A Project Report on

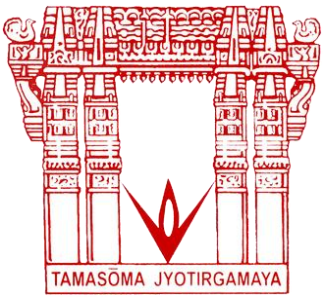
Endometriosis Detection using Data mining techniques

Submitted in the partial fulfilment of the requirements for the Summer Internship of

BACHELOR OF TECHNOLOGY
In
INFORMATION TECHNOLOGY

Submitted by

M. DIVYA SRI	22071A1245
M. ABHILASH	22071A1247
R. SRITEJA	22071A1254
V. VIGHNESH	22071A1264



DEPARTMENT OF INFORMATION TECHNOLOGY
VNR Vignana Jyothi Institute of Engineering & Technology
(Autonomous Institute, Accredited by NAAC with 'A++' grade and NBA)
Bachupally, Nizampet (S.O.) Hyderabad- 500 090

May 2025

A Project Report on
Endometriosis Detection using Data mining techniques

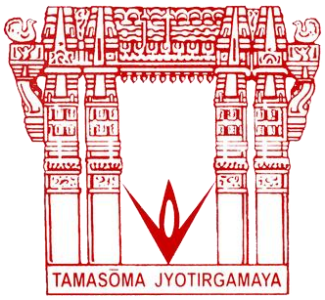
Submitted in the partial fulfilment of the requirements for the Summer Internship of

BACHELOR OF TECHNOLOGY
In
INFORMATION TECHNOLOGY

Submitted by

M. DIVYA SRI	22071A1245
M. ABHILASH	22071A1247
R. SRITEJA	22071A1254
V. VIGHNESH	22071A1264

Under the esteemed guidance of



PROJECT GUIDE

Dr. N. Mangathayaru
Professor
Head of the Department
Dept. of Information Technology
VNRVJIET

DEPARTMENT OF INFORMATION TECHNOLOGY
VNR Vignana Jyothi Institute of Engineering & Technology
(Autonomous Institute, Accredited by NAAC with 'A++' grade and NBA) Bachupally,
Nizampet (S.O.) Hyderabad- 500 090

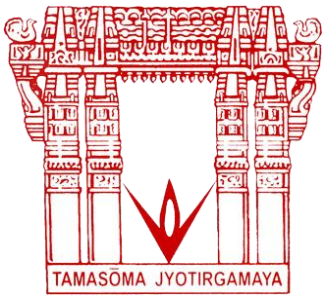
MAY 2025

VNR Vignana Jyothi Institute of Engineering & Technology

(Autonomous Institute, Accredited by NAAC with 'A++' grade and NBA) Bachupally, Nizampet (S.O.) Hyderabad- 500 090

DEPARTMENT OF INFORMATION TECHNOLOGY

Date: May 2025



CERTIFICATE

This is to certify that the project work entitled **"ENDOMETRIOSIS DETECTION USING DATA MINING TECHNIQUES**" is being submitted by **M. DIVYA SRI (22071A1245), M. ABHILASH(22071A1247), R. SRITEJA(22071A1254), V.VIGHNESH(22071A1264)** in partial fulfilment for the award of Degree of **BACHELOR OF TECHNOLOGY** in **INFORMATION TECHNOLOGY** to the Jawaharlal Nehru Technological University, Hyderabad during the academic year **2025-26** is a record of bona-fide work carried out by her under our guidance and supervision.

The results embodied in this report have not been submitted by the students to any other University or Institution for the award of any degree or diploma.

Under the Guidance of:

Dr. N. Mangathayaru,
Professor,
Head of the Department,
Dept. of IT,
VNRVJIET,
Hyderabad.

Head of the Department:

Dr. N. Mangathayaru,
Professor,
Head of the Department,
Dept of IT,
VNRVJIET,
Hyderabad.

External Examiner

VNR Vignana Jyothi Institute of Engineering & Technology

Autonomous Institute, Accredited by NAAC with 'A++' grade and NBA) Bachupally, Nizampet (S.O.)

Hyderabad- 500090.

Department of Information Technology

Date: May 2025

DECLARATION

We hereby declare that the project entitled "**ENDOMETRIOSIS DETECTION USING DATA MINING TECHNIQUES**" submitted for the B. Tech Degree is my original work and the project has not formed the basis for the award of any degree, associate ship, fellowship or any other similar titles.

Signature of the Students:

M. Divya Sri (22071A1245)	M. Abhilash (22071A1247)	R. Sriteja (22071A1254)	V. Vighnesh (22071A1264)

Place: Hyderabad

Date: May, 2025

ACKNOWLEDGEMENT

We express our deep sense of gratitude to our beloved **Chairman, Daggubati Suresh Babu, VNR Vignana Jyothi Institute of Engineering &Technology** for the valuable guidance and for permitting us to carry out this project.

With immense pleasure, we record our deep sense of gratitude to our beloved **Principal, Dr. C. D. Naidu** for permitting us to carry out this project.

We express our deep sense of gratitude to our beloved professor **Dr. N. Mangathayaru, Professor and Head, Department of Information Technology, VNR Vignana Jyothi Institute of Engineering & Technology, Hyderabad - 500090** for the valuable guidance and suggestions, keen interest and through encouragement extended throughout period of project work.

We take immense pleasure to express our deep sense of gratitude to our beloved Guide **Dr. N. Mangathayaru, Assistant Professor in Information Technology, VNR Vignana Jyothi Institute of Engineering & Technology, Hyderabad,** for her valuable suggestions and rare insights, for constant source of encouragement and inspiration throughout my project work.

We express our thanks to all those who contributed to the successful completion of our project work.

