test, which measures statistical dependence between categorical features and the target variable using p-values, where a lower p-value indicated higher feature significance. As illustrated in Figure 5.2, this test confirmed that most features contained valuable diagnostic information to detect endometriosis. Among the least significant features were loss of appetite, sickness, migraines, abnormal uterine bleeding, and fever. Notably, the chi-square test results were largely consistent with those obtained from the correlation matrix, reinforcing the importance of features such as menstrual pain, painful cramps, fatigue, and abdominal pain.

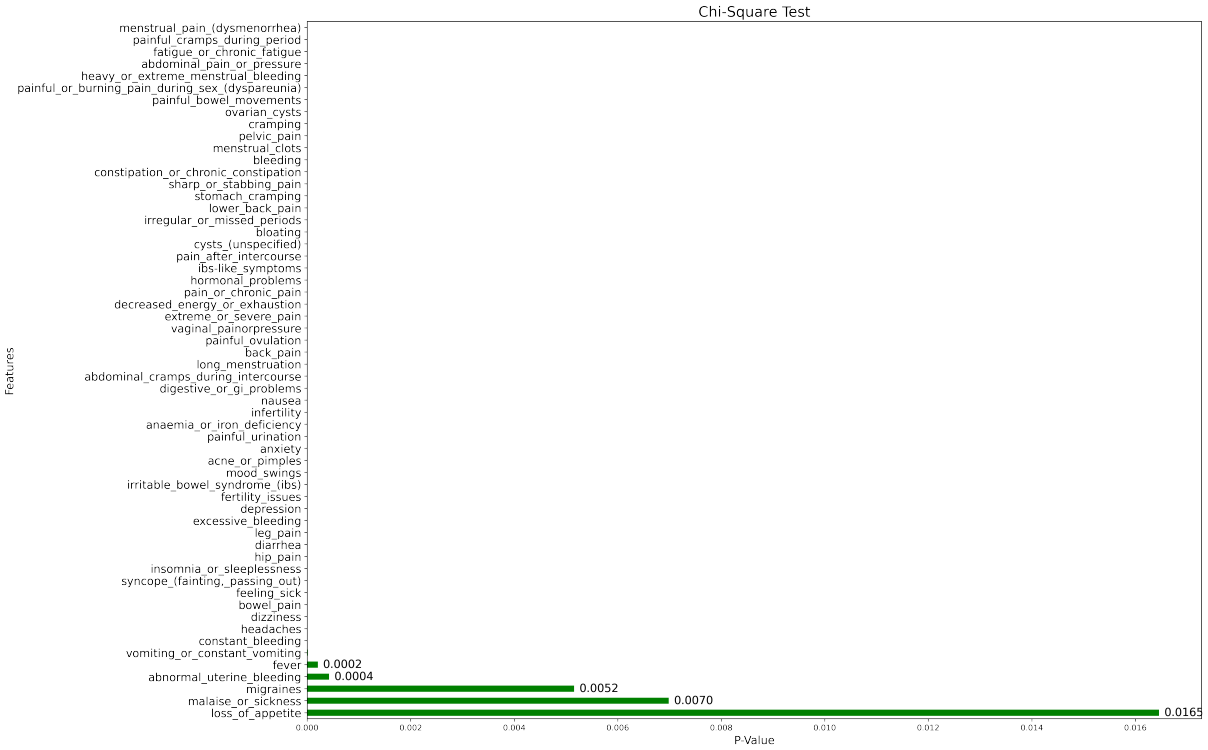


Figure 5.2 Chi-Square Test

The final filter-based technique employed during this study was feature importance analysis, which assesses the contribution of each feature to the predictive model. Unlike the chi-square test and correlation matrix, this method does not assume specific data distributions, making it highly adaptable across different datasets. The feature importance rankings, depicted in Figure 5.3, largely aligned with the previous filtering algorithms, with menstrual pain being identified as the most relevant feature, followed closely by painful cramps during the period, cramping, fatigue and pelvic pain. Conversely, the least informative features included leg pain, abnormal uterine bleeding, fever and loss of appetite, all of which had 0% importance scores. These results are consistent with the findings of the correlation matrix and chi-square test, underscoring the robustness of the identified key features.

Feature importance scores may vary depending on the ML algorithm used, necessitating careful interpretation in alignment with model-specific characteristics. To further investigate the significance of model-based features, feature importance was recorded for the six ML models developed in this research. The LR model identified fatigue as the most important feature, closely followed by bowel pain, ovarian cysts and menstrual pain. It also classified digestive problems, acne and hip pain as the three least informative features for detecting endometriosis. SVM associated the diagnosis of endometriosis closely with bowel pain, irregular period and bleeding. However, features such as pain and insomnia were deemed unimportant. The RF, DT, and XGBoost models supported the general importance feature rankings by deeming

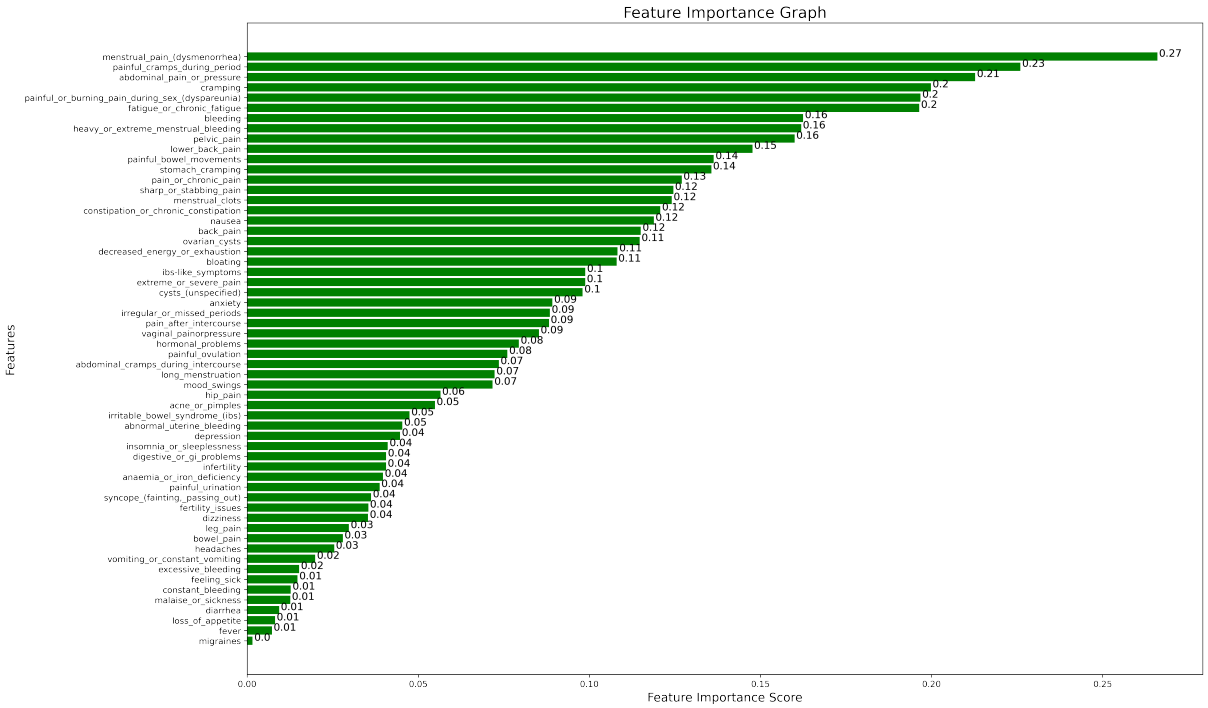


Figure 5.3 Feature Importance

menstrual pain as the 100% most influential feature within the dataset. RF highlighted fatigue as the second most informative feature with a 70% importance score, while declaring that fertility issues and digestive problems are irrelevant to the detection of the disease. Meanwhile, XGBoost ranked constant bleeding as the second most relevant feature, but with a significantly lower importance score of 21%. This model also excluded five features, including fertility issues, excessive bleeding, syncope, menstrual clots, irritable bowel syndrome, and digestive problems, to have no relation to the target variable. The DT feature importance showed a sharp decline in feature importance scores following the most relevant feature, with the fatigue symptom being ranked second at only 16% importance. Notably, this algorithm reported that 21 features, including leg pain, syncope and fertility issues, were irrelevant to the target variable, resulting in a 0% correlation score. AdaBoost categorised the features into four groups with importance scores of 100%, 67%, 33%, and 0%. It solely identified irregular periods as the most relevant feature, followed by menstrual pain, cramping and fatigue in the second percentile. Meanwhile, insomnia, digestive problems and syncope were some of the features classified as uninformative by the model, which is consistent with the feature importance results of the previous models.

The findings of these filtering approaches provide critical insights into feature relevance for diagnosing endometriosis. While some models resulted in different feature importance rankings, menstrual pain and fatigue consistently emerged as key predictive features, whereas syncope and digestive problems were frequently deemed

irrelevant.

The comprehensive list of symptom correlations and feature importance values, as calculated by the correlation matrix, chi-square test, and feature importance analysis, is presented in the Appendix C, along with the corresponding bar charts of the model-specific feature importance rankings.

5.2.2 Evaluation of Baseline Models and Feature Selection Techniques

The evaluation process of the ML models involved systematically testing multiple feature engineering techniques and hyperparameter optimisation strategies to effectively assess their impact on their predictive capabilities. Each variation of the models was rigorously evaluated using key performance metrics, including accuracy, precision, recall, F1-score, and AUC-ROC. Additionally, standard deviation values were reported to account for variability, while tools such as ROC and PR curves were utilised for further visual assessment. This analysis primarily focuses on the most relevant results derived during this research, while supplementary findings, such as the second model trained using PCA with 29 component models, have been relegated to the Appendix B due to their lower performance. Moreover, the list of features selected by the feature selection methods as well as the ROC and Precision-Recall curve plots are presented in Appendix C and Appendix D, respectively.

The base models served as the benchmark for this dissertation. These models were implemented using default parameters, as outlined in Chapter 4, and were trained iteratively on the entire symptom dataset. Without the application of the feature selection methods or hyperparameter tuning, the base models exhibited strong classification performance, demonstrating the potential for AI-driven diagnosis of endometriosis using self-reported patient symptoms. The evaluation metrics of these models are summarised in Table 5.1, with the best-performing model for each metric highlighted in bold for better readability.

Although all base models achieved high predictive accuracy, exceeding 90% in most cases, certain models performed better than others. Specifically, the LR, RF, XGBoost and AdaBoost models attained over 90% in all evaluation metrics, whereas DT and SVM exhibited slightly lower performance but remained above 85%. Notably, XGBoost and AdaBoost achieved the highest recall score of 95.69%, while LR demonstrated the lowest standard division score of 0.82, suggesting higher model stability. Among all base models, RF emerged as the top-performing model, excelling in four out of six metrics, including accuracy, precision, F1-score, and AUC-ROC, and maintaining a recall score of 94.83%, which, although not the highest, was still considered strong. LR was the second-best performing model, exhibiting consistently high performances across all metrics while maintaining low variability as demonstrated

in the standard deviation score.

Table 5.1 Performance Metrics of Base Models

Model	**Metrics**					
	Accuracy	Precision	Recall	F1 Score	AUC-ROC	Std. Dev.
Logistic Regression	94.59	95.61	93.97	94.78	94.62	**0.82**
Random Forest	**95.95**	**97.35**	94.83	**96.07**	**96.00**	2.27
XGBoost	94.59	94.07	**95.69**	94.87	94.54	1.53
Decision Tree	89.64	89.74	90.52	90.13	89.60	1.00
SVM	90.54	88.62	93.97	91.21	90.38	1.95
AdaBoost	93.24	91.74	**95.69**	93.67	93.13	1.01

To enhance model performance, various feature selection strategies were explored, including FFS, BFS, and PCA. These techniques aim to refine the feature set by retaining only the most informative symptoms while eliminating redundant or less predictive variables.

Applying FFS to the base models yielded the results presented in Table 5.2, which revealed that LR showed slight improvement in performance and consistently outperformed other models, achieving the highest accuracy, precision, F1-score, and AUC-ROC each above 95%. SVM reported the highest recall score of this variation at 96.55%, while AdaBoost exhibited the lowest standard deviation at 1.42, which was slightly higher than the baseline score, indicating performance stability might have decreased after applying FFS. Although RF maintained strong performance, a slight decline is noted in its evaluation metrics, suggesting that feature selection did not enhance its predictive capabilities, likely due to RF's inherent ability to perform feature selection internally. In contrast, XGBoost and DT exhibited noticeable performance deterioration, particularly in accuracy, with DT being the lowest-performing model, achieving assessment scores between 81% and 84%. Additionally, DT had the highest standard deviation of 2.29, further signifying model instability. Moreover, SVM and AdaBoost demonstrated balanced performance with evaluation metrics similar to their baseline variation.

Table 5.2 Performance Metrics of FFS Models

Model	**Metrics**					
	Accuracy	Precision	Recall	F1 Score	AUC-ROC	Std. Dev.
Logistic Regression	**95.05**	**95.65**	94.83	**95.24**	**95.06**	2.35
Random Forest	94.14	94.78	93.97	94.37	94.15	2.11
XGBoost	91.89	92.24	92.24	92.24	91.88	1.47
Decision Tree	81.98	82.20	83.62	82.91	81.90	2.29
SVM	92.34	89.60	**96.55**	92.95	92.14	2.20
AdaBoost	94.59	94.83	94.83	94.83	94.58	**1.42**

In terms of feature importance, FFS consistently selected 29 distinct features to

be used to train each of the ML models. Specifically, this approach selected dysmenorrhea as a key predictive factor across all models, alongside symptoms such as painful cramps, ovarian cysts, constant bleeding, digestive problems, sickness, decreased energy and fertility issues. Conversely, 12 out of the 58 symptoms were not selected by any models, including leg pain, bloating and headaches. This aligns with the results obtained through the filter methods during data analysis.

BFS was the second feature selection strategy applied to the baseline models with the aim of enhancing model performance rates. This approach resulted in exceptional performance improvement in most models, with all algorithms attaining above 90% assessments as reported in Table 5.3. Although DT showed significant improvement in key metrics compared to its base and FFS model variations, suggesting that BFS was particularly beneficial for tree-based models, this algorithm remained the lowest performing. AdaBoost exhibited a slight performance decline, indicating that BFS did not enhance its feature selection process. Notably, however, this model maintained an above 90% performance in all metrics and attained the highest recall score in this experiment of 94.83%, demonstrating its reliability and consistency. The SVM model also showed a slight decline in performance when compared to its FFS variant, suggesting that eliminating redundant features was less beneficial than forward selection but still more effective than training using the entire dataset. Despite slight assessment variations, the RF and XGBoost models retained their high performance metric scores and low standard deviation values, demonstrating their strong classification abilities, model stability and reliability. LR remained a top performer, attaining the highest accuracy, precision, F1-score and AUC-ROC score in this experimental process. Notably, however, this model exhibited a marginal decrease in its key performance metrics alongside an increase in standard deviation, suggesting increased sensitivity to the BFS method. The second and third best-performing models after applying BFS included the XGBoost and RF models, which attained strong competitive performances in all metrics with low standard deviation scores.

Table 5.3 Performance Metrics of BFS Models

Model	**Metrics**					
	Accuracy	Precision	Recall	F1 Score	AUC-ROC	Std. Dev.
Logistic Regression	**94.14**	**94.78**	93.97	**94.37**	**94.15**	2.39
Random Forest	92.79	93.86	92.24	93.04	92.82	1.46
XGBoost	92.34	92.31	93.10	92.70	92.31	1.37
Decision Tree	90.54	91.30	90.52	90.91	90.54	1.64
SVM	91.44	90.08	93.97	91.98	91.32	**1.25**
AdaBoost	92.34	90.91	**94.83**	92.83	92.23	2.31

Consistent with the FFS strategy, this approach also selected 29 features from the original 58 symptoms found in the dataset to use in model training. Aligning with

the findings of the FFS and filtering methods, BFS selected dysmenorrhea as the most informative feature in the dataset, along with ovarian cysts and chronic fatigue, including them in all six model training subdatasets. Moreover, symptoms such as sharp or stabbing pain were eliminated from the training dataset by the algorithm for all models.