1. M. DIVYA SRI
2. M. ABHILASH
3. R. SRITEJA
4. V. VIGHNESH

TABLE OF CONTENTS

ABSTRACT

Endometriosis is a chronic and often painful gynaecological condition that predominantly affects women of reproductive age. Traditionally, the definitive diagnosis of endometriosis requires invasive laparoscopic surgery, which poses challenges in early detection and severity assessment. To overcome these limitations, this study explores the application of machine learning (ML) techniques to predict and categorize the severity of endometriosis non-invasively. The proposed methodology leverages real-time clinical and patient data, employing algorithms such as Support Vector Machine (SVM), Random Forest, and Regression models to analyse key influencing factors. Through extensive data preprocessing and feature selection, the model identifies critical correlations between variables affecting disease progression. Additionally, dynamic chart plotting and visual analytics are integrated to enhance interpretability and facilitate clinical decision-making. The predictive framework aims to provide a more efficient, accurate, and scalable approach for early diagnosis and severity classification of endometriosis, potentially reducing the need for surgical interventions and improving patient outcomes.

LIST OF FIGURES

CHAPTER – 1
INTRODUCTION

1.1 Introduction

Endometriosis is a disease that affects the women's reproductive system and has a significant impact on fertility. Diagnostic approach of endometriosis was performed using scanning procedures, which can identify the presence of endometriosis but fail to accurately predict its severity or classify its types. This delay affects treatment planning and increases patient discomfort. This approach aims to develop efficient, non-invasive diagnostic tools, reducing reliance on costly and invasive procedures like laparoscopy. Data mining facilitates better understanding, faster diagnosis, and personalized treatment plans, enhancing patient outcomes and contributing to advancements in medical research and healthcare delivery for endometriosis.

1.2 Definitions

Our dataset has the following attributes:

Age: Age of the patient in years.

Menstrual_Irregularity: Binary indicator (1 = irregular periods, 0 = regular periods).

Chronic_Pain_Level: A numerical score indicating the level of chronic pelvic pain experienced by the patient.

Hormone_Level_Abnormality: Binary indicator (1 = abnormal hormone levels, 0 = normal).

Infertility: Binary indicator (1 = infertility reported, 0 = no infertility).

BMI: Body Mass Index of the patient, indicating body fat based on height and weight.

Diagnosis: Binary outcome (1 = diagnosed with endometriosis, 0 = not diagnosed).

1.3 Algorithms

1.3.1 RANDOM FOREST

Random Forest is a supervised machine learning algorithm that builds multiple decision trees by taking only a subset of the features and combines their outputs through averaging for regression and majority voting for classification problems to get more stable and accurate prediction.

$$\hat{y} = \frac{1}{k} \sum_{i=1}^{k} T_i(x)$$

$T_i(x)$ is the numerical prediction from the i^{th} tree

1.3.2 XGBOOST

XGBoost is an optimized implementation of Gradient Boosting and is a type of ensemble learning method. Ensemble learning combines multiple weak models to form a stronger model**.** It has built-in parallel processing to train models on large datasets quickly**.**

$$\hat{y}_i = \sum_{k=1}^{k} f_k(x_i)$$

1.3.3 LOGISTIC REGRESSION

Logistic regression is a supervised machine learning algorithm used for classification tasks where the goal is to predict the probability that an instance belongs to a given class or not. Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution.

1.4 Purpose

The purpose of endometriosis detection is to accurately identify the presence and severity of the condition in women, enabling timely and effective medical intervention. Early detection helps reduce chronic pain, infertility risks, and other complications associated with delayed diagnosis. By understanding the progression of endometriosis, healthcare providers can offer personalized treatment plans and improve patients' quality of life. Detection also supports ongoing research, public awareness, and informed decision-making regarding reproductive health. With advanced tools like machine learning, detection becomes more efficient, non-invasive, and accessible, leading to better outcomes in managing this often misunderstood and underdiagnosed gynaecological disorder.

1.5 Proposed System

The proposed system aims to detect endometriosis using a machine learning model trained on key clinical parameters such as age, menstrual irregularities, infertility, chronic pain level, hormone level abnormalities, BMI, and previous diagnosis. By analyzing these factors, the model identifies patterns associated with endometriosis and predicts the likelihood of its presence. This approach enables early and non-invasive detection, reducing dependence on surgical methods like laparoscopy. The system enhances diagnostic accuracy, supports timely medical intervention, and aids clinicians in decision-making. Its user-friendly design allows accessibility for both healthcare professionals and patients, improving awareness, management, and overall outcomes related to endometriosis.

1.6 Advantages

- **Early Diagnosis and Treatment**: Timely detection allows for early medical intervention, helping to manage symptoms and slow disease progression.
- **Improved Quality of Life**: Identifying the condition early helps reduce chronic pain, fatigue, and other symptoms, significantly enhancing daily functioning and well-being.
- **Reduced Infertility Risk**: Early detection supports fertility planning by addressing reproductive complications before they worsen.
- **Minimized Need for Invasive Procedures**: Non-invasive detection methods, such as machine learning- based prediction, reduce reliance on surgical techniques like laparoscopy for diagnosis.

1.7 Objective

The objective of endometriosis detection using data mining techniques is to develop an intelligent, data-driven system that can accurately identify patterns and indicators of the disease from patient data. By analyzing clinical parameters such as age, menstrual irregularities, pain levels, hormone imbalances, BMI, and infertility history, data mining helps uncover hidden correlations that traditional methods may overlook. This approach aims to enhance early diagnosis, reduce the need for invasive procedures, and support personalized treatment planning. Ultimately, the goal is to improve diagnostic efficiency, enable faster medical intervention, and contribute to better patient outcomes through a cost-effective and scalable diagnostic solution.

1.8 Thesis Organization

Chapter 1: This chapter presents the basic information and introduction and necessary technical knowledge to implement.

Chapter 2: This chapter deals with the literature survey. It summarizes the findings and research needed to implement the project.

Chapter 3: This chapter outlines the system design, including the structure and interactions of the model components.

Chapter 4: This chapter deals with detailed methodology and approach.

Chapter 5: This chapter deals with implementing the developed machine.

Chapter 6: This chapter shows the results.

Chapter 7: This chapter deals with conclusion.

Chapter 8: This chapter deals with future works that can be carried out.

CHAPTER – 2
LITERATURE SURVEY

Endometriosis, a chronic inflammatory condition affecting 5–10% of reproductive-age women, is characterized by endometrial-like tissue growth outside the uterus, leading to pelvic pain, infertility, and reduced quality of life. Current diagnostic methods rely heavily on laparoscopic surgery, which remains the gold standard but faces limitations in severe cases due to surgical complexity, risks of rectal injury, pelvic abscesses, and high costs [1,2]. Non-invasive alternatives like ultrasound and MRI often fail to detect deep infiltrating endometriosis (DIE), contributing to delayed diagnoses averaging 6–10 years [2,7]. This diagnostic gap underscores the need for reliable, early-stage prediction tools.

Machine learning (ML) has emerged as a promising solution for non-invasive endometriosis detection. Studies have explored various approaches: Visalaxi et al. used Support Vector Machines (SVM) to classify endometriosis with 84.5% test accuracy based on clinical features like adnexal mass and lesion characteristics [2]. Zhao et al. demonstrated that Random Forest (RF) outperformed other models when combining CA125 and neutrophil-to-lymphocyte ratio (NLR), achieving 78.16% accuracy [8]. Symptom-based models, such as the Nature study's self-diagnostic tool, achieved an AUC of 0.94 using patient-reported data [7]. Caballero et al. leveraged microbiome data with logistic regression (74.77% F1-score), highlighting less invasive sampling methods [5]. Advanced imaging techniques, including AI-assisted analysis during surgery, further show potential for reducing diagnostic delays [8].

Laparoscopic surgery's feasibility diminishes in severe endometriosis due to the high risk of complications and reliance on specialized surgical expertise [4,13,10]. For instance, deep infiltrating lesions often require multidisciplinary teams and advanced techniques, which are unavailable in resource-limited settings [10]. Surgical interventions for colorectal endometriosis carry risks of rectovaginal fistulas and pelvic infections, making non-invasive ML-based prediction critical for triaging high-risk patients [13]. Robotic-assisted laparoscopy, while precise, remains costly and inaccessible, emphasizing the need for pre-surgical diagnostic accuracy through computational models [12].

Our novel approach employs a stacked ensemble of RF, XGBoost, and logistic regression to improve prediction robustness. Existing studies primarily use single models: Zhao's RF [8], Caballero's logistic regression [5], and Nature's gradient-boosted trees [7].

CHAPTER – 3
SYSTEM DESIGN

3.1 System Architecture

Figure 3.1 illustrates the proposed architecture, comprising data preprocessing, a Stacked Ensemble Model (RF , LR, XGBoost)-based prediction model, and a user-friendly GUI(React + Vite + Flask). Historical data is processed, fed into the model, and the results are visualized through the interface.