PCA was the final feature engineering algorithm employed with the aim to reduce dataset dimensionality to enhance model performance and generalisability, as well as mitigate the risk of overfitting. Table 5.4 presents the performance of the ML models after applying PCA with 58 components. Notably, this technique shows a general decline in performance scores across the models when compared to the baseline models and previous feature selection results. While the DT model showed minor improvements compared to its base version, BFS yielded superior results, indicating that PCA was detrimental to its performance and reduced its predictive power. AdaBoost and SVM also exhibited deterioration in their performance when compared to previous results. Additionally, RF demonstrated a significant decrease in key performance metrics. Notably, however, this model attained the highest recall score in this variation of 96.55%. Meanwhile, XGBoost also demonstrated a decrease in its evaluation metrics. However, it maintained strong overall predictive performance and was the second-best-performing model in this experimental process. Nonetheless, LR remained the best-performing model, achieving results comparable to its base and BFS variations while maintaining the lowest standard deviation. With the highest evaluation metrics in accuracy, precision, F1-score and AUC-ROC, this algorithm demonstrated its adaptability to dimensionality reduction.

Table 5.4 Performance Metrics of PCA Models with 58 Components

Model	**Metrics**					
	Accuracy	Precision	Recall	F1 Score	AUC-ROC	Std. Dev.
Logistic Regression	**94.59**	**95.61**	93.97	**94.78**	**94.62**	0.82
Random Forest	90.99	87.50	**96.55**	91.80	90.73	1.71
XGBoost	91.89	92.24	92.24	92.24	91.88	1.55
Decision Tree	90.09	91.23	89.66	90.43	90.11	3.33
SVM	90.54	88.62	93.97	91.21	90.38	1.95
AdaBoost	90.09	90.52	90.52	90.52	90.07	1.85

5.2.3 Impact of Hyperparameter Tuning

Following the evaluation of base models and feature selection techniques, hyperparameter tuning was conducted to further refine model performance. Table 5.5 presents the performance of models after hyperparameter tuning on the baseline models. The LR model showed no significant improvements except for an increase in

standard deviation, implying a decrease in model stability. Meanwhile, RF experienced a slight decline in key metrics compared to its base model but maintained a strong, balanced performance above the 94 percentile. XGBoost exhibited minor variations in its predictive performance, and AdaBoost demonstrated marginal improvements across all metrics. The DT model showed moderate improvement but remained the lowest-performing model with persistent accuracy, precision and AUC-ROC marginally below 90%. Notably, SVM exhibited a substantial performance boost, averaging 95% across all metrics and outperforming other models in four out of six evaluation criteria, making it the top-performing model in this comparison. However, while this model attained the highest accuracy, precision, F1-score and AUC-ROC scores after hyperparameter tuning, it also reported the highest standard deviation, indicating that the model may be unstable. XGBoost followed closely, achieving the highest recall and lowest standard deviation, indicating enhanced model stability.

Table 5.5 Performance Metrics of Fine-Tuned Base Models

Model	**Metrics**					
	Accuracy	Precision	Recall	F1 Score	AUC-ROC	Std. Dev.
Logistic Regression	94.59	95.61	93.97	94.78	94.62	1.96
Random Forest	94.59	94.83	94.83	94.83	94.58	2.36
XGBoost	94.59	93.33	**96.55**	94.92	94.50	**1.03**
Decision Tree	89.64	89.08	91.38	90.21	89.56	2.04
SVM	**95.05**	**95.65**	94.83	**95.24**	**95.06**	2.67
AdaBoost	94.14	93.28	95.69	94.47	94.07	1.34

Table 5.6 presents the evaluation of the FFS-based models after applying hyperparameter tuning. This process led to the deterioration of most of the model's predictive abilities. AdaBoost and SVM showed identical results before and after tuning, indicating that these were already well optimised. Despite an increase in its performance after tuning, DT remained the weakest performer across the modelling algorithms, with metrics below the 85 percentile and the highest standard deviation. LR and XGBoost exhibited lower performances compared to their FFS counterparts, suggesting that hyperparameter tuning was not always beneficial when combined with FFS, as observed in the performance decline. However, their respective standard deviation experienced slight improvement, indicating better model stability. Meanwhile, RF exhibited slight improvements in performance, attaining the best performance scores in three out of six assessment criteria, including accuracy, precision and AUC-ROC. Notably, AdaBoost also achieved the highest metrics in three of the six evaluation criteria, including accuracy, F1-score and standard deviation. Thereby, RF and AdaBoost were considered as the top-performing models of this experiment.

Similarly, the application of hyperparameter tuning on the BFS model variants resulted in slight performance drops, but to a lesser extent than that experienced by

Table 5.6 Performance Metrics of Fine-Tuned FFS Models

Model	**Metrics**					
	Accuracy	Precision	Recall	F1 Score	AUC-ROC	Std. Dev.
Logistic Regression	94.14	94.02	94.83	94.42	94.11	2.37
Random Forest	**94.59**	**95.61**	93.97	94.78	**94.62**	1.59
XGBoost	91.44	92.17	91.38	91.77	91.44	1.62
Decision Tree	83.33	84.96	82.76	83.84	83.36	3.32
SVM	92.34	89.60	**96.55**	92.95	92.14	2.20
AdaBoost	**94.59**	94.83	94.83	**94.83**	94.58	**1.42**

the FFS models. The results of this process is listed in Table 5.7. SVM and AdaBoost maintained their pre-tuning performance of above 90%, highlighting their consistency and robustness. The DT model exhibited minor improvements in performance when compared to its base and FFS models, suggesting that BFS was suited for this type of ML algorithm. Meanwhile, LR experienced slight declines in its predictive performance but retained strong scores above 93%, demonstrating the models' resilience. Additionally, the RF model showed a slight decrease in all key metrics with the exception of the precision score, indicating that BFS may not be the most suitable feature selection method for this algorithm. XGBoost, however, exhibited enhanced performance, excelling in four out of six of the evaluation criteria, including accuracy, precision, F1-score and AUC-ROC, thereby making it the top-performing modelling algorithm of this variation.

Table 5.7 Performance Metrics of Fine-Tuned BFS Models

Model	**Metrics**					
	Accuracy	Precision	Recall	F1 Score	AUC-ROC	Std. Dev.
Logistic Regression	93.69	93.97	93.97	93.97	93.68	2.45
Random Forest	92.34	94.59	90.52	92.51	92.43	3.23
XGBoost	**94.59**	**95.61**	93.97	**94.78**	**94.62**	1.54
Decision Tree	88.74	88.24	90.52	89.36	88.65	1.70
SVM	91.44	90.08	93.97	91.98	91.32	**1.25**
AdaBoost	92.34	90.91	**94.83**	92.83	92.23	2.31

Finally, hyperparameter tuning on the PCA algorithm had a notable impact on model performance, albeit with varying degrees of effectiveness across different classifiers as presented in Table 5.8. The DT model demonstrated improved predictive performance compared to its base model and feature selection-based counterparts, suggesting that feature redundancy reduction and hyperparameter optimisation enhanced its classification ability. However, despite this relative improvement, DT remained the lowest-performing algorithm among all models implemented in this study, with most evaluation metrics unable to surpass the 90% threshold. LR exhibited a decline in overall performance, except for a minor improvement in its standard

deviation score, indicating increased stability. Conversely, RF demonstrated minimal improvement in predictive performance but experienced a higher standard deviation, suggesting reduced model stability compared to its base variant. Both XGBoost and AdaBoost maintained competitive performance levels but exhibited slight variations in their evaluation metrics, coupled with an increase in standard deviation, indicating reduced robustness following hyperparameter tuning. Notably, SVM emerged as the top-performing model in this comparison with an average evaluation score of 95% across all metrics. However, despite achieving the highest classification scores, it retained the same predictive performance as its base counterpart while exhibiting an increase in standard deviation.

Table 5.8 Performance Metrics of Fine-Tuned PCA Models with 58 Components

Model	**Metrics**					
	Accuracy	Precision	Recall	F1 Score	AUC-ROC	Std. Dev.
Logistic Regression	90.54	88.62	93.97	91.21	90.38	**1.13**
Random Forest	91.89	89.52	**95.69**	92.50	91.71	2.51
XGBoost	92.34	91.60	93.97	92.77	92.27	2.39
Decision Tree	89.64	92.66	87.07	89.78	89.76	1.76
SVM	**95.05**	**95.65**	94.83	**95.24**	**95.06**	2.36
AdaBoost	92.79	94.64	91.38	92.98	92.86	1.47

5.2.4 Comparative Analysis of Top-Performing Models

Table 5.9 presents the highest-performing models across all experimental variations conducted in this research. As observed in the previous subsection of this document, LR accounted for three of the top-performing models across eight experimental configurations, highlighting its consistent predictive capability. RF and SVM each produced two top-performing models for the highest-ranking set, while XGBoost contributed one.

Among these models, the RF base variant before hyperparameter tuning demonstrated superior classification performance, achieving top scores in five out of six key evaluation metrics, including accuracy, precision, recall, F1-score and AUC-ROC. Despite LR's dominance across multiple experiments, its FFS variant demonstrated the best recall score of 94.83% along with three other models, while its PCA variation reported the lowest standard deviation across all models. Further analysis revealed the fine-tuned SVM baseline model and its PCA variant produced competitive results, achieving classification scores exceeding the 95 percentile, making them comparable to the LR FFS model. Notably, however, these models reported an increased standard deviation score, suggesting that while SVM remained a high-performing model, its predictions were less stable relative to RF and LR in specific

configurations. The tuned XGBoost BFS model also exhibited excellent performance metrics across all evaluation criteria, with the second-best standard deviation score, providing a balance between the model's stability and predictive performance. Overall, these findings highlight RF and LR as the most effective classifiers for endometriosis detection, with varying strengths in different aspects of predictive performance.

Table 5.9 Performance Metrics of Top ML Models

Model	**Metrics**					
	Accuracy	Precision	Recall	F1 Score	AUC-ROC	Std. Dev.
Random Forest (Base)	**95.95**	**97.34**	**94.83**	**96.07**	**96.00**	2.27
Logistic Regression (FFS)	95.05	95.65	**94.83**	95.24	95.06	2.35
Logistic Regression (BFS)	94.14	94.78	93.97	94.37	94.15	2.39
Logistic Regression (PCA)	94.59	95.61	93.97	94.78	94.62	**0.82**
SVM Hyp	95.05	95.65	**94.83**	95.24	95.06	2.67
Random Forest (FFS Hyp)	94.59	95.61	93.97	94.78	94.62	1.59
XGBoost (BFS Hyp)	94.59	95.61	93.97	94.78	94.62	1.54
SVM (PCA Hyp)	95.05	95.65	**94.83**	95.24	95.06	2.36

5.2.5 Comparative Analysis with Literature Review Models

Bendifallah et al. [13] reported the performance metrics of various models, including sensitivity, specificity, F1-score, and AUC-ROC. Consistent with the findings of this study, their results indicated that the DT model exhibited the lowest performance across all metrics, whereas RF, LR, and XGBoost demonstrated superior predictive capabilities. Notably, the LR model achieved the highest recall of 95% across all modelling algorithms, while LR and XGBoost attained the highest F1-score of 92%. As reported in Table 5.9, the LR model in this study also achieved the 95% recall successfully, and all top-performing models surpassed the 92% F1-score threshold. Additionally, Bendifallah et al. [13] reported an AUC-ROC score of 93% for XGBoost, a benchmark that was surpassed by the top models in this study. Notably, it is acknowledged that the dataset used in this research differs from that of Bendifallah et al. [13]. Consequently, while the comparative analysis highlights similar trends, direct comparisons should be interpreted with caution. Nevertheless, both studies reinforce the potential effectiveness of LR, RF, and XGBoost in detecting endometriosis based on self-reported symptom data.

Goldstein and Cohen [42] conducted an evaluation using the same dataset as this study, assessing the performance of DT, RF, GBC, and AdaBoost in predicting endometriosis. Their study employed a similar set of evaluation metrics, including accuracy, recall, specificity, precision, F1-score, and AUC-ROC. As presented in Table 5.10, the models implemented in this study demonstrated competitive or

superior performance compared to those reported by Goldstein and Cohen [42]. A key distinction between the two studies lies in the methodological enhancements applied in this research. Specifically, the use of feature engineering and hyperparameter tuning contributed to improved model performance. Notably, the baseline models developed in this study outperformed most of the results from this research paper, while the final optimised models achieved even higher performance metrics. These findings underscore the effectiveness of feature selection and hyperparameter optimisation in enhancing predictive performance for endometriosis detection. Furthermore, this study identified 24 features with high predictive value for diagnosing endometriosis, aligning with and reinforcing the findings made in this research. This consistency underscores the robustness of the selected features and their relevance in improving model reliability for clinical applications.

Table 5.10 Comparison of Performance Metrics Against [42]

Metrics	Base Models				Model Results from [42]			
	RF	XGB	DT	AdaBoost	RF	XGB	DT	AdaBoost
Accuracy	95.05%	94.59%	89.64%	93.24%	93.00%	92.80%	87.60%	93.70%
Precision	97.30%	94.07%	89.74%	91.74%	94.50%	94.20%	88.00%	94.40%
Recall	93.10%	95.69%	90.52%	95.69%	92.40%	92.40%	89.00%	93.90%
F1 Score	95.15%	94.87%	90.13%	93.67%	93.40%	93.20%	88.50%	94.10%
AUC	95.14%	94.54%	89.60%	93.13%	93.00%	92.80%	87.50%	93.70%

5.2.6 Final Remarks

One of the key insights derived from this experimental phase was that hyperparameter tuning did not universally enhance model performance. Additionally, while tuning improved the baseline models in most cases, its application to the feature engineering algorithms, yielded inconsistent results. Certain models, such as LR and SVM, benefitted from optimisation, whereas others, such as RF, experienced a decline in performance. Feature selection techniques prior to hyperparameter tuning exhibited mixed effects. Although the SVM model benefitted from PCA transformation, other models showed limited improvement or significant performance deterioration. These findings underscore the importance of model-specific tuning strategies and the need for tailored optimisation approaches when applying AI-driven techniques for medical diagnostics.

DT consistently struggled across all experimental variations, reinforcing its unsuitability for endometriosis detection compared to more advanced ML models. RF and SVM exhibited high variability, demonstrating strong performance across multiple configurations but underperforming in select variations. Notably, AdaBoost emerged as the most resilient and stable model, consistently maintaining above 90%

classification accuracy across all experimental conditions. XGBoost, despite facing challenges with its FFS-based feature selection variation, remained one of the highest-performing models, further emphasising its robustness in structured medical datasets. Overall, LR emerged as the most consistently high-performing model, achieving top-ranking classification scores across various configurations.

Feature selection played a crucial role in optimising the ML models by identifying the most informative predictors of endometriosis and eliminating irrelevant or redundant features from the dataset. By analysing the results of all the feature selection techniques employed in the study, including the correlation matrix, chi-square test, feature importance, FFS and BFS, the features that provide the most accurate predictive results for diagnosing endometriosis were determined. The key predictors of endometriosis, selected by all methods, included menstrual pain (dysmenorrhea), painful cramps, and fatigue. Fertility issues, ovarian cysts, decreased energy levels, bowel pain and digestive problems were some of the most selected symptoms by these techniques. Meanwhile, features such as acne, fever, syncope and insomnia were rarely selected by these algorithms, signifying their unimportance to diagnosing the disease.

5.3 Deep Learning Model Assessment

The evaluation of various DL architectures was conducted across multiple experimental setups to gain crucial insights into their effectiveness and efficiency. Specifically, the assessment followed a structured methodology that included evaluating baseline models, the application of hyperparameter tuning, the implementation of data augmentation, and a combination of both strategies. Each DL architecture was assessed using the same key metrics as the ML models to ensure consistency. Additionally, the training duration of each model was also recorded. Moreover, the accuracy and loss curves were plotted to visualise the learning trends and are illustrated in Appendix E.

Although the GLENDA dataset includes four annotated categories of endometriosis, including peritoneal, ovarian, uterine and DIE, this study restricted the DL evaluation to a binary classification task distinguishing pathological from non-pathological laparoscopic images. This design choice was made for several methodological and data-driven reasons. First, the distribution of cases across the four subtypes was highly imbalanced, with substantially fewer annotated samples for certain classes, particularly uterine and DIE lesions. Conducting a reliable multiclass experiment under such imbalance would risk severe overfitting and unreliable performance estimates. Second, the primary aim of this dissertation was to develop an AI-driven tool for early detection and initial diagnostic support by detecting

endometrial lesions in laparoscopic images rather than subtype characterisation. Establishing a robust binary diagnostic baseline is a necessary precursor to subsequent research that may expand toward fine-grained classification once sufficient data is available. Finally, due to the constraint where frames containing multiple types of pathology were assigned a single label based on the lesion with the largest coverage area, the reliability of the sub-class labels is inherently reduced. Hence, for the purpose of this dissertation, the evaluation process consists solely of binary classification.