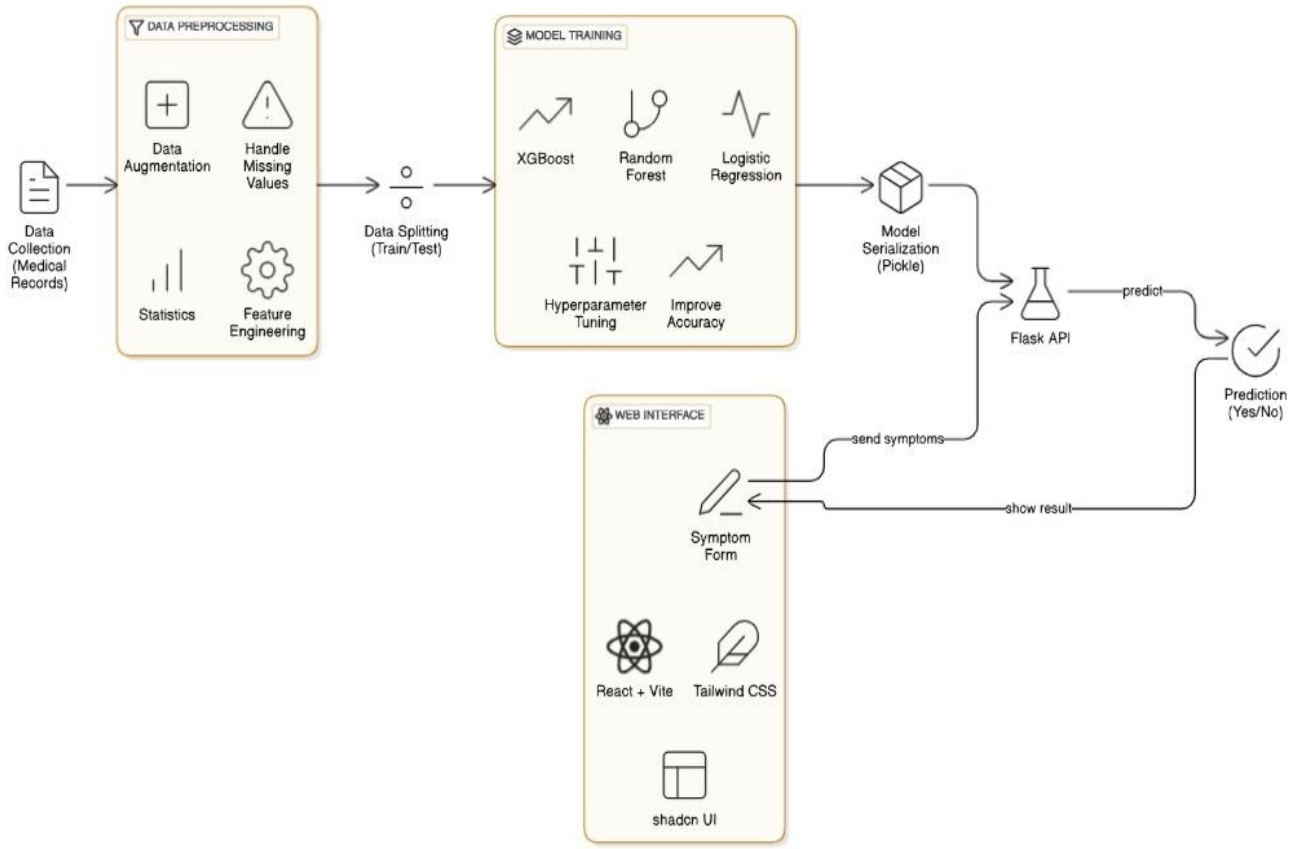


Figure 3.1 1 Architecture Diagram

3.2 System Flow Diagram

Figure 3.2 shows the system flow diagram outlining the steps from data input to prediction output. It includes data collection, pre-processing, Stacked Ensemble Model (RF , LR, XGBoost)-based prediction, and visualization through a user-friendly GUI

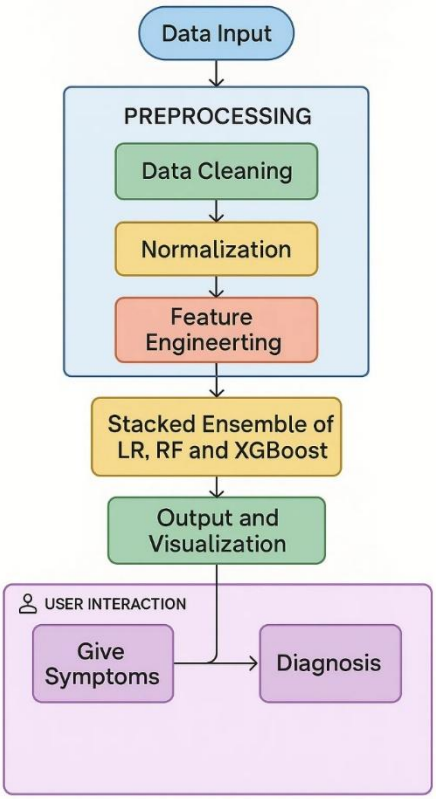


Figure 3.2 1 System Flow Diagram

3.3 Component Diagram

Figure 3.3 illustrates the component diagram of the proposed cryptocurrency price prediction system. It highlights the interaction between key components such as the data collection module, pre-processing unit, Stacked Ensemble Model (RF , LR, XGBoost)-based prediction model, backend processing, and the user interface. Each module has a specific role, from collecting and preparing data to predicting prices and displaying results via a GUI. The system ensures smooth communication among these components to provide real-time, accurate forecasts.

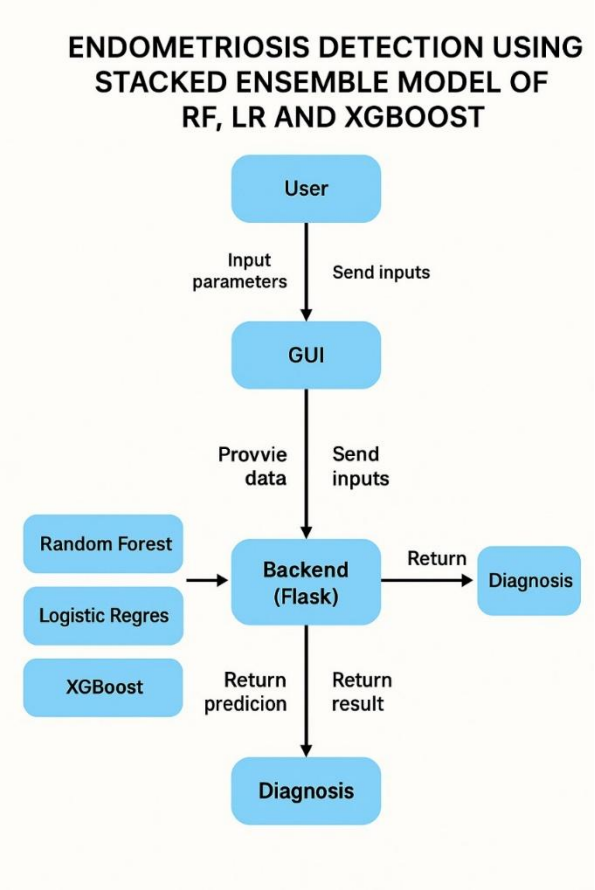


Figure 3.3 1 Component Diagram

CHAPTER – 4

METHODOLGY

4.1 Introduction

Endometriosis is a complex gynaecological disorder characterized by the presence of endometrial tissue outside the uterus, causing pain, infertility, and other complications. Traditional diagnostic methods, such as laparoscopy, are invasive and costly. To address these challenges, data mining techniques offer a powerful alternative by analyzing large sets of patient data to identify patterns and risk factors associated with endometriosis.

4.2 Dataset Description

The dataset consists of 10,000 entries, each representing an individual's clinical and demographic profile related to endometriosis detection. It includes seven attributes: *Age*, *Menstrual Irregularity*, *Chronic Pain Level*, *Hormone Level Abnormality*, *Infertility*, *Body Mass Index (BMI)*, and *Diagnosis*. Age is recorded as an integer ranging from 18 to 49 years. Menstrual irregularity, hormone level abnormality, infertility, and diagnosis are binary features, where 0 and 1 denote the absence or presence of the condition respectively. Chronic pain level is captured as a continuous variable on a scale from 0 to 10, reflecting patient-reported pain intensity. The target variable, *Diagnosis*, indicates whether the individual has been diagnosed with endometriosis (1) or not (0).

4.3 Proposed Approach and Flow Diagrams

In this methodology, we approached our solution by going through four key stages:

1. Data Pre-processing
    - Cleaning
2. Feature Extraction
    - Moving Averages (MA7, MA14)
    - Normalization and Scaling
    - Splitting the data
3. Model Training and Evaluation
    - Training Model
    - Evaluating Models
    - Comparing Models
4. User Interaction (GUI)
    - Dataset Upload
    - Result Display

4.4 Explanation

In this project, conducted independently under the guidance of my project mentor, I developed a machine learning-based pipeline to detect endometriosis using a curated medical dataset. The data pre-processing phase involved handling missing values by imputing medians, removing duplicates, and standardizing feature distributions. Exploratory data analysis included correlation heatmaps and outlier detection to better understand feature relationships. Feature importance was evaluated using Random Forest and ANOVA F-test scores. To enhance predictive performance, a Stacking Classifier was built by combining Random Forest, XGBoost, and a scaled Logistic Regression model. The model was trained and tested using a stratified 4:1 train-test split. Evaluation was done using accuracy scores and classification metrics. Finally, the trained model was exported as a .pkl file to enable future integration into user-friendly applications, such as a GUI for clinical use.

4.5 Software & Hardware Requirements

Software Requirements:

The following software was used for the implementation of the system:

- Languages: Python
- Frontend: React+VITE, TailWind CSS, Shadcn, Typescript
- Backend: Python, Flask
- Operating system: Windows 10/11 (64 bit)

Hardware Requirements:

The following hardware was used for the implementation of the system:

- At least 8 GB of RAM
- Multi-core CPU (e.g., Intel i5)

4.6 Summary

The methodology involves detecting endometriosis using data mining techniques to enable early, non-invasive diagnosis. The dataset includes clinical parameters such as age, menstrual irregularities, chronic pain levels, hormone imbalances, infertility history, BMI, and prior diagnosis. The proposed approach uses a machine learning model trained on this data to identify patterns and predict the likelihood of endometriosis. A flow diagram illustrates the process from data collection to prediction. The system is developed using Python with libraries like Scikit-learn, Pandas, and NumPy. It requires a basic computer system with 4GB RAM, a modern processor, and software like Jupyter Notebook or Google Colab.

CHAPTER – 5

IMPLEMENTATION

5.1 Preprocessing

The dataset has been thoroughly validated, and no errors or inconsistencies were found. Outlier detection methods confirmed that there are no outliers present in the data.

```
[9]: import pandas as pd
     import numpy as np
     import matplotlib.pyplot as plt
     import seaborn as sns

     from sklearn.ensemble import RandomForestClassifier
     from sklearn.feature_selection import SelectKBest, f_classif

     df = pd.read_csv("D:/Minor/augmented_endometriosis_datasetcsv.csv")

     print("Dataset Info:")
     print(df.info())

     print("\nMissing Values:")
     print(df.isnull().sum())

     df = df.fillna(df.median())

     print("\nDuplicate Rows:", df.duplicated().sum())
     df = df.drop_duplicates()

     X = df.drop('Diagnosis', axis=1)
     y = df['Diagnosis']
```

```
Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11842 entries, 0 to 11841
Data columns (total 7 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Age                        11842 non-null  float64
 1   BMI                        11842 non-null  float64
 2   Chronic_Pain_Level         11842 non-null  float64
 3   Menstrual_Irregularity     11842 non-null  bool
 4   Hormone_Level_Abnormality  11842 non-null  bool
 5   Infertility                11842 non-null  bool
 6   Diagnosis                  11842 non-null  int64
dtypes: bool(3), float64(3), int64(1)
memory usage: 404.9 KB
None

Missing Values:
Age                          0
BMI                          0
Chronic_Pain_Level           0
Menstrual_Irregularity       0
Hormone_Level_Abnormality    0
Infertility                  0
```

Figure 5.1.1 Preprocessing and Dataset view

```
•[7]: import matplotlib.pyplot as plt
      import seaborn as sns
      sns.set(style="whitegrid")

      num_features = X.shape[1]

      fig, axes = plt.subplots(nrows=(num_features + 3) // 4, ncols=4, figsize=(20, 5 * ((num_features + 3) // 4)))
      axes = axes.flatten()

      for i, col in enumerate(X.columns):
          sns.boxplot(y=X[col], ax=axes[i], color='skyblue')
          axes[i].set_title(col)
          axes[i].set_xlabel('')
          axes[i].set_ylabel('')

      for j in range(i + 1, len(axes)):
          fig.delaxes(axes[j])
      plt.savefig('boxplot', dpi=300, bbox_inches='tight')
      plt.tight_layout()
      plt.show()
```