5.3.1 Evaluation of Base Architectures

The results of the base model evaluation are summarised in Table 5.11, with the optimal results marked in bold to enhance readability. Although most of the DL architectures achieved high performance rates using the original medical image dataset before data augmentation, some models struggled with the classification task proposed in this dissertation.

The hybrid InceptionResNetV2 architecture exhibited the lowest performance across all evaluation metrics, achieving an accuracy of only 52%, an AUC-ROC score of 50%, and failing to register any meaningful precision, recall, or F1-score. This underperformance may be attributed to numerous reasons, including overfitting or the model's inability to extract relevant features. This is followed by the ResNetV2 and InceptionV3 frameworks, which achieved relatively high performance but remained slightly below the top-tier models, with certain assessment criteria below the 90% threshold. Meanwhile, the NasNetMobile and Xception architectures demonstrated strong classification capabilities with key evaluation metrics above 90% and training durations of less than two hours.

Among the most efficient architectures, the lightweight CNN MobileNetV3 Small and Large variants demonstrated excellent model performance, consistently achieving over 99% scores across all evaluation metrics while maintaining the shortest training durations of 16 and 27 minutes, respectively. These characteristics suggest their suitability for real-time applications where computational efficiency is a priority. ResNet50 and EfficientNetV2B0 emerged as the top-performing architectures in this comparison, achieving near-perfect classification results exceeding 99% while maintaining reasonable training durations that balance their high performance rates. In particular, ResNet50 attained the highest evaluation metrics in four out of six assessment criteria, including accuracy, recall, F1-score and AUC-ROC, making it the best-performing model in this configuration. Additionally, this model achieved a precision of 99.96%, second only to EfficientNetV2B0's perfect 100%, and a training duration of less than two hours. Notably, the remaining of EfficientNetV2B0's performance metrics were slightly below that of ResNet50, thereby positioning it as the

second-best model overall. Moreover, while VGG16 demonstrated strong performance in this classification task, with all scores exceeding 99.6%, its extended training time of 3 hours and 21 minutes rendered it less suitable for practical applications compared to EfficientNetV2B0 and ResNet50 models, which delivered similar or better performances in efficient training durations and reduced computation cost.

Table 5.11 Performance Metrics of Base Architectures

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC	Training Duration
MobileNetV3 Small	99.84%	99.96%	99.72%	99.84%	99.84%	**0h 16m 14.98s**
MobileNetV3 Large	99.77%	99.80%	99.72%	99.76%	99.76%	0h 27m 6.40s
NASNetMobile	98.11%	98.33%	97.74%	98.04%	98.10%	1h 44m 16.69s
EfficientNetV2B0	99.90%	**100.00%**	99.80%	99.90%	99.90%	1h 17m 48.77s
DenseNet121	99.18%	99.03%	99.27%	99.15%	99.19%	2h 39m 32.41s
ResNet50	**99.92%**	99.96%	**99.88%**	**99.92%**	**99.92%**	1h 35m 19.53s
ResNet50V2	89.37%	91.03%	86.47%	88.69%	89.27%	1h 8m 59.23s
InceptionV3	91.75%	97.37%	85.18%	90.87%	91.52%	1h 13m 6.71s
Xception	94.94%	98.22%	91.16%	94.55%	94.81%	1h 17m 12.70s
InceptionResNetV2	51.78%	0.00%	0.00%	0.00%	49.98%	1h 59m 51.11s
VGG16	99.65%	99.68%	99.60%	99.64%	99.65%	3h 21m 0.27s

5.3.2 Impact of Hyperparameter Tuning

To further enhance model performance, hyperparameter tuning was applied following the transfer learning process. Notably, due to computational and time constraints, the number of different hyperparameter combinations that could be explored by each algorithm was limited to five. Additionally, an early stopping mechanism was implemented to monitor the validation loss with a patience parameter of three that restores the best-performing weights once the tuning process is complete. Afterwards, the final models were trained using the optimised parameters determined by the tuning algorithm for 20 epochs with a 30% validation split. Notably, the training duration of these architectures was not recorded as the training process was less computationally expensive.

As presented in Table 5.12, although the baseline architectures demonstrated strong evaluation metrics, the application of hyperparameter tuning generally led to improvements in performance across most of the models. However, certain models, such as DenseNet and ResNet50, exhibited slight declines in performance. While these models maintained evaluation metrics above the 98% threshold, the deterioration in performance suggests that finetuning could lead to suboptimal results for some architectures. Meanwhile, ResNet5V2 experienced a significant drop in accuracy and recall, highlighting potential model instability introduced through hyperparameter

adjustments. Moreover, despite hyperparameter tuning, the InceptionResNetV2 architecture remained the poorest-performing architecture in this configuration. Although slight improvements were observed in its accuracy and AUC-ROC metrics, the model was still unable to consistently correctly classify endometriosis through laparoscopic images, reaffirming its unsuitability for this specific task. Conversely, VGG16, Xception and InceptionV3 benefitted from hyperparameter tuning, demonstrating enhanced predictive capabilities.

Among the best-performing modelling algorithms of this configuration, the MobileNet frameworks further enhanced their already high classification performance, achieving near-perfect evaluation scores of above 99% across all metrics. These models attained the highest precision rate of 99.96%, along with the EfficientNetV2B0 architecture. Moreover, after hyperparameter tuning, EfficientNetV2B0 emerged as the top-performing model, resulting in the highest evaluation metrics in all of the recorded assessment criteria. With near-perfect scores of 99.96% in precision, 100% in recall and 99.98% in accuracy, F1-score and AUC-ROC, this model indicated its superior ability to identify all positive cases with 100% certainty. Therefore, the top-performing models following hyperparameter tuning were EfficientNetV2B0, MobileNetV3 Large and ResNet50, respectively.

Table 5.12 Performance Metrics of Fine-Tuned DL Architectures

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
MobileNetV3 Small	99.86%	**99.96%**	99.75%	99.86%	99.86%
MobileNetV3 Large	99.96%	**99.96%**	99.96%	99.96%	99.96%
NASNetMobile	97.66%	97.58%	97.50%	97.54%	97.66%
EfficientNetV2B0	**99.98%**	**99.96%**	**100.00%**	**99.98%**	**99.98%**
DenseNet121	98.70%	98.41%	98.85%	98.63%	98.70%
ResNet50	99.81%	99.80%	99.80%	99.80%	99.80%
ResNet50V2	77.48%	98.86%	53.20%	69.17%	76.32%
InceptionV3	94.24%	90.92%	97.62%	94.15%	94.40%
Xception	97.86%	98.62%	96.84%	97.73%	97.81%
InceptionResNetV2	52.50%	0.00%	0.00%	0.00%	50.00%
VGG16	99.67%	99.79%	99.51%	99.65%	99.66%

5.3.3 Effects of Data Augmentation

Data augmentation was introduced to improve model generalisability by increasing dataset variability. However, due to the already large size of the image dataset, computational constraints limited the extent of augmentation. Initially, three augmented images were intended to be created per original image, but this resulted in memory overload errors. Despite optimisation attempts, only one augmented image

per original was feasible for training. Additionally, some models encountered memory-based or time-out errors during the training process, reducing the number of architectures included in this experiment to eight. Namely, the models that were trained and assessed for this experiment included MobileNetV3 Small and Large, EfficientNetV2B0, DenseNet121, ResNet50V2, InceptionV3, Xception and the hybrid architecture InceptionResNetV2. The VGG16, ResNet50 and NasNetMobile architectures were excluded due to computational limitations.

As illustrated in Table 5.13, data augmentation yielded mixed results. The InceptionResNetV2 model continued to perform poorly, reaffirming its unsuitability for this classification problem. Consistent with its previous performances, the InceptionResNetV2 architecture continued to perform poorly, resulting in an accuracy of 53%, AUC-ROC of 50%, and precision, recall and F1-score of 0%, further reaffirming its unsuitability for this classification problem. ResNet50V2 exhibited a significant decline in performance, with its accuracy dropping from 89% to 67%, suggesting that this model did not generalise well under data augmentation. Inception and Xception achieved good results, with evaluation metric scores ranging between 78% and 93%. However, their dramatically increased training durations of 2.5 and nearly 6 hours, respectively, rendered them inefficient.

DenseNet121 maintained high performance, exceeding 95% across all evaluation criteria, but its training duration of nearly 6 hours and slight performance decline made it less practical compared to its previous variant. Conversely, MobileNetV3 Small and Large, along with EfficientNetV2B0, preserved their high classification accuracy, exceeding the 99% performance rates. With the best results in five out of six evaluation criteria and a consistent near-perfect 99.8% score in all metrics, MobileNetv3 Large was notably the top-performing model in this experiment, followed closely by MobileNetV3 Small and EfficientNetV2B0. However, the trade-off in training time must be considered, particularly for MobileNetV3 Small, where training duration increased from 16 minutes to nearly 2 hours, raising questions about whether the slight performance gain justified the increased computational cost.

Table 5.13 Performance Metrics of Data Augmented DL Architectures

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC	Training Duration
MobileNetV3 Small	99.6%	99.7%	99.4%	99.6%	99.6%	1h 54m 35.74s
MobileNetV3 Large	**99.8%**	**99.8%**	**99.8%**	**99.8%**	**99.8%**	1h 49m 30.03s
EfficientNetV2B0	99.5%	99.3%	99.6%	99.5%	99.5%	1h 44m 11.72s
DenseNet121	95.7%	95.7%	95.2%	95.4%	95.7%	5h 8m 3.79s
ResNet50V2	67.0%	76.3%	43.4%	55.3%	65.7%	**1h 14m 49.16s**
InceptionV3	82.9%	84.4%	78.3%	81.2%	82.7%	2h 30m 42.44s
Xception	89.8%	92.7%	85.0%	88.7%	89.5%	5h 48m 56.86s
InceptionResNetV2	52.9%	0.00%	0.00%	0.00%	50.00%	2h 50m 25.18s

Following data augmentation, hyperparameter tuning was applied, resulting in the findings presented in Table 5.14. As expected based on previous performances, the InceptionResNetV2 model continued to exhibit poor classification performance. While ResNet50V2, InceptionV3, and Xception showed slight improvements, their performance remained below the 90% threshold, making them less competitive compared to their base models and the remaining architectures of this experiment. DenseNet121 exhibited mixed results, enhancing its precision but declining in other metrics. Notably, however, this model still produced performance scores in the 90 percentile. Although the MobileNetV3 and EfficientNetV2B0 frameworks remained the top three models in this experiment, with evaluation metrics consistently exceeding 99%, these models are still slightly inferior to their non-augmented model variants. However, with near-perfect evaluation scores ranging between 99.7% and 99.9%, MobileNetV3 Large and EfficientNetV2B0 both attained 4 out of 5 of the best performance scores across the board. Therefore, MobileNetV3 Large was the top-performing model after fine-tuning, closely followed by EfficientNetV2B0.

Table 5.14 Performance Metrics of Fine-Tuned Data Augmented DL Architectures

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
MobileNetV3 Small	99.1%	99.0%	99.2%	99.1%	99.1%
MobileNetV3 Large	**99.8%**	**99.9%**	99.7%	**99.8%**	**99.8%**
EfficientNetV2B0	**99.8%**	99.7%	**99.8%**	**99.8%**	**99.8%**
DenseNet121	95.2%	96.7%	93.3%	95.0%	95.2%
ResNet50V2	85.4%	83.5%	86.9%	85.2%	85.5%
InceptionV3	83.2%	86.4%	77.4%	81.6%	83.0%
Xception	88.4%	95.3%	79.9%	86.9%	88.1%
InceptionResNetV2	51.8%	0.00%	0.00%	0.00%	50.0%

5.3.4 Comparative Analysis of Top-Performing Models

Table 5.15 presents the highest-performing models across all experimental configurations conducted on the DL architectures. Among these, the MobileNetV3 Lare architecture accounted for two of the four top-performing models, demonstrating its robustness across multiple experimental setups. Meanwhile, EfficientNetV2B0 and ResNet50 each contributed one top-performing model within this highest-ranking set. Notably, EfficientNetV2B0 consistently delivered excellent results across all configurations, closely following MobileNetV3 Large in the data augmentation experiments and the ResNet50 in the baseline configuration.

Although all of the models listed in Table 5.15 attained evaluation metrics exceeding the 99% threshold, the EfficientNetV2B0 architecture after fine-tuning emerged as the best-performing model, excelling across all five assessment criteria. This architecture achieved a recall of 100%. Precision of 99.96% and accuracy, F1-score and AUC-ROC of 99.98%, making it the most reliable choice for endometriosis classification. This is followed by the baseline ResNet50 architecture, which produced high performance rates above 99.92% across all metrics. While the augmented MobileNetV3 Large model was the third-best performing architecture in this comparison, it should be noted that this architecture's baseline variation attained an accuracy of 99.96% after hyperparameter tuning. This suggests that augmentation-based models may not be effective in this classification problem.

Table 5.15 Performance Metrics of Top-Performing DL Architectures

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
ResNet50	99.92%	99.96%	99.88%	99.92%	99.92%
EfficientNetV2B0 Tuned	**99.98%**	**99.96%**	**100.00%**	**99.98%**	**99.98%**
MobileNetV3 Large Augmented	99.8%	99.8%	99.8%	99.8%	99.8%
MobileNetV3 Large Augmented & Tuned	99.8%	99.9%	99.7%	99.8%	99.8%

5.3.5 Comparison with Literature Review Models

Visalaxi and Muthu [43] also utilised the GLENDA dataset to detect endometriosis using transfer learning with several CNN architectures, including VGG16, ResNet50, InceptionV3, Xception, and InceptionResNetV2. When compared to the models implemented in this study, nearly all base models outperformed those reported by Visalaxi and Muthu [43], with the exception of InceptionResNetV2. This can be observed in Table 5.16, where the baseline results of this study are presented against the findings of the research in [43]. The superior performance of this architecture in their study, despite poor results in this dissertation, remains unclear, as both used the

same dataset and transfer learning techniques. However, this occurrence may be due to differences in the employed data preprocessing steps, training strategies and hyperparameter tuning.

The research in [43] identified ResNet50 as the top-performing architecture, reporting an accuracy of 90%, precision of 83%, recall of 82%, F1-score of 82%, and an AUC of 78%. Comparatively, the ResNet50 base model in this study achieved significantly higher performance scores, both with and without hyperparameter tuning. Notably, due to computational constraints, a comparison with the data augmentation variant of ResNet50 could not be made. However, the findings of that experiment show that some models managed to surpass these results, including MobileNetV3 and EfficientNetV2B0.

Table 5.16 Comparison of Performance Metrics Against [43]

Architecture	Base Models		Model Results from [43]	
	Accuracy	Recall (Sensitivity)	Accuracy	Recall (Sensitivity)
ResNet50	99.92%	99.88%	91%	82%
InceptionV3	91.75%	85.18%	84%	80%
Xception	94.94%	91.16%	83.5%	78%
InceptionResNetV2	51.78%	0.00%	88%	75%
VGG16	99.65%	99.60%	80%	76%

5.3.6 Final Remarks

A key insight from this study is that while hyperparameter tuning is conventionally applied to optimise models and enhance predictive performance, similarly to the results observed for the ML evaluation, its impact varied across different architectures. Notably, ResNet50V2 and InceptionV3 experienced performance deterioration rather than improvement, suggesting that certain architectures may be inherently less adaptable to hyperparameter modifications in this classification task. Furthermore, data augmentation produced mixed results. While it improved the robustness of some architectures, it led to decreased accuracy in others. The combined application of data augmentation and hyperparameter tuning yielded the most consistent improvements, though these enhancements came at a higher computational and memory cost. These trade-offs must be carefully considered to ensure an optimal balance between model performance and training efficiency.

Among these models, MobileNetV3 and EfficientNetV2B0 frameworks consistently demonstrated superior performance in all model variants, striking an optimal balance between predictive accuracy and computational efficiency. In contrast, models such as InceptionResNetV2 and ResNet50V2 struggled across similar configurations, indicating that their architectural complexities may not have been

well-suited to the dataset or the specific classification task. For real-world applications that demand both high accuracy and computational efficiency, the fine-tuned EfficientNetV2B0 or MobileNetV3 Large architectures are the most suitable choices for endometriosis detection. However, in resource-constrained environments, MobileNetV3 Small provides a practical trade-off between classification accuracy and training time, making it a viable alternative.

5.4 Conclusion

This chapter provided a comprehensive evaluation of the ML and DL models used for the classification of endometriosis using the self-reported patient symptoms and GLENDA laparoscopic image dataset. The evaluation followed a systematic approach, beginning with the evaluation plan and progressing through a detailed assessment of various feature engineering techniques, hyperparameter tuning strategies, and model performance comparisons against literature-based benchmarks. The findings from this evaluation have yielded valuable insights into the impact of feature selection, model tuning, and data augmentation on predictive performance.