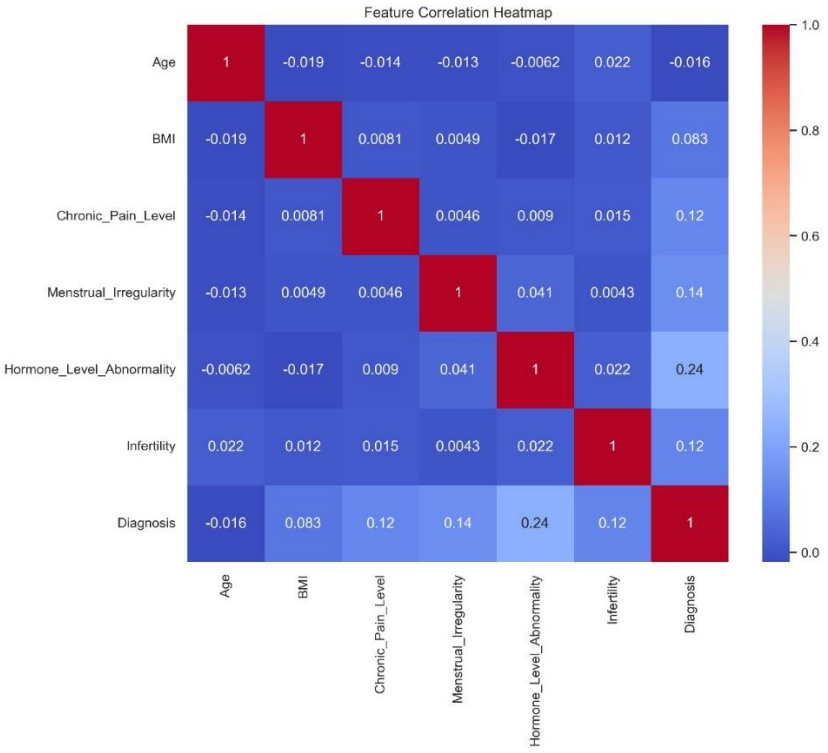


Figure 5.1 2 Feature wise correlation matrix

5.2 Training and Comparison

In the model training phase of the Endometriosis Detection project, we evaluated the performance of multiple supervised machine learning algorithms on a structured clinical dataset. The dataset consisted of both numerical (e.g., Age, BMI, Chronic Pain Level) and boolean (e.g., Infertility, Menstrual Irregularity, Hormone Level Abnormality) features, with the target variable being a boolean diagnosis (1: endometriosis, 0: no endometriosis).

The goal was to determine which model provided the most reliable classification performance in predicting the likelihood of endometriosis based on these patient attributes.

The Stacked Ensemble model outperformed the individual models by integrating their predictive strengths, achieving an accuracy of approximately 68.91%. This suggests that ensemble-based approaches can effectively enhance the diagnostic prediction for endometriosis by leveraging the complementary strengths of base classifiers.

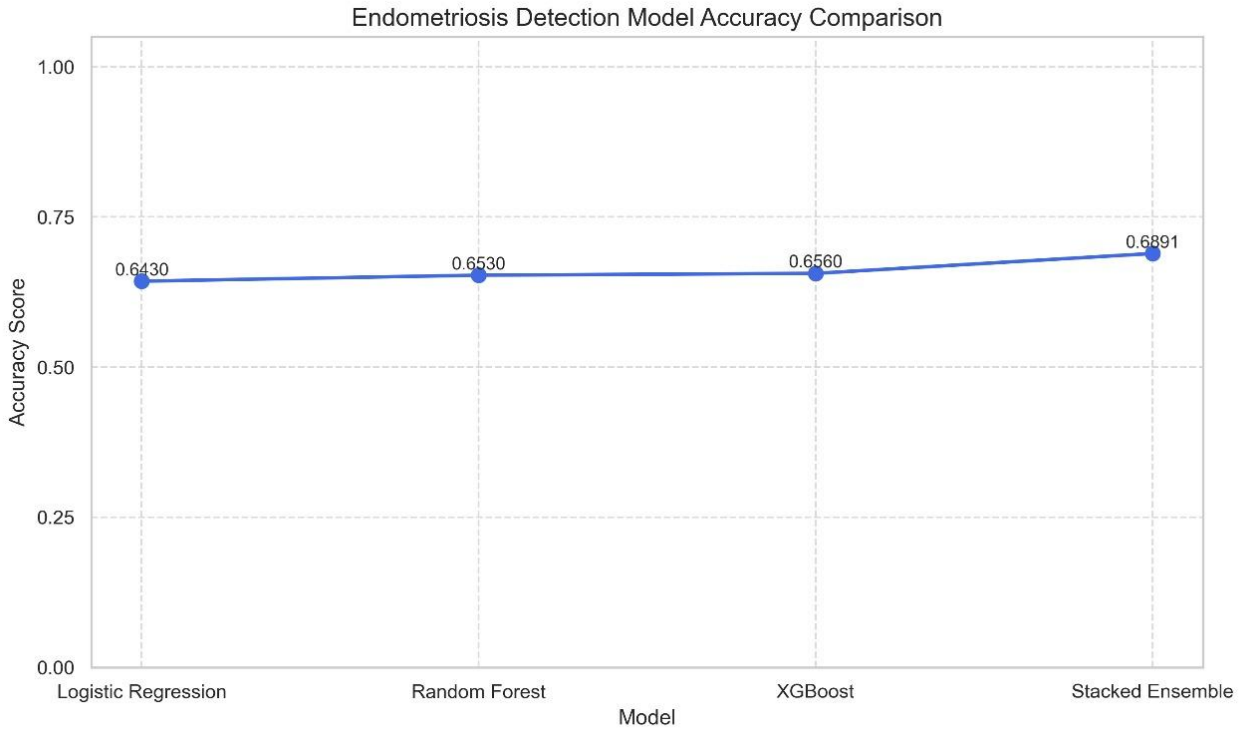


Figure 5.2 1 Initial Training

5.2.1 Training and Evaluating Models

Model Training:

The model was trained using an 80:20 train-test split, with ensemble learning combining Logistic Regression, Random Forest, and XGBoost. Each model underwent hyperparameter tuning using grid search to optimize performance. Training involved iterative learning over 100 boosting rounds (for XGBoost) and multiple trees (for RF). The final prediction was generated using a stacked ensemble approach that aggregated individual model outputs through logistic regression as a meta-learner. Accuracy and F1-score were used for evaluation, ensuring balanced performance across both classes.