In the ML evaluations, SVM and LR consistently outperformed other models, while DT exhibited the weakest performance. Feature selection and hyperparameter tuning had varying effects, with some models benefiting while others, particularly PCA-based models, showed minimal improvement. For the DL assessment, EfficientNetV2B0 and MobileNetV3 Large emerged as the top-performing architectures. In addition, while data augmentation improved model robustness in some cases, it did not always justify the increased computational cost. These results contribute to the growing body of research on AI-driven diagnostic tools for endometriosis, demonstrating the feasibility of applying transfer learning and feature engineering techniques to improve disease detection.

6 Conclusion

This dissertation has explored the potential of AI-powered tools for the early detection and diagnosis of endometriosis, leveraging both self-reported patient symptoms and laparoscopic medical imagery data. The primary objective was to develop non-invasive, robust ML and DL modelling algorithms capable of accurately diagnosing the disease at an earlier stage, thereby reducing diagnostic delays and improving patient care. Furthermore, this research conducted a comprehensive comparative analysis to assess the performance of the developed models in order to determine AI-driven approaches most suitable for clinical applications and to support healthcare professionals to effectively and efficiently detect endometriosis.

This concluding chapter revisits and evaluates the aims and objectives established in Chapter 1, assessing the extent to which each objective was successfully achieved. In addition, the primary research question—How can AI techniques be employed to effectively and efficiently detect and diagnose endometriosis based on clinical and imagery data at early stages?—is addressed based on the findings of this research. The limitations encountered throughout this study, including medical data acquisition challenges and computational constraints, are acknowledged and discussed. Furthermore, several prospective future research directions for extending this work are outlined. Finally, this dissertation concludes with a succinct summary and final reflection.

6.1 Revisiting the Aim and Objectives

The principal aim of this dissertation was to investigate, develop, and evaluate AI-driven approaches for the early detection of endometriosis using clinical and imaging data to enhance detection accuracy, efficiency and overall patient outcomes. This was systematically addressed through the four objectives presented below.

Objective 1: Research and investigate various AI techniques effective in disease diagnostics

The first objective was attained through the extensive literature review presented in Chapter 3, where a broad spectrum of AI-powered methodologies in the field of medical disease diagnostics were examined. This review emphasised the applicability of ML and DL models in disease prediction and classification, with a particular focus on their effectiveness in diagnosing endometriosis. Furthermore, this research identified the current state-of-the-art AI technologies with high potential for endometriosis detection, including various ML and DL frameworks. Key insights were

gained into the types of data utilised in these predictive models, including clinical variables, self-reported patient symptoms, medical images and genetic markers. Moreover, various model evaluation techniques, such as classification metrics and performance assessment methodologies, were explored. The findings from this review provided a structured foundation for the dataset acquisition, model implementation, and comparative evaluation undertaken in this study.

Objective 2: Attain and preprocess the clinical and imaging datasets

The acquisition of the medical datasets posed a significant challenge due to privacy and ethical concerns, as well as the limited availability of publicly accessible data. However, after an exhaustive search, both clinical and image datasets were successfully obtained. The clinical dataset consisted of self-reported patient symptom data sourced from the study by Goldstein and Cohen [42] and required minimal preprocessing. Nevertheless, extensive feature engineering techniques were applied to assess the predictive performance, including FFS, BFS and PCA. Additionally, filter-based feature selection methods such as feature importance analysis, chi-square testing, and correlation matrices were employed to assess the predictive value of various features. Meanwhile, the laparoscopic imagery dataset GLENDA was retrieved from the ITEC Datasets repository [44]. This dataset underwent preprocessing procedures such as resizing and normalisation to align with model input requirements. Moreover, data augmentation techniques were applied as an experimental measure to evaluate and compare their impact on model performance.

Objective 3: Implement several ML and DL endometriosis diagnostic models

This objective was accomplished through the implementation of a diverse set of ML and DL models tailored for endometriosis detection. Six distinct ML classifiers were developed for detecting endometriosis using self-reported symptoms, including LR, RF, DT, SVM, XGBoost, and AdaBoost. Additionally, variations incorporating feature selection techniques, such as FFS, BFS, and PCA, were explored to optimise predictive performance. For DL-based classification, eleven pre-trained CNN architectures were employed, including VGG16, ResNet50, ResNet50V2, DenseNet121, InceptionV3, Xception, InceptionResNetV2, MobileNetV3 Small and Large, NASNetMobile, and EfficientNetV2B0. These models were fine-tuned using transfer learning techniques to detect endometrial lesions in laparoscopic images. Additionally, hyperparameter tuning was conducted to optimise model performance while ensuring computational efficiency.

Objective 4: Evaluate the effectiveness and efficiency of the developed models and perform a comparative analysis The final objective was fulfilled through the rigorous model performance evaluation and comparative analysis detailed in Chapter 5. Several classification metrics were employed to assess model effectiveness, including accuracy, precision and recall. Additionally, visualisation tools such as ROC curves, PR curves, accuracy and loss graphs were utilised to further validate the findings. The comparative analysis identified the most effective models for clinical application, confirming that AI-driven approaches can significantly enhance the early detection and diagnosis of endometriosis. Additionally, a comparative assessment against previously published models reaffirmed the competitiveness and reliability of the developed methodologies.

By attaining these objectives, this dissertation was able to successfully address the primary research question, demonstrating that AI techniques can significantly enhance the early detection and diagnostics of endometriosis through symptom-based analysis and medical image interpretation. Furthermore, this study also identified high-performing ML and DL models that not only attain high accuracy but also maintain computational efficiency, making them suitable for real-world clinical deployment.

6.2 Limitations

Despite the promising results attained by the ML and DL models, several limitations were encountered throughout this research. One major challenge was the limited availability of public medical datasets, as many existing studies utilise private, non-publicly accessible data. This constraint restricted the diversity of training samples, potentially impacted model generalisation, and led to a prolonged search to obtain the necessary datasets for this study.

Computational limitations also posed significant challenges during this dissertation, particularly when developing the DL algorithms. Due to hardware restrictions and time constraints, training CNN models from scratch was deemed infeasible. Therefore, this study relied on pre-trained models with transfer learning to refine the architectures to detect endometrial lesions in laparoscopic images. Furthermore, resource limitations restricted the extent of hyperparameter tuning on the DL models, limiting the number of allowed tuning tests for each architecture. Moreover, this limitation constrained the experiments with data augmentation, as certain models became untrainable due to the significant increase in data, resulting in memory-based or time-out errors.

6.3 Future Work

Given the findings and limitations of this dissertation, several avenues for future research can be proposed. One such future work for the ML modelling algorithms would be the deployment of this system into clinical applications to aid healthcare workers in real-time diagnostic assistance. Additionally, another prospect would be to develop and publish a mobile application available to the public where users input self-reported symptoms to receive a likelihood score for endometriosis. This application could also be integrated with female health tracking applications, thus enhancing the diseases' awareness.

With respect to the DL modelling algorithms, lesion localisation could be added to the models to extend the DL classifiers to not only classify the pathology but also localise the endometrial lesions using bounding boxes, thereby enhancing interpretability and clinical applicability. Moreover, these architectures could be further enhanced to differentiate between endometriosis subtypes, including superficial, deep infiltrating and ovarian, which could provide more granular diagnostic insights.

The integration of multimodal AI systems is another prospective future work that could be studied. By combining clinical and imaging data in a unified AI model, an enhanced predictive diagnostic tool could be developed to further improve patient outcomes and reduce diagnostic delays.

6.4 Final Remarks

This dissertation has successfully demonstrated the feasibility and effectiveness of AI-driven approaches in the early diagnosis of endometriosis. By leveraging ML and DL techniques, this research has contributed to the field of AI-assisted medical diagnostics, offering a non-invasive, automated solution for identifying endometriosis. The promising performance of the developed models underscores the potential of AI in detecting endometriosis, reducing diagnostic delays and improving patient outcomes. Ultimately, this work serves as a foundation for future AI-driven endometriosis diagnostic systems and highlights the transformative potential of AI in medical applications.

References

[1] K. E. Nnoaham, L. Hummelshoj, P. Webster, T. d'Hooghe, F. de Cicco Nardone, C. de Cicco Nardone, C. Jenkinson, S. H. Kennedy, and K. T. Zondervan, "Impact of endometriosis on quality of life and work productivity: A multicenter study across ten countries," *Fertility and Sterility*, vol. 96, no. 2, pp. 366–373, 2011. DOI: 10.1016/j.fertnstert.2011.05.090.

[2] K. T. Zondervan, C. M. Becker, and S. A. Missmer, "Endometriosis," *New England Journal of Medicine*, vol. 382, no. 13, pp. 1244–1256, 2020. DOI: 10.1056/NEJMra1810764.

[3] K. Zieliński, D. Drabczyk, M. Kunicki, D. Drzyzga, A. Kloska, and J. Rumiński, "Evaluating the risk of endometriosis based on patients' self-assessment questionnaires," *Reproductive Biology and Endocrinology*, vol. 21, no. 1, pp. 1–13, 2023. DOI: 10.1186/s12958-023-01156-9.

[4] A. Hine, J. Bowles, and T. Webber, "The need for a non-invasive technology for endometriosis detection and care," in *Stud Health Technol Inform.*, 2023. DOI: 10.3233/SHTI230073.

[5] F. Chollet. "Keras." (2015), [Online]. Available: https://github.com/fchollet/keras.

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

[7] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16, ACM, 2016, pp. 785–794, ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785.

[8] E. Pascoal, J. Wessels, M. Aas-Eng, M. Abrao, G. Condous, D. Jurkovic, M. Espada, C. Exacoustos, S. Ferrero, S. Guerriero, G. Hudelist, M. Malzoni, S. Reid, S. Tang, C. Tomassetti, S. Singh, T. Van den Bosch, and M. Leonardi, "Strengths and limitations of diagnostic tools for endometriosis and relevance in diagnostic test accuracy research," *Ultrasound Obstet Gynecol*, vol. 60, no. 3, pp. 309–327, 2022. DOI: 10.1002/uog.24892.

[9] E. J. Kleczyk, T. Yadav, and S. Amirtharaj, "Applying machine learning algorithms to predict endometriosis onset," in *Endometriosis*, G. A. Gonçalves, Ed., IntechOpen, 2021, ch. 7. DOI: 10.5772/intechopen.101391.

[10] M. A. Al-Antari, “Artificial intelligence for medical diagnostics—existing and future ai technology!” *Diagnostics*, vol. 13, no. 4, p. 688, 2023. DOI: 10.3390/diagnostics13040688.

[11] I. Sadrehaghighi, *Artificial intelligence (ai) & machine learning (ml/dl/nns)*, Mar. 2024. DOI: 10.13140/RG.2.2.20926.05444.

[12] I. H. Sarker, “Machine learning: Algorithms, real-world applications and research directions,” *SN Computer Science*, vol. 2, no. 160, 2021. DOI: 10.1007/s42979-021-00592-x.

[13] S. Bendifallah, A. Puchar, S. Suisse, L. Delbos, M. Poilblanc, P. Descamps, F. Golfier, C. Touboul, Y. Dabi, and E. Daraï, “Machine learning algorithms as new screening approach for patients with endometriosis,” *Scientific Reports*, vol. 12, no. 1, Jan. 2022. DOI: 10.1038/s41598-021-04637-2.

[14] S. Bomrah, M. Uddin, U. Upadhaya, M. Komorowski, J. Priya, E. Dhar, S.-C. Hsu, and S. A. Shabbir, “A scoping review of machine learning for sepsis prediction-feature engineering strategies and model performance: A step towards explainability,” *Critical care (London, England)*, vol. 28, no. 1, p. 180, May 2024. DOI: 10.1186/s13054-024-04948-6.

[15] N. Pudjihartono, T. Fadason, A. Kempa-Liehr, and J. O’Sullivan, “A review of feature selection methods for machine learning-based disease risk prediction,” *Frontiers in Bioinformatics*, vol. 2, Jun. 2022. DOI: 10.3389/fbinf.2022.927312.

[16] S. Alsenan, I. Al-Turaiki, and A. Hafez, “Feature extraction methods in quantitative structure–activity relationship modeling: A comparative study,” *IEEE Access*, vol. 8, pp. 78 737–78 752, Apr. 2020. DOI: 10.1109/ACCESS.2020.2990375.

[17] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *Arxiv*, Sep. 2014. DOI: 10.48550/arXiv.1409.1556.

[18] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[19] K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks,” *Arxiv*, Mar. 2016. DOI: 10.48550/arXiv.1603.05027.

[20] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269. DOI: 10.1109/CVPR.2017.243.

[21] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, " Rethinking the Inception Architecture for Computer Vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2016, pp. 2818–2826. DOI: 10.1109/CVPR.2016.308.

[22] F. Chollet, " Xception: Deep Learning with Depthwise Separable Convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA: IEEE Computer Society, Jul. 2017, pp. 1800–1807. DOI: 10.1109/CVPR.2017.195.

[23] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, Feb. 2017. DOI: 10.1609/aaai.v31i1.11231.

[24] M. Tan and Q. V. Le, "Efficientnetv2: Smaller models and faster training," in *International Conference on Machine Learning*, Apr. 2021. DOI: 10.48550/arXiv.2104.00298.

[25] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for mobilenetv3," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1314–1324. DOI: 10.1109/ICCV.2019.00140.

[26] M. Iman, H. R. Arabnia, and K. Rasheed, "A review of deep transfer learning and recent advancements," *Technologies*, vol. 11, no. 2, p. 40, Mar. 2023, ISSN: 2227-7080. DOI: 10.3390/technologies11020040.

[27] M. Hossin and S. M.N, "A review on evaluation metrics for data classification evaluations," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, pp. 01–11, Mar. 2015. DOI: 10.5121/ijdkp.2015.5201.

[28] Z. Vujovic, "Classification model evaluation metrics," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, pp. 599–606, Jul. 2021. DOI: 10.14569/IJACSA.2021.0120670.

[29] S. Alowais, S. Alghamdi, N. Alsuhebany, T. Alqahtani, A. Alshaya, S. Almohareb, A. Aldairem, M. Alrashed, K. Saleh, H. Badreldin, M. Al Yami, S. Al Harbi, and A. Albekairy, "Revolutionizing healthcare: The role of artificial intelligence in clinical practice," *BMC Medical Education*, vol. 23, no. 1, Sep. 2023. DOI: 10.1186/s12909-023-04698-z.

[30] V. R. Umapathy, S. Rajinikanth B, R. D. Samuel Raj, S. Yadav, S. A. Munavarah, P. A. Anandapandian, A. V. Mary, K. Padmavathy, and A. R, "Perspective of artificial intelligence in disease diagnosis: A review of current and future endeavours in the medical field," *Cureus*, vol. 15, no. 9, Sep. 2023, ISSN: 2168-8184. DOI: 10.7759/cureus.45684.

[31] B. Dungate, D. R. Tucker, E. Goodwin, and P. J. Yong, "Assessing the utility of artificial intelligence in endometriosis: Promises and pitfalls," *Women's Health*, vol. 20, no. 3, Jan. 2024, ISSN: 1745-5065. DOI: 10.1177/17455057241248121.

[32] B. Sivajohan, M. Elgendi, C. Menon, C. Allaire, P. Yong, and M. A. Bedaiwy, "Clinical use of artificial intelligence in endometriosis: A scoping review," *npj Digital Medicine*, vol. 5, no. 1, Aug. 2022, ISSN: 2398-6352. DOI: 10.1038/s41746-022-00638-1.

[33] M. Szubert, A. Rycerz, and J. R. Wilczyński, "How to improve non-invasive diagnosis of endometriosis with advanced statistical methods," *Medicina*, vol. 59, no. 3, p. 499, Mar. 2023, ISSN: 1648-9144. DOI: 10.3390/medicina59030499.

[34] U. Tore, A. Abilgazym, A. Asunsolo-del-Barco, M. Terzic, Y. Yemenkhan, A. Zollanvari, and A. Sarria-Santamera, "Diagnosis of endometriosis based on comorbidities: A machine learning approach," *Biomedicines*, vol. 11, no. 11, p. 3015, Nov. 2023, ISSN: 2227-9059. DOI: 10.3390/biomedicines11113015.

[35] N. Zhao, T. Hao, F. Zhang, Q. Ni, D. Zhu, Y. Wang, Y. Shi, and X. Mi, "Application of machine learning techniques in the diagnosis of endometriosis," *BMC Women's Health*, vol. 24, no. 491, Sep. 2024. DOI: 10.1186/s12905-024-03334-2.

[36] F.-L. Pei, J.-J. Jia, S.-H. Lin, X.-X. Chen, L.-Z. Wu, Z.-X. Lin, B.-W. Sun, and C. Zeng, "Construction and evaluation of endometriosis diagnostic prediction model and immune infiltration based on efferocytosis-related genes," *Frontiers in Molecular Biosciences*, vol. 10, Jan. 2024, ISSN: 2296-889X. DOI: 10.3389/fmolb.2023.1298457.

[37] NCBI, *Gene expression omnibus (geo)*, Accessed: 2025-03-19, 2025. [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/.

[38] Q. Chen, Y. Jiao, Z. Yin, X. Fu, S. Guo, J. Xiang, and Y. Wang, "Establishment of a novel glycolysis-immune-related diagnosis gene signature for endometriosis by machine learning," *J Assist Reprod Genet.*, vol. 40, no. 5, pp. 1147–1161, May 2023. DOI: 10.1007/s10815-023-02769-0.

[39] H. Zhang, H. Zhang, H. Yang, A. N. Shuid, D. Sandai, and X. Chen, "Machine learning-based integrated identification of predictive combined diagnostic biomarkers for endometriosis," *Frontiers in Genetics*, vol. 14, Nov. 2023, ISSN: 1664-8021. DOI: 10.3389/fgene.2023.1290036.

[40] S. Akter, D. Xu, S. C. Nagel, J. J. Bromfield, K. Pelch, G. B. Wilshire, and T. Joshi, "Machine learning classifiers for endometriosis using transcriptomics and methylomics data," *Frontiers in Genetics*, vol. 10, Sep. 2019, ISSN: 1664-8021. DOI: 10.3389/fgene.2019.00766.

[41] Ziwig, *Ziwig – transforming women's health with salivary rna and ai*, Accessed: 2025-03-19, 2025. [Online]. Available: https://ziwig.com/en/home/.

[42] A. Goldstein and S. Cohen, "Self-report symptom-based endometriosis prediction using machine learning," *Scientific Reports*, vol. 13, no. 1, Apr. 2023. DOI: 10.1038/s41598-023-32761-8.

[43] S. Visalaxi and T. S. Muthu, "Automated prediction of endometriosis using deep learning," *International Journal of Nonlinear Analysis and Applications*, vol. 12, no. 2, pp. 2403–2416, Jan. 2021. DOI: 10.22075/ijnaa.2021.5383.

[44] A. Leibetseder, S. Kletz, K. Schoeffmann, S. Keckstein, and J. Keckstein, "GLENDA: gynecologic laparoscopy endometriosis dataset," in *MultiMedia Modeling - 26th International Conference, MMM 2020, Daejeon, South Korea, January 5-8, 2020, Proceedings, Part II*, ser. Lecture Notes in Computer Science, vol. 11962, Springer, 2020, pp. 439–450. DOI: 10.1007/978-3-030-37734-2_36.

[45] A. Leibetseder, K. Schoeffmann, J. Keckstein, and S. Keckstein, "Endometriosis detection and localization in laparoscopic gynecology," *Multimedia Tools and Applications*, vol. 81, pp. 6191–6215, Feb. 2022. DOI: 10.1007/s11042-021-11730-1.

[46] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Cham: Springer International Publishing, 2014, pp. 740–755, ISBN: 978-3-319-10602-1. DOI: 10.1007/978-3-319-10602-1_48.

[47] FEMaLe, *Female in a nutshell*, Accessed: 2025-03-19, 2025. [Online]. Available: https://findingendometriosis.eu/about/female-in-a-nutshell/.

[48] Imagendo, *Imagendo – supporting endometriosis awareness and research*, Accessed: 2025-03-19, 2025. [Online]. Available: https://imagendo.org.au/.

[49] The Pandas Development Team, *Pandas-dev/pandas: Pandas*, version 1.5.1, Feb. 2020. DOI: 10.5281/zenodo.3509134.

[50] T. A. Caswell, M. Droettboom, A. Lee, E. S. de Andrade, T. Hoffmann, J. Hunter, J. Klymak, E. Firing, D. Stansby, N. Varoquaux, J. H. Nielsen, B. Root, R. May, P. Elson, J. K. Seppänen, D. Dale, J.-J. Lee, D. McDougall, A. Straw, P. Hobson, hannah, C. Gohlke, T. S. Yu, E. Ma, A. F. Vincent, S. Silvester, C. Moad, N. Kniazev, E. Ernest, and P. Ivanov, *Matplotlib/matplotlib: Rel: V3.4.3*, version v3.4.3, Aug. 2021. DOI: 10.5281/zenodo.5194481.

[51] M. L. Waskom, "Seaborn: Statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021. DOI: 10.21105/joss.03021.

[52] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020. DOI: 10.1038/s41586-020-2649-2.

[53] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[54] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Y. Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng, *TensorFlow: Large-scale machine learning on heterogeneous systems*, Software available from tensorflow.org, 2015. [Online]. Available: https://www.tensorflow.org/.

Appendix A Software and Libraries

The work for this dissertation was conducted within a customised Anaconda environment, using Jupyter Notebooks and the Python programming language. The Python libraries listed below have been employed extensively to facilitate model development, performance evaluation, and data visualisation.

- **NumPy [41]:** NumPy was primarily used in DL modelling for array transformation of the image dataset.
- **Pandas [34]:** This library was used throughout the dissertation for data manipulation, preprocessing and the handling of the DataFrames.
- **OpenCV [42]:** OpenCV is a Computer Vision library that was utilised to augment and preprocess the image dataset.
- **Matplotlib [35]:** Matplotlib is a data visualisation tool that was vital in creating all plots and graphs in this study.
- **Seaborn [36]:** This library was used in conjunction with Matplotlib to further improve the interpretability of the plots and graphs.
- **Scikit-learn [37]:** This library played a central role in multiple phases of the development process, including model initialisation, preprocessing, feature selection, training and evaluation.
- **XGBoost [38]:** The XGBoost library was specific used to initialise and train the XGBoost classifier.
- **TensorFlow [39]:** This library is used along with Keras to build and train the CNN architectures for the DL modelling.
- **Keras [40]:** Keras an API built on TensorFlow that was used to facilitate the building and training of the DL models of this study.

Appendix B Machine Learning Model Results

This Appendix presents the results of the ML models trained on the PCA algorithms with 29 components as well as after applying hyperparameter tuning.

Table B.1 Performance Metrics of PCA Model with 29 Components

Model	Metrics					
	Accuracy	Precision	Recall	F1 Score	ROC AUC	Std. Dev.
Logistic Regression	94.14	95.58	93.10	94.32	94.19	2.11
Random Forest	91.44	89.43	94.83	92.05	91.28	1.76
XGBoost	92.34	92.31	93.10	92.70	92.31	1.75
Decision Tree	87.84	86.78	90.52	88.61	87.71	2.05
SVM	95.05	95.65	94.83	95.24	95.06	1.88
AdaBoost	91.89	92.98	91.38	92.17	91.92	2.13

Table B.2 Performance Metrics of Fine-Tuned PCA Models with 29 Components

Model	Metrics					
	Accuracy	Precision	Recall	F1 Score	ROC AUC	Std. Dev.
Logistic Regression	93.69	95.54	92.24	93.86	93.76	1.83
Random Forest	90.09	89.17	92.24	90.68	89.99	1.53
XGBoost	89.64	89.74	90.52	90.13	89.60	1.68
Decision Tree	88.29	90.91	86.21	88.50	88.39	1.97
SVM	94.14	94.78	93.97	94.37	94.15	2.45
AdaBoost p	91.44	92.17	91.38	91.77	91.44	1.22

Appendix C Machine Learning Feature Selection Results

This Appendix presents the results of the feature selection algorithms applied on the self-reported patient symptom dataset during ML modelling. The findings of the Correlation Matrix, Chi-Square Test and General Feature Importance are detailed in this document. Additionally, the Model-Based Feature Importance Graphs and their corresponding results are detailed. Finally, the list of features selected by both the Forward Feature Selection (FFS) and Backward Feature Selection (BFS) for each ML algorithm are presented.

C.1 Correlation Matrix

Table C.1 Correlation Matrix Results

Feature	Correlation Value
Menstrual pain (Dysmenorrhea)	0.713039
Painful cramps during period	0.611767
Cramping	0.602516
Fatigue / Chronic fatigue	0.598330
Heavy / Extreme menstrual bleeding	0.594275
Bleeding	0.569073
Pelvic pain	0.567080
Abdominal pain / pressure	0.563149
Painful / Burning pain during sex (Dyspareunia)	0.552379
Painful bowel movements	0.495025
Ovarian cysts	0.491360
Back pain	0.489094
Bloating	0.484099
Lower back pain	0.483313
Sharp / Stabbing pain	0.482743
Menstrual clots	0.478226
Stomach cramping	0.465247
Decreased energy / Exhaustion	0.454389
Pain / Chronic pain	0.453660
Irregular / Missed periods	0.436597
Cysts	0.436543

Feature	Correlation Value
Pain after Intercourse	0.431117
Painful ovulation	0.423201
IBS-like symptoms	0.418389
Extreme / Severe pain	0.417709
Constipation / Chronic constipation	0.414955
Hormonal problems	0.409041
Nausea	0.401442
Abdominal Cramps during Intercourse	0.377198
Vaginal Pain/Pressure	0.371865
Anxiety	0.366846
Digestive / GI problems	0.364051
Long menstruation	0.363956
Infertility	0.337573
Acne / pimples	0.335421
Mood swings	0.332986
Anaemia / Iron deficiency	0.328626
Painful urination	0.322389
Irritable Bowel Syndrome (IBS)	0.312430
Depression	0.309255
Excessive bleeding	0.298194
Diarrhea	0.291039
Feeling sick	0.273422
Hip pain	0.272833
Leg pain	0.266174
Insomnia / Sleeplessness	0.261053
Dizziness	0.253644
Fertility Issues	0.252305
Bowel pain	0.248704
Syncope (fainting, passing out)	0.241837
Headaches	0.236963
Constant bleeding	0.218317
Vomiting / constant vomiting	0.199626
Migraines	0.163858
Loss of appetite	0.157223
Abnormal uterine bleeding	0.151067
Malaise / Sickness	0.130251
Fever	-0.142806

C.2 Chi-Square Test

Table C.2 Chi-Square Test Results

Feature	P-Value
Menstrual pain (Dysmenorrhea)	0.0000000
Painful cramps during period	0.0000000
Fatigue or chronic fatigue	0.0000000
Abdominal pain or pressure	0.0000000
Heavy or extreme menstrual bleeding	0.0000000
Painful or burning pain during sex (Dyspareunia)	0.0000000
Painful bowel movements	0.0000000
Ovarian cysts	0.0000000
Cramping	0.0000000
Pelvic pain	0.0000000
Menstrual clots	0.0000000
Bleeding	0.0000000
Constipation or chronic constipation	0.0000000
Sharp or stabbing pain	0.0000000
Stomach cramping	0.0000000
Lower back pain	0.0000000
Irregular or missed periods	0.0000000
Bloating	0.0000000
Cysts	0.0000000
Pain after intercourse	0.0000000
IBS-like symptoms	0.0000000
Hormonal problems	0.0000000
Pain or chronic pain	0.0000000
Decreased energy or exhaustion	0.0000000
Extreme or severe pain	0.0000000
Vaginal pain/pressure	0.0000000
Painful ovulation	0.0000000
Back pain	0.0000000
Long menstruation	0.0000000
Abdominal cramps during intercourse	0.0000000
Digestive or GI problems	0.0000000
Nausea	0.0000000
Infertility	0.0000000

Feature	P-Value
Anaemia or iron deficiency	0.0000000
Painful urination	0.0000000
Anxiety	0.0000000
Acne or pimples	0.0000000
Mood swings	0.0000000
Irritable bowel syndrome (IBS)	0.0000000
Fertility issues	0.0000000
Depression	0.0000000
Excessive bleeding	0.0000000
Leg pain	0.0000000
Diarrhea	0.0000000
Hip pain	0.0000000
Insomnia or sleeplessness	0.0000000
Syncope (fainting, passing out)	0.0000000
Feeling sick	0.0000000
Bowel pain	0.0000000
Dizziness	0.0000015
Headaches	0.0000023
Constant bleeding	0.0000027
Vomiting or constant vomiting	0.0000126
Fever	0.0002056
Abnormal uterine bleeding	0.0004225
Migraines	0.0051575
Malaise or sickness	0.0069860
Loss of appetite	0.0164577

C.3 General Model Feature Importance

Table C.3 General Model Feature Importance Results

Feature	Feature Importance Value
Menstrual pain (Dysmenorrhea)	0.2629270
Painful cramps during period	0.2258320
Cramping	0.1966471
Fatigue or chronic fatigue	0.1813170
Bleeding	0.1793968
Pelvic pain	0.1776877
Abdominal pain or pressure	0.1742195
Heavy or extreme menstrual bleeding	0.1519892
Painful or burning pain during sex (Dyspareunia)	0.1433973
Painful bowel movements	0.1373453
Lower back pain	0.1324505
Back pain	0.1312131
Menstrual clots	0.1270394
Irregular or missed periods	0.1265204
Ovarian cysts	0.1250150
Pain after intercourse	0.1237204
Sharp or stabbing pain	0.1219097
Stomach cramping	0.1184357
Extreme or severe pain	0.1083518
Pain or chronic pain	0.1030041
Constipation or chronic constipation	0.1016999
Cysts	0.0971805
Anaemia or iron deficiency	0.0961477
Hormonal problems	0.0942230
IBS-like symptoms	0.0933913
Bloating	0.0920850
Long menstruation	0.0908480
Painful urination	0.0720355
Infertility	0.0712786
Decreased energy or exhaustion	0.0693257
Nausea	0.0689219
Painful ovulation	0.0674035
Vaginal pain or pressure	0.0617363

Feature	Feature Importance Value
Abdominal cramps during intercourse	0.0582973
Digestive or GI problems	0.0535346
Mood swings	0.0469268
Bowel pain	0.0435719
Anxiety	0.0427989
Irritable bowel syndrome (IBS)	0.0424385
Feeling sick	0.0411288
Diarrhea	0.0376182
Depression	0.0370352
Hip pain	0.0349941
Fertility issues	0.0334801
Syncope (Fainting, passing out)	0.0295857
Insomnia or sleeplessness	0.0283056
Excessive bleeding	0.0228295
Migraines	0.0205476
Vomiting or constant vomiting	0.0086506
Constant bleeding	0.0079285
Headaches	0.0067188
Acne or pimples	0.0049049
Dizziness	0.0032364
Malaise or sickness	0.0016613
Leg pain	0.0
Abnormal uterine bleeding	0.0
Fever	0.0
Loss of appetite	0.0