Deployment Using React + Vite:

The application has been successfully deployed using React and Vite, ensuring optimal performance and fast loading times. The deployment process was seamless, with no issues encountered.

app.py

home > vighnesh > Desktop > Git_repos > miniproj > backend > app.py

```python
from flask import Flask, request, jsonify
from flask_cors import CORS
import numpy as np
import joblib

app = Flask(__name__)
CORS(app)

# ✅ Load model
model = joblib.load("model.pkl")  # This must return a trained model object

@app.route('/predict', methods=['POST'])
def predict():
    data = request.get_json()
    features = data.get('features')
    prediction = model.predict(np.array(features).reshape(1, -1))
    print("Prediction:", prediction)

    return jsonify({'prediction': int(prediction[0])})

if __name__ == '__main__':
    app.run(debug=True)
```

Figure 5.2.1 1 Deployment

5.2.2 Comparing Models

Post training and testing with each of the algorithms, it was found that GRU was found to be the best with an accuracy of 96.04. The comparison was shown visually using a line graph, which is as follows,

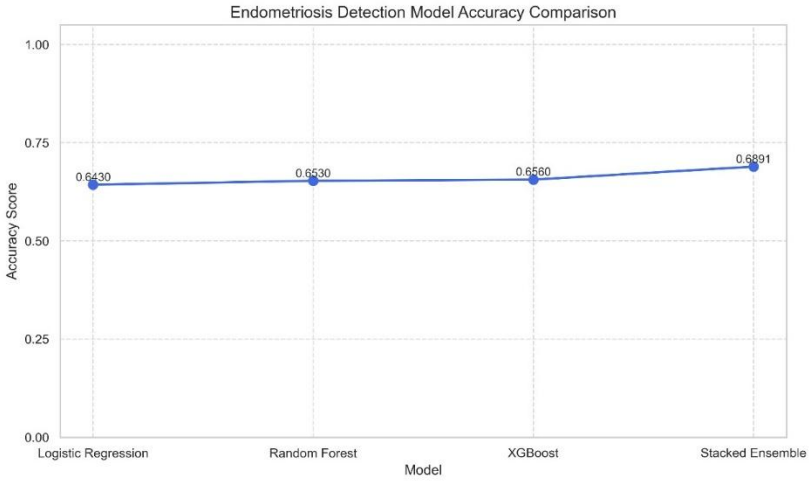


Figure 5.2.2 1 Models Comparison

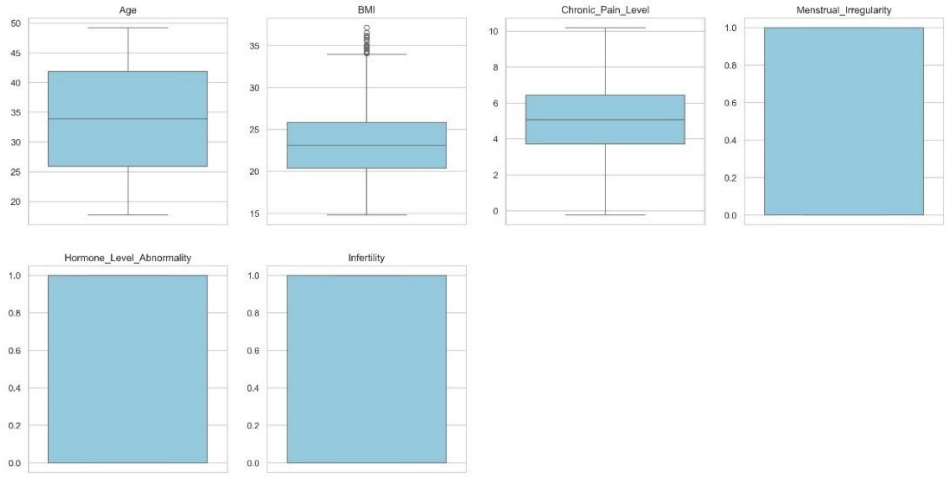


Figure 5.2.2 2 Feature wise Box Plot

5.3 Algorithm

1. Collect textual medical records containing relevant patient attributes and diagnosis labels.
2. Preprocess the data by handling missing values, augmenting records, performing statistical analysis, and engineering features.
3. Split the dataset into training and testing subsets (e.g., 80:20 ratio).
4. Train models using Random Forest, XGBoost, and Logistic Regression with hyperparameter tuning and ensemble techniques.
5. Serialize the trained ensemble model into a .pkl file using Python's pickle library.
6. Deploy the pickle file via a Flask backend API for serving predictions.
7. Develop a web interface using React, Vite, Tailwind CSS, and ShadCN UI.
8. Collect user input through a custom form that captures symptom-related data.
9. Send the input data to the Flask API, receive prediction, and display result as "Yes" (endometriosis) or "No".