C.4 Model Feature Importance Graphs

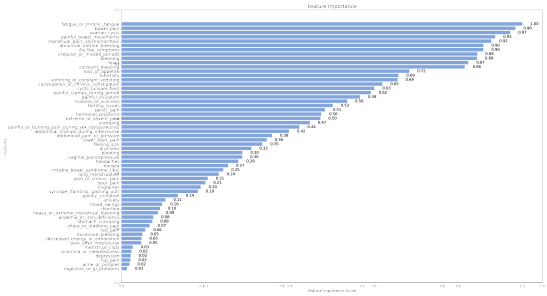


Figure C.1 Logistic Regression Feature Importance

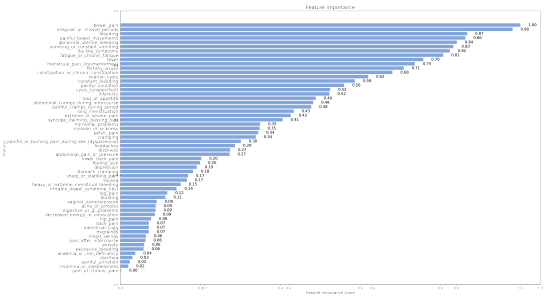


Figure C.2 SVM Feature Importance

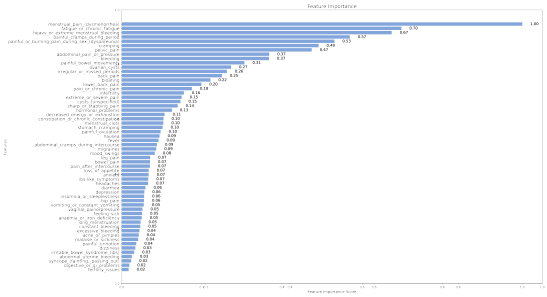


Figure C.3 Random Forest Feature Importance

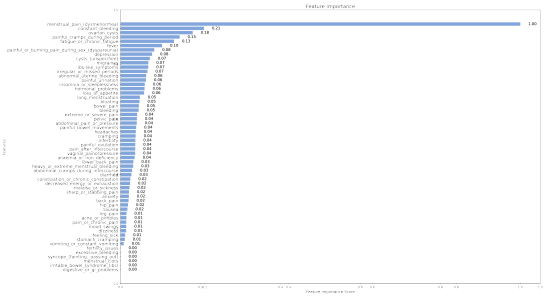


Figure C.4 XGBoost Feature Importance

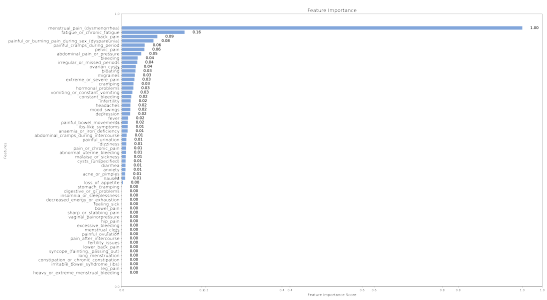


Figure C.5 Decision Tree Feature Importance

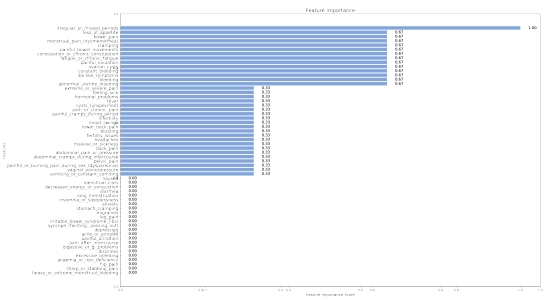


Figure C.6 AdaBoost Feature Importance

Table C.4 Model-Based Feature Importance Results

Feature	LR	SVM	RF	XGB	DT	Ada
Abdominal cramps during intercourse	0.4175	0.4820	0.0878	0.0289	0.0123	0.3333
Abdominal pain or pressure	0.3753	0.2727	0.3682	0.0405	0.0485	0.3333
Abnormal uterine bleeding	0.9030	0.8413	0.0254	0.0641	0.0106	0.6667
Acne or pimples	0.0192	0.0879	0.0424	0.0147	0.0084	0.0000
Anaemia or iron deficiency	0.0786	0.0370	0.0499	0.0353	0.0147	0.0000
Anxiety	0.1091	0.0591	0.0671	0.0207	0.0094	0.0000
Back pain	0.2087	0.0707	0.2499	0.0200	0.0888	0.3333
Bleeding	0.8868	0.8671	0.3676	0.0453	0.0398	0.6667
Bloating	0.3012	0.1123	0.2213	0.0474	0.0341	0.3333
Bowel pain	0.9828	1.0000	0.0704	0.0458	0.0000	0.6667
Constant bleeding	0.8562	0.5865	0.0462	0.2087	0.0235	0.6667
Constipation or chronic constipation	0.6506	0.6800	0.1043	0.0245	0.0000	0.6667
Cramping	0.4695	0.3382	0.4914	0.0378	0.0294	0.6667
Cysts	0.6307	0.5244	0.1466	0.0732	0.0096	0.3333
Decreased energy or exhaustion	0.0492	0.0858	0.1071	0.0238	0.0000	0.0000
Depression	0.0223	0.1906	0.0564	0.0789	0.0209	0.0000
Diarrhea	0.0954	0.0295	0.0592	0.0272	0.0094	0.0000
Digestive or GI problems	0.0131	0.0877	0.0189	0.0000	0.0000	0.0000
Dizziness	0.3236	0.2738	0.0339	0.0140	0.0112	0.0000
Excessive bleeding	0.0515	0.0578	0.0440	0.0000	0.0000	0.0000
Extreme or severe pain	0.4959	0.4256	0.1496	0.0416	0.0315	0.3333
Fatigue or chronic fatigue	1.0000	0.8073	0.6983	0.1338	0.1568	0.6667
Feeling sick	0.3509	0.1989	0.0506	0.0115	0.0000	0.3333
Fertility issues	0.5268	0.7082	0.0172	0.0000	0.0000	0.3333
Fever	0.8654	0.7573	0.0919	0.1038	0.0157	0.3333
Headaches	0.2909	0.2865	0.0656	0.0381	0.0215	0.3333
Heavy or extreme menstrual bleeding	0.0903	0.1505	0.6739	0.0316	0.0000	0.0000
Hip pain	0.0217	0.0763	0.0554	0.0187	0.0000	0.0000
Hormonal problems	0.4980	0.3492	0.1252	0.0602	0.0288	0.3333
IBS-like symptoms	0.9028	0.8237	0.0662	0.0694	0.0149	0.6667
Infertility	0.6907	0.5225	0.1551	0.0371	0.0224	0.3333
Insomnia or sleeplessness	0.0239	0.0192	0.0559	0.0608	0.0000	0.0000
Irregular or missed periods	0.8877	0.9806	0.2624	0.0672	0.0387	1.0000
Irritable bowel syndrome (IBS)	0.2534	0.1400	0.0309	0.0000	0.0000	0.0000
Leg pain	0.0589	0.1169	0.0709	0.0147	0.0000	0.0000
Long menstruation	0.2421	0.4331	0.0469	0.0482	0.0000	0.0000

Feature	LR	SVM	RF	XGB	DT	Ada
Loss of appetite	0.7179	0.4884	0.0673	0.0590	0.0030	0.6667
Lower back pain	0.3619	0.2021	0.1986	0.0322	0.0000	0.3333
Malaise or sickness	0.5627	0.3479	0.0409	0.0222	0.0101	0.3333
Menstrual clots	0.0273	0.0706	0.1038	0.0000	0.0000	0.0000
Menstrual pain (Dysmenorrhea)	0.9225	0.7361	1.0000	1.0000	1.0000	0.6667
Migraines	0.1969	0.0706	0.0861	0.0694	0.0327	0.0000
Mood swings	0.1010	0.0636	0.0825	0.0142	0.0214	0.3333
Nausea	0.2653	0.1658	0.0944	0.0173	0.0083	0.0000
Ovarian cysts	0.9700	0.6193	0.2729	0.1805	0.0351	0.6667
Pain after intercourse	0.0485	0.0623	0.0702	0.0367	0.0000	0.0000
Pain or chronic pain	0.2149	0.0015	0.1752	0.0143	0.0106	0.3333
Painful bowel movements	0.9324	0.8625	0.3063	0.0390	0.0151	0.6667
Painful cramps during period	0.6219	0.4769	0.5693	0.1478	0.0573	0.3333
Dyspareunia	0.4435	0.3009	0.5311	0.0843	0.0791	0.3333
Painful ovulation	0.5947	0.5594	0.0971	0.0368	0.0000	0.6667
Painful urination	0.1399	0.0233	0.0363	0.0635	0.0114	0.0000
Pelvic pain	0.5073	0.3444	0.4744	0.0409	0.0564	0.3333
Sharp or stabbing pain	0.0697	0.1688	0.1396	0.0210	0.0000	0.0000
Stomach cramping	0.0763	0.1811	0.1020	0.0106	0.0000	0.0000
Syncope (Fainting, passing out)	0.1899	0.4054	0.0235	0.0000	0.0000	0.0000
Vaginal pain or pressure	0.3008	0.0905	0.0517	0.0366	0.0000	0.3333
Vomiting or constant vomiting	0.6886	0.8328	0.0549	0.0080	0.0264	0.3333

C.5 Features Selected by Feature Selection Methods

Note that for all the ML models, 29 distinct features were selected for both FFS and BFS feature selection strategies.

C.5.1 Logistic Regression

Features selected by FFS: heavy or extreme menstrual bleeding, menstrual pain (dysmenorrhea), irregular or missed periods, abdominal pain or pressure, infertility, painful cramps during period, long menstruation, constipation or chronic constipation, vomiting or constant vomiting, fatigue or chronic fatigue, syncope (fainting, passing out), mood swings, depression, bleeding, fertility issues, ovarian cysts, constant bleeding, digestive or GI problems, anaemia or iron deficiency, vaginal pain or pressure, bowel pain, anxiety, dizziness, malaise or sickness, abnormal uterine bleeding, decreased energy or exhaustion, abdominal cramps during intercourse, acne or pimples, loss of appetite.

Features selected by BFS: dysmenorrhea, dyspareunia, pelvic pain, irregular or missed periods, back pain, painful bowel movements, infertility, painful cramps during period, constipation or chronic constipation, vomiting or constant vomiting, fatigue or chronic fatigue, painful ovulation, extreme or severe pain, bleeding, lower back pain, ovarian cysts, constant bleeding, IBS-like symptoms, vaginal pain or pressure, sharp or stabbing pain, bowel pain, cysts , dizziness, abnormal uterine bleeding, fever, feeling sick, abdominal cramps during intercourse, insomnia or sleeplessness, loss of appetite.

C.5.2 Random Forest

Features selected by FFS: dysmenorrhea, pelvic pain, menstrual clots, infertility, painful cramps during period, constipation or chronic constipation, vomiting or constant vomiting, fatigue or chronic fatigue, painful ovulation, irritable bowel syndrome (IBS), syncope (fainting, passing out), mood swings, bleeding, lower back pain, fertility issues, ovarian cysts, constant bleeding, digestive or GI problems, IBS-like symptoms, excessive bleeding, sharp or stabbing pain, cysts , dizziness, malaise or sickness, abnormal uterine bleeding, hormonal problems, decreased energy or exhaustion, abdominal cramps during intercourse, insomnia or sleeplessness.

Features selected by BFS: dysmenorrhea, irregular or missed periods, cramping, back pain, painful bowel movements, nausea, infertility, constipation or chronic constipation, vomiting or constant vomiting, fatigue or chronic fatigue, migraines, extreme or severe pain, leg pain, bleeding, lower back pain, ovarian cysts, headaches, constant bleeding, pain after intercourse, hip pain, anxiety, cysts , dizziness, abnormal

uterine bleeding, fever, feeling sick, abdominal cramps during intercourse, acne or pimples, loss of appetite.

C.5.3 XGBoost

Features selected by FFS: heavy or extreme menstrual bleeding, dysmenorrhea, irregular or missed periods, cramping, painful bowel movements, infertility, painful cramps during period, diarrhea, long menstruation, fatigue or chronic fatigue, stomach cramping, irritable bowel syndrome (IBS), syncope (fainting, passing out), depression, bleeding, lower back pain, fertility issues, ovarian cysts, constant bleeding, digestive or GI problems, IBS-like symptoms, excessive bleeding, anaemia or iron deficiency, hip pain, vaginal pain or pressure, sharp or stabbing pain, cysts , fever, loss of appetite.

Features selected by BFS: heavy or extreme menstrual bleeding, dysmenorrhea, dyspareunia, irregular or missed periods, cramping, abdominal pain or pressure, painful bowel movements, infertility, painful cramps during period, vomiting or constant vomiting, fatigue or chronic fatigue, painful ovulation, extreme or severe pain, leg pain, bleeding, ovarian cysts, headaches, constant bleeding, excessive bleeding, vaginal pain or pressure, bowel pain, cysts , abnormal uterine bleeding, fever, hormonal problems, bloating, decreased energy or exhaustion, abdominal cramps during intercourse, loss of appetite.

C.5.4 Decision Tree

Features selected by FFS: dysmenorrhea, irregular or missed periods, cramping, painful bowel movements, infertility, painful cramps during period, diarrhea, long menstruation, stomach cramping, irritable bowel syndrome (IBS), syncope (fainting, passing out), lower back pain, fertility issues, ovarian cysts, constant bleeding, pain after intercourse, digestive or GI problems, IBS-like symptoms, anaemia or iron deficiency, vaginal pain or pressure, sharp or stabbing pain, anxiety, cysts , malaise or sickness, abnormal uterine bleeding, fever, bloating, abdominal cramps during intercourse, insomnia or sleeplessness.

Features selected by BFS: heavy or extreme menstrual bleeding, dysmenorrhea, dyspareunia, pelvic pain, irregular or missed periods, abdominal pain or pressure, back pain, painful bowel movements, nausea, infertility, painful cramps during period, pain or chronic pain, diarrhea, vomiting or constant vomiting, fatigue or chronic fatigue, painful ovulation, extreme or severe pain, syncope (fainting, passing out), depression, bleeding, ovarian cysts, excessive bleeding, bowel pain, anxiety, dizziness, abnormal uterine bleeding, fever, abdominal cramps during intercourse, loss of appetite.

C.5.5 SVM

Features selected by FFS: heavy or extreme menstrual bleeding, dysmenorrhea, pelvic pain, irregular or missed periods, cramping, back pain, painful bowel movements, painful cramps during period, long menstruation, constipation or chronic constipation, fatigue or chronic fatigue, painful ovulation, migraines, extreme or severe pain, irritable bowel syndrome (IBS), mood swings, bleeding, fertility issues, ovarian cysts, painful urination, constant bleeding, digestive or GI problems, excessive bleeding, anaemia or iron deficiency, bowel pain, malaise or sickness, fever, decreased energy or exhaustion, acne or pimples.

Features selected by BFS: heavy or extreme menstrual bleeding, dysmenorrhea, dyspareunia, pelvic pain, irregular or missed periods, cramping, abdominal pain or pressure, painful bowel movements, infertility, long menstruation, constipation or chronic constipation, vomiting or constant vomiting, fatigue or chronic fatigue, painful ovulation, mood swings, bleeding, fertility issues, ovarian cysts, headaches, constant bleeding, digestive or GI problems, IBS-like symptoms, bowel pain, cysts , malaise or sickness, abnormal uterine bleeding, fever, abdominal cramps during intercourse, loss of appetite.

C.5.6 AdaBoost

Features selected by FFS: heavy or extreme menstrual bleeding, dysmenorrhea, painful bowel movements, nausea, menstrual clots, painful cramps during period, fatigue or chronic fatigue, painful ovulation, stomach cramping, extreme or severe pain, leg pain, depression, bleeding, lower back pain, fertility issues, ovarian cysts, constant bleeding, pain after intercourse, digestive or GI problems, excessive bleeding, anaemia or iron deficiency, vaginal pain or pressure, sharp or stabbing pain, anxiety, dizziness, malaise or sickness, abnormal uterine bleeding, decreased energy or exhaustion, abdominal cramps during intercourse.

Features selected by BFS: dysmenorrhea, dyspareunia, pelvic pain, irregular or missed periods, cramping, back pain, painful bowel movements, painful cramps during period, pain or chronic pain, constipation or chronic constipation, vomiting or constant vomiting, fatigue or chronic fatigue, painful ovulation, mood swings, bleeding, fertility issues, ovarian cysts, headaches, constant bleeding, IBS-like symptoms, bowel pain, cysts , dizziness, malaise or sickness, abnormal uterine bleeding, fever, hormonal problems, abdominal cramps during intercourse, loss of appetite.

Appendix D Machine Learning Plots

This Appendix illustrates the ROC and PR plots of all six ML classifiers for the base, FFS, BFS, and PCA model variations before and after hyperpameter tuning.

D.1 Base Models

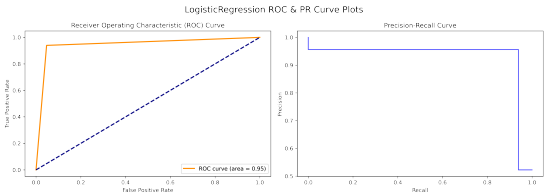


Figure D.1 Logistic Regression

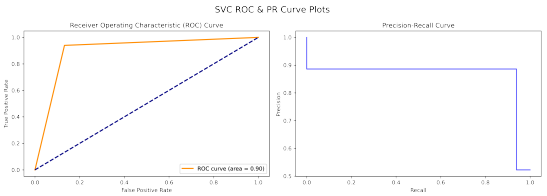


Figure D.2 SVM

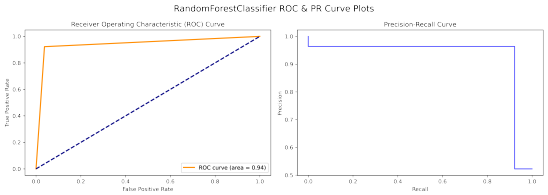


Figure D.3 Random Forest

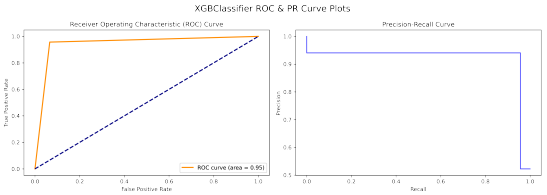


Figure D.4 XGBoost

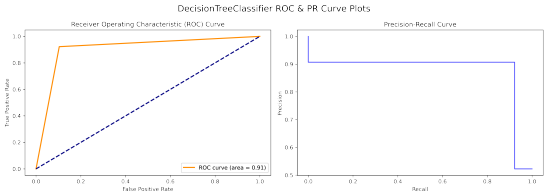


Figure D.5 Decision Tree

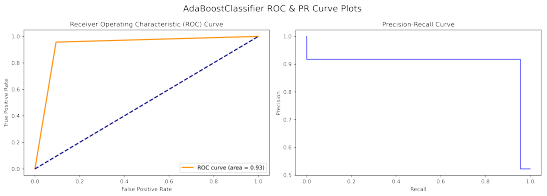


Figure D.6 AdaBoost

D.2 FFS-Based Models

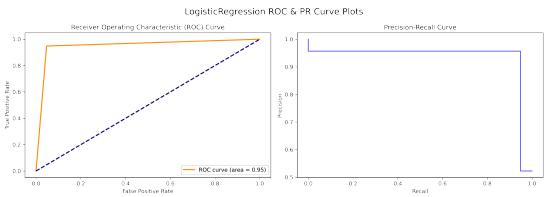


Figure D.7 Logistic Regression

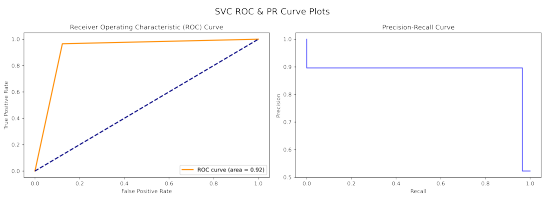


Figure D.8 SVM

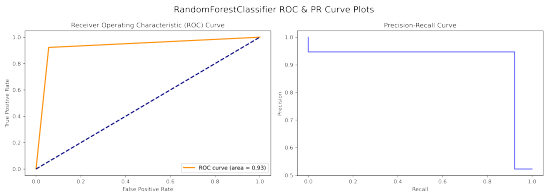


Figure D.9 Random Forest

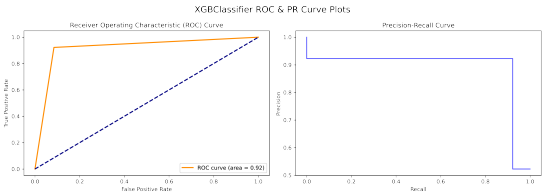


Figure D.10 XGBoost

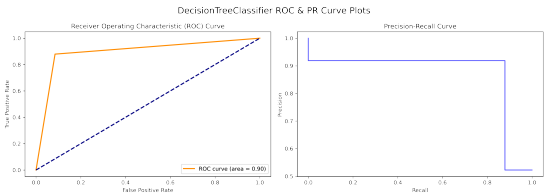


Figure D.11 Decision Tree

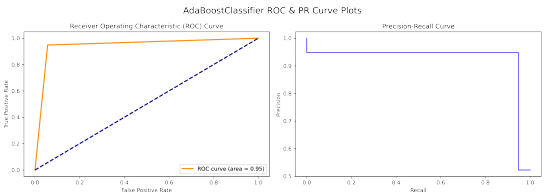


Figure D.12 AdaBoost

D.3 BFS-Based Models

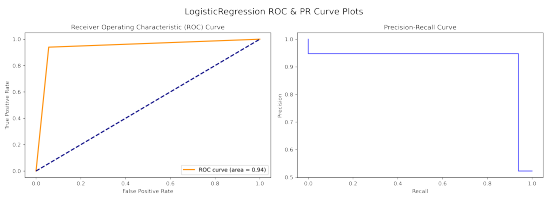


Figure D.13 Logistic Regression

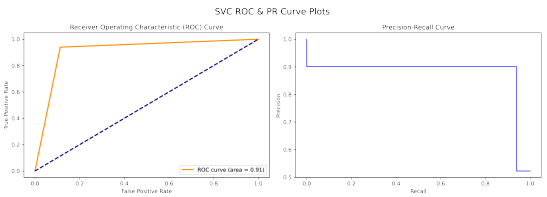


Figure D.14 SVM

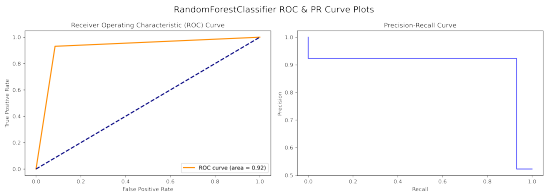


Figure D.15 Random Forest

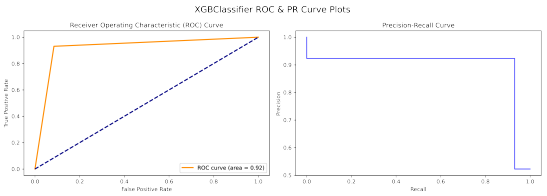


Figure D.16 XGBoost

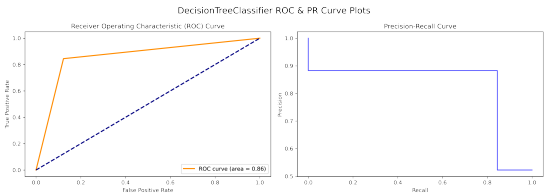


Figure D.17 Decision Tree

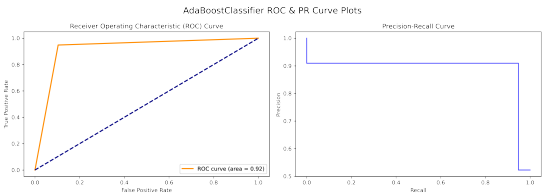


Figure D.18 AdaBoost

D.4 PCA-Based Models with 29 Components

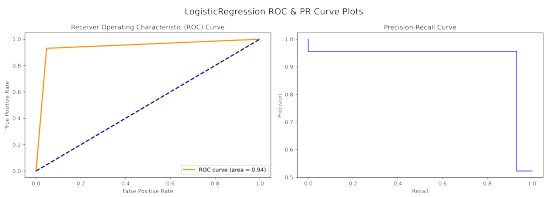


Figure D.19 Logistic Regression

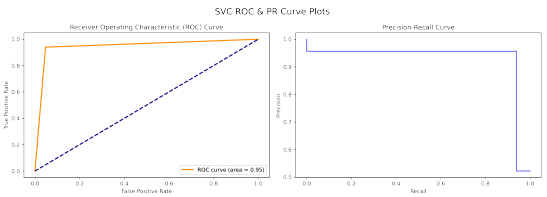


Figure D.20 SVM

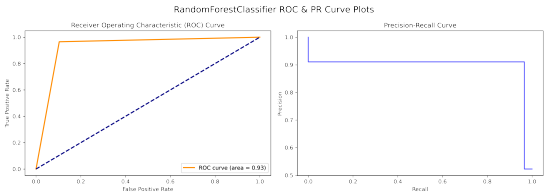


Figure D.21 Random Forest

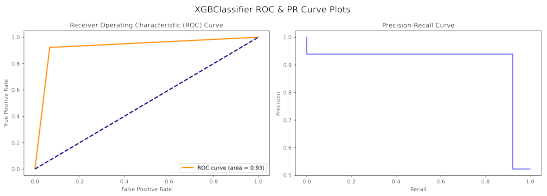


Figure D.22 XGBoost

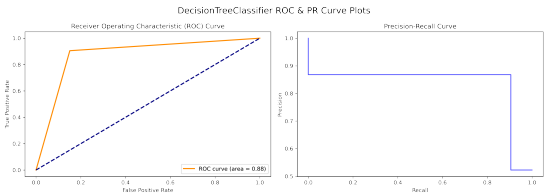


Figure D.23 Decision Tree

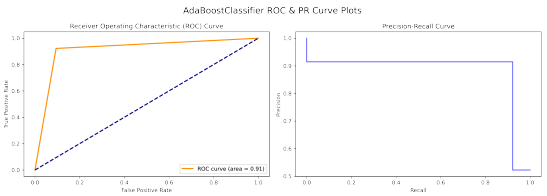


Figure D.24 AdaBoost

D.5 PCA-Based Models with 58 Components

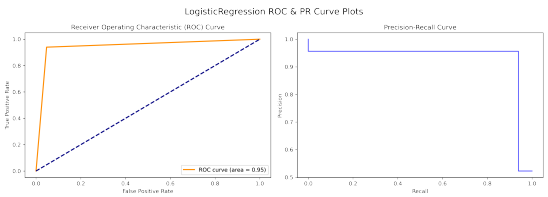


Figure D.25 Logistic Regression

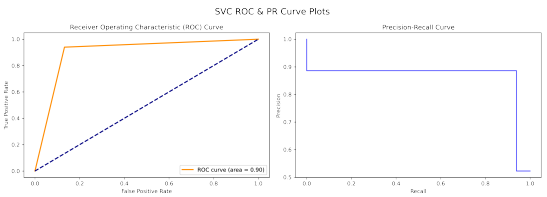


Figure D.26 SVM

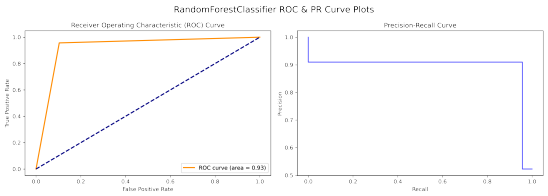


Figure D.27 Random Forest

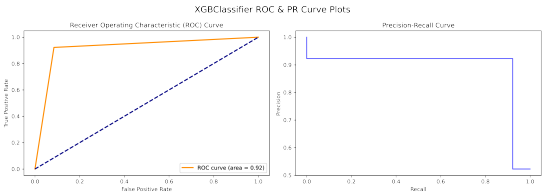


Figure D.28 XGBoost

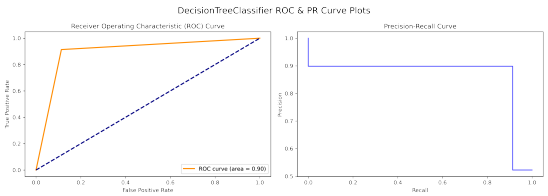


Figure D.29 Decision Tree

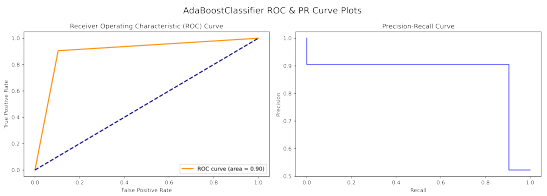


Figure D.30 AdaBoost

D.6 Base Models after Hyperparameter Tuning

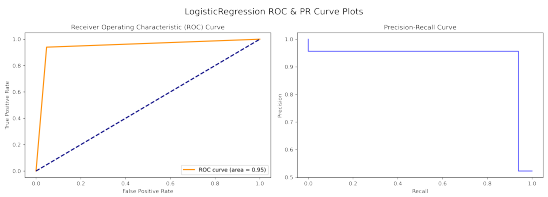


Figure D.31 Logistic Regression

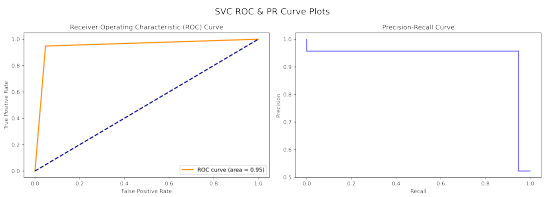


Figure D.32 SVM

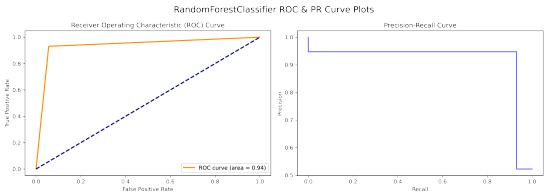


Figure D.33 Random Forest

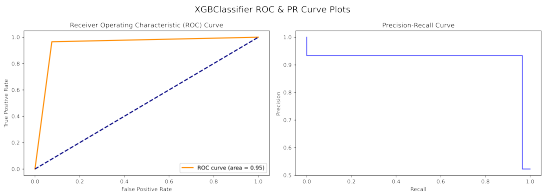


Figure D.34 XGBoost

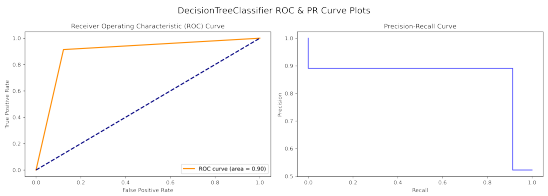


Figure D.35 Decision Tree

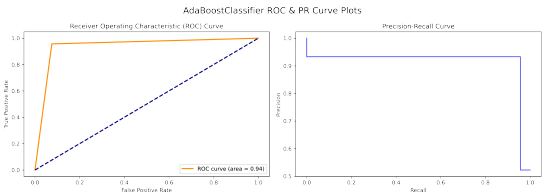


Figure D.36 AdaBoost

D.7 FFS-Based Models after Hyperparameter Tuning

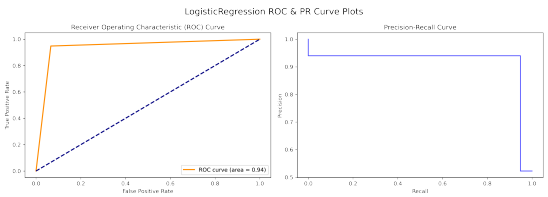


Figure D.37 Logistic Regression

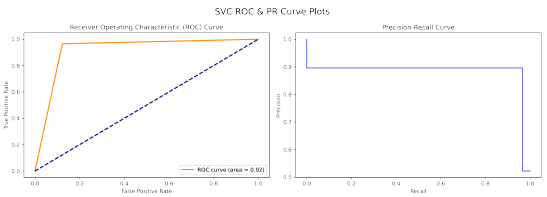


Figure D.38 SVM

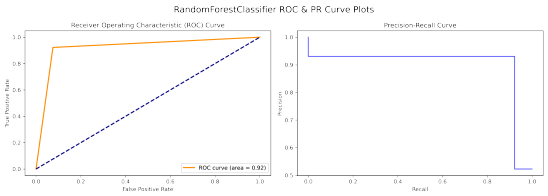


Figure D.39 Random Forest

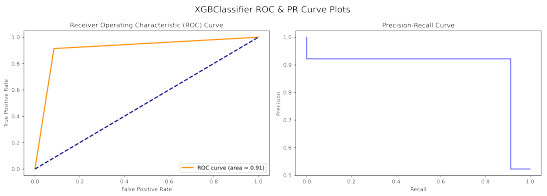


Figure D.40 XGBoost

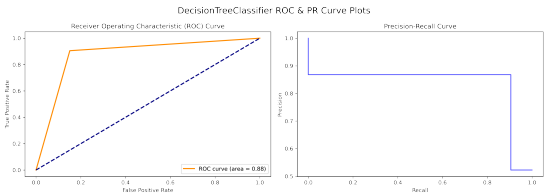


Figure D.41 Decision Tree

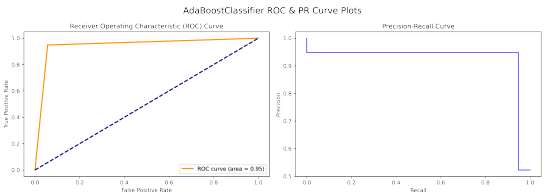


Figure D.42 AdaBoost

D.8 BFS-Based Models after Hyperparameter Tuning

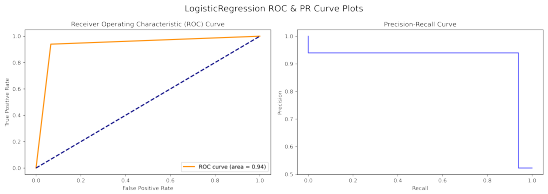


Figure D.43 Logistic Regression

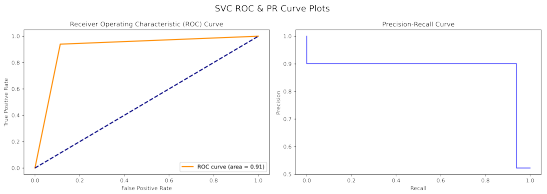


Figure D.44 SVM

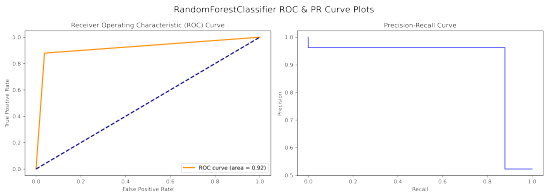


Figure D.45 Random Forest

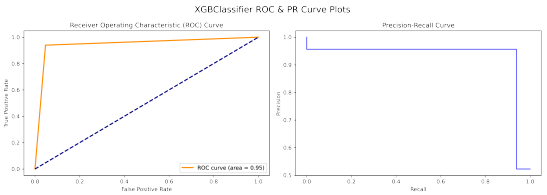


Figure D.46 XGBoost

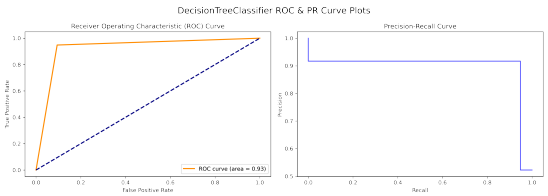


Figure D.47 Decision Tree

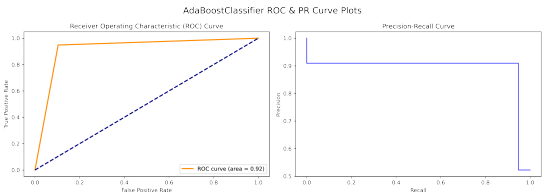


Figure D.48 AdaBoost

D.9 PCA-Based Models with 29 Components after Hyperparameter Tuning

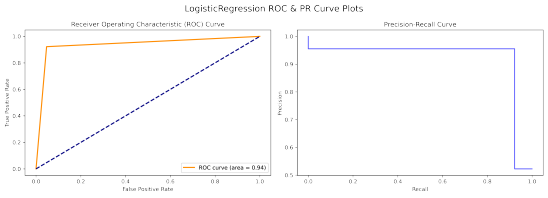


Figure D.49 Logistic Regression

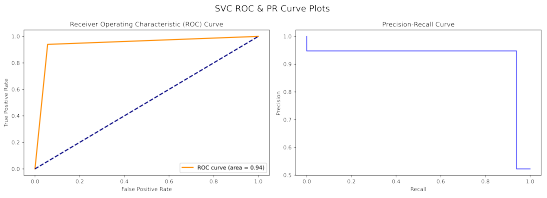


Figure D.50 SVM

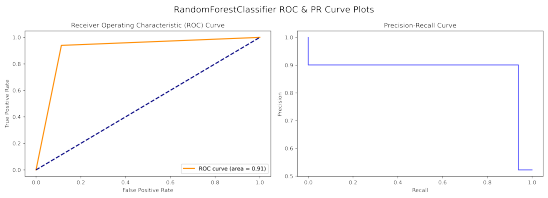


Figure D.51 Random Forest

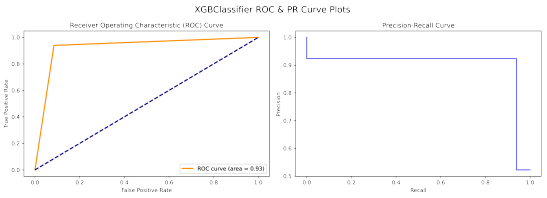


Figure D.52 XGBoost

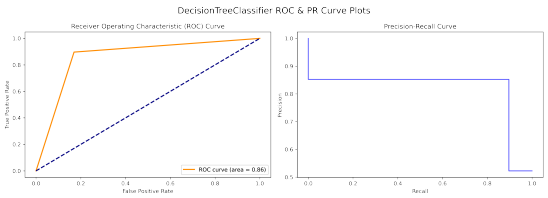


Figure D.53 Decision Tree

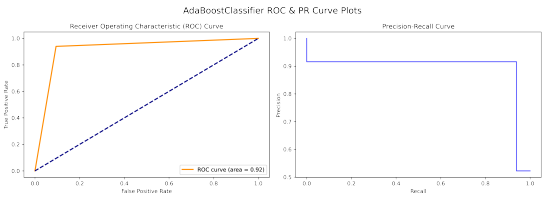


Figure D.54 AdaBoost

D.10 PCA-Based Models with 58 Components after Hyperparameter Tuning

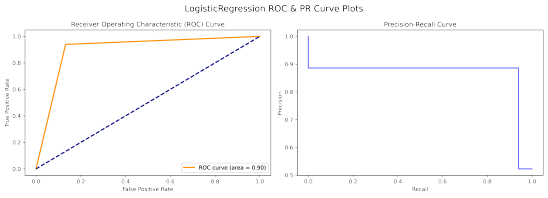


Figure D.55 Logistic Regression

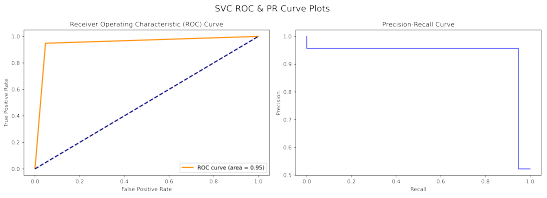


Figure D.56 SVM

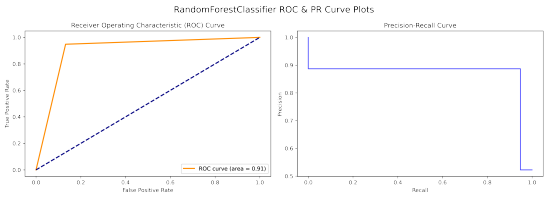


Figure D.57 Random Forest

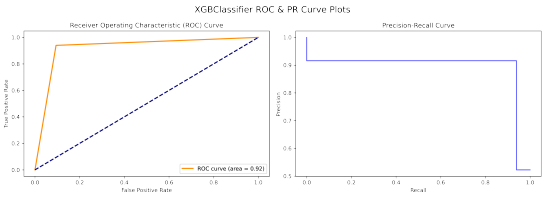


Figure D.58 XGBoost

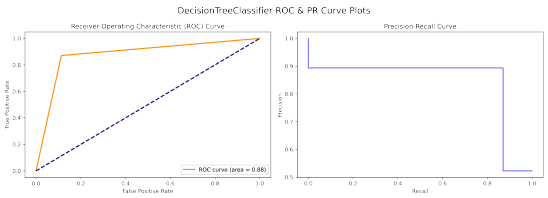


Figure D.59 Decision Tree

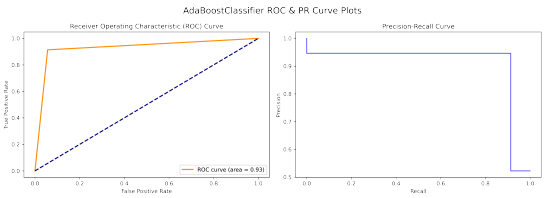


Figure D.60 AdaBoost

Appendix E Deep Learning Plots

This Appendix illustrates the accuracy and loss plots of the DL architectures for the base and data augmented model variations.

E.1 Base Models

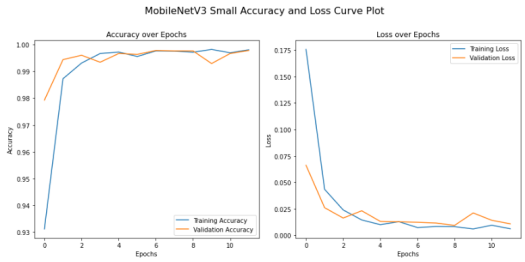


Figure E.1 MobileNetv3 Small

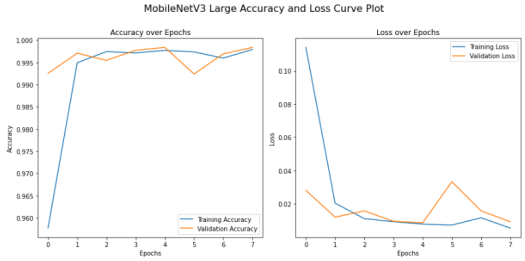


Figure E.2 MobileNetv3 Large

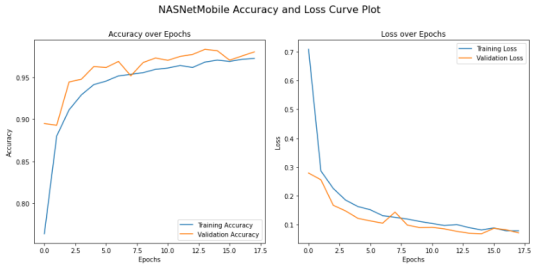


Figure E.3 NASNetMobile

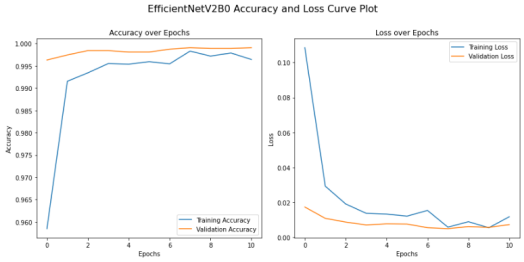


Figure E.4 EfficientNetV2B0

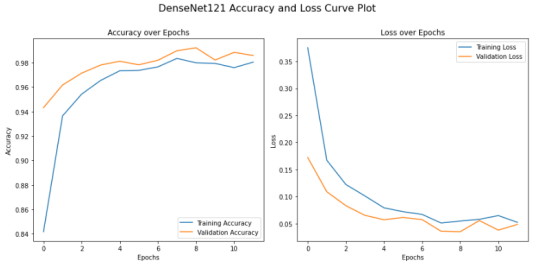


Figure E.5 DenseNet121

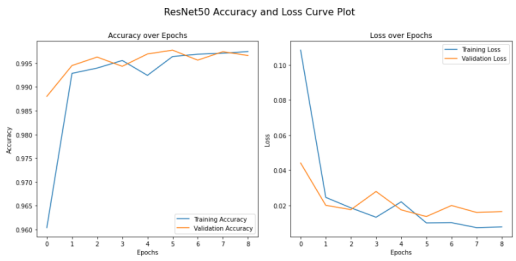


Figure E.6 ResNet50

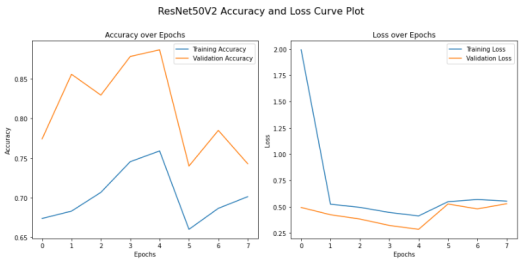


Figure E.7 ResNet50V2

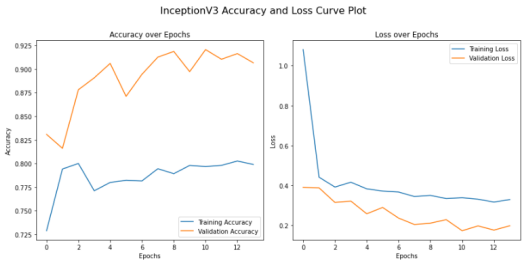


Figure E.8 InceptionV3

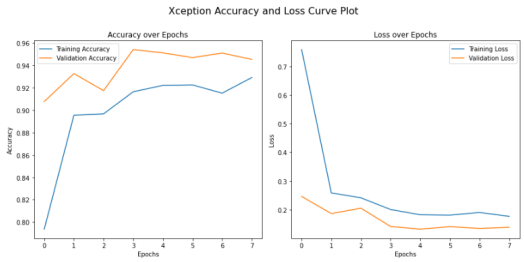


Figure E.9 Xception

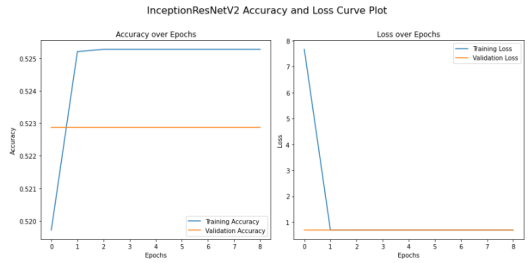


Figure E.10 InceptionResNetV2

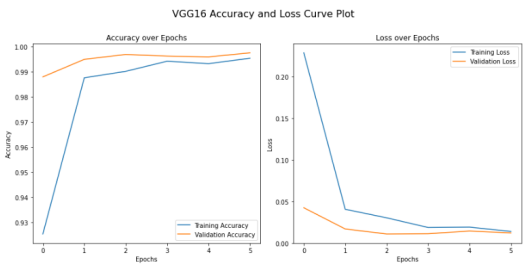


Figure E.11 VGG16

E.2 Data Augmented Models

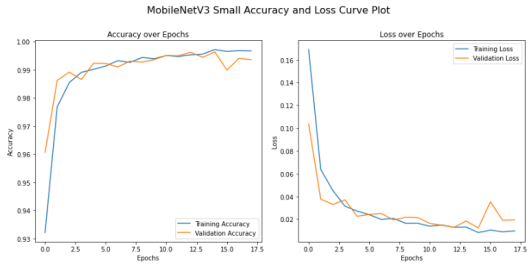


Figure E.12 MobileNetV3 Small

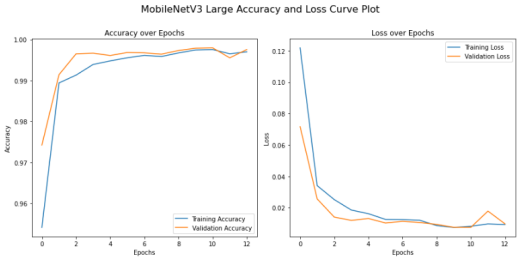


Figure E.13 MobileNetV3 Small

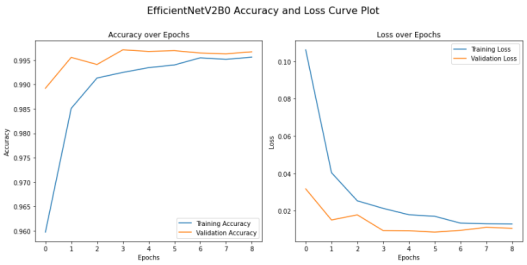


Figure E.14 EfficientNetV2B0

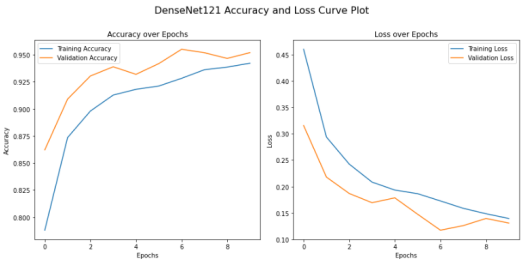


Figure E.15 DenseNet121

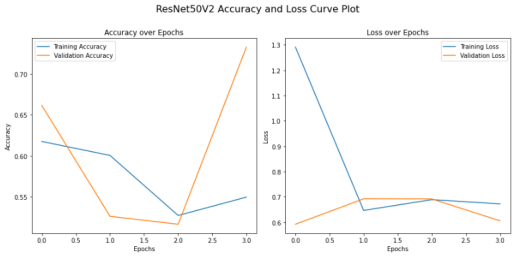


Figure E.16 ResNet50V2

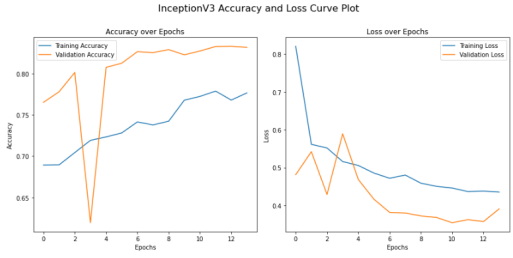


Figure E.17 InceptionV3

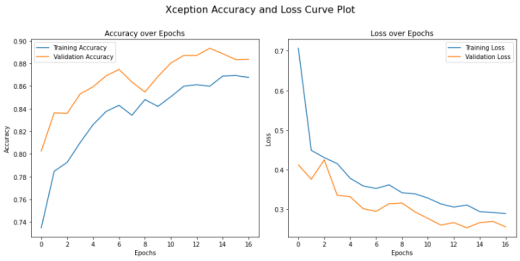


Figure E.18 Xception

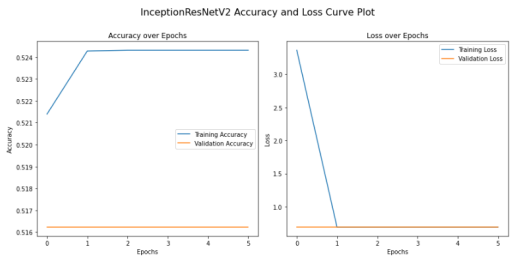


Figure E.19 InceptionResNetV2