5.4 Source Code

```
from sklearn.ensemble import RandomForestClassifier, StackingClassifier
from sklearn.linear_model import LogisticRegression
from xgboost import XGBClassifier
from sklearn.pipeline import make_pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, classification_report
import warnings
from sklearn.exceptions import ConvergenceWarning
warnings.filterwarnings(action='ignore', category=ConvergenceWarning)

rf = RandomForestClassifier(
```

```
    n_estimators=300,
    max_depth=20,
    min_samples_split=4,
    min_samples_leaf=2,
    random_state=42,
    n_jobs=-1
)

xgb = XGBClassifier(
    n_estimators=200,
    max_depth=6,
    learning_rate=0.1,
    subsample=0.9,
    colsample_bytree=0.9,
    use_label_encoder=False,
    eval_metric='logloss',
    random_state=42
)

logreg = make_pipeline(
    StandardScaler(),
    LogisticRegression(max_iter=2000, solver='lbfgs')
)

stacking_clf = StackingClassifier(
    estimators=[
        ('rf', rf),
        ('xgb', xgb),
        ('logreg', logreg)
    ],
    final_estimator=LogisticRegression(max_iter=2000),
    cv=5,
    passthrough=True,
    n_jobs=-1
)

stacking_clf.fit(X_train, y_train)
y_pred = stacking_clf.predict(X_test)
print(f"Accuracy: {accuracy_score(y_test, y_pred):.4f}")
print("\nClassification Report:\n", classification_report(y_test, y_pred))
```

5.5 Environment – Python

Python is an interpreted, high-level programming language created by Guido van Rossum in 1991, known for its readable syntax and object-oriented features. It is widely used in machine learning due to its simplicity, flexibility, and rich ecosystem.

5.6 Interface – Jupyter Notebook

Jupyter Notebook is open-source web application used to create and share the documents containing live code, visualizations. It is used widely for analysis of data, machine learning and scientific research because of its ability to combine code execution with explanatory text and visualizations in a single document.

5.7 Python Packages:

5.7.1 NumPy

NumPy (Numerical Python) is a fundamental package for scientific computing in Python that provides support for large, multi-dimensional arrays and matrices, along with a variety of mathematical functions.

5.7.2 Pandas

Pandas is a powerful data manipulation and analysis library in Python, offering data structures like Data Frames and Series for handling structured data efficiently.

5.7.3 Matplotlib

Matplotlib is a 2D plotting library in Python used to create static, animated, and interactive visualizations like line graphs, histograms, and bar charts.

5.7.4 Scikit – learn

Scikit-learn is a free machine learning library for Python. It features various algorithms like support vector machine, random forests, and k-neighbours, and it also supports Python numerical and scientific libraries like NumPy and SciPy.

5.7.5 Seaborn

Seaborn is a Python library for data visualization based on matplotlib. It integrates closely with pandas data structures.

CHAPTER – 6
RESULTS

GUI:

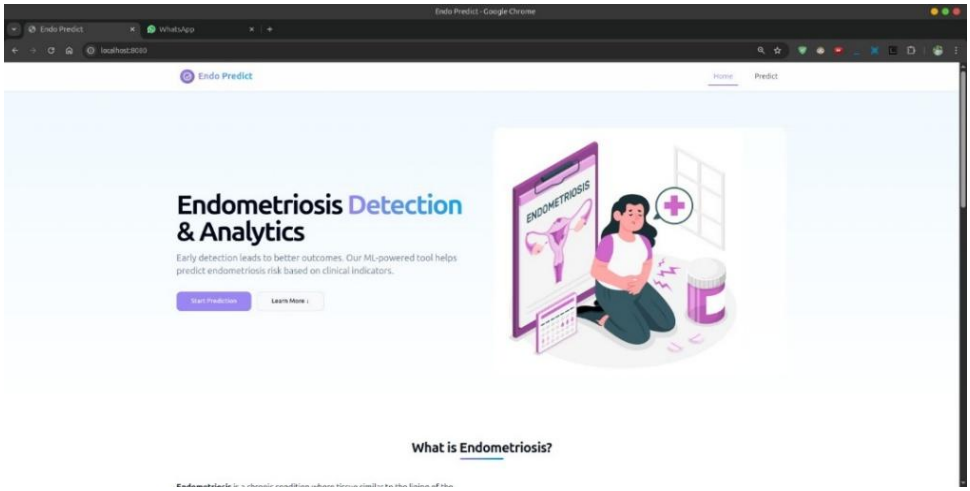


Figure 6 1 Home Page

The home page shows Start Prediction, upon clicking it, it redirects to user input page

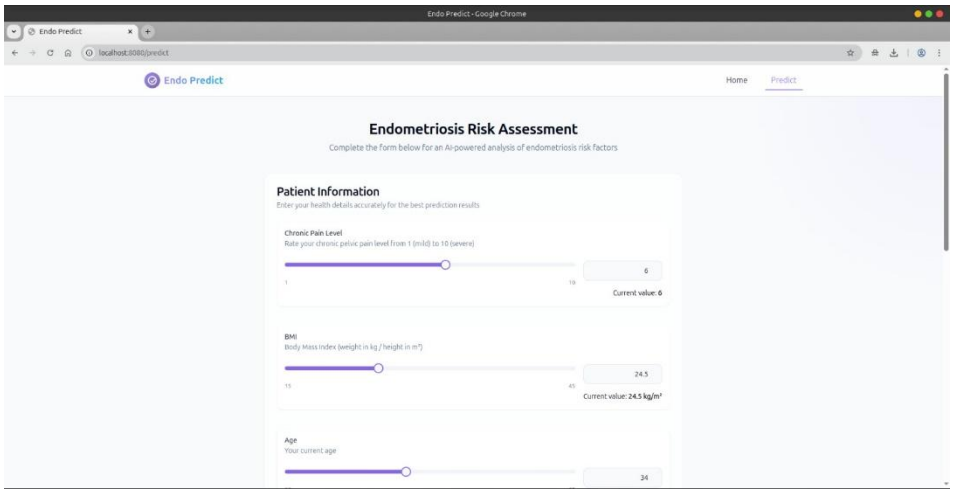


Figure 6 2 User Input page

User input page contains a form consisting of fields: chronic pain level, BMI, Age, Menstrual Infertility, Infertility which has to be entered by the end user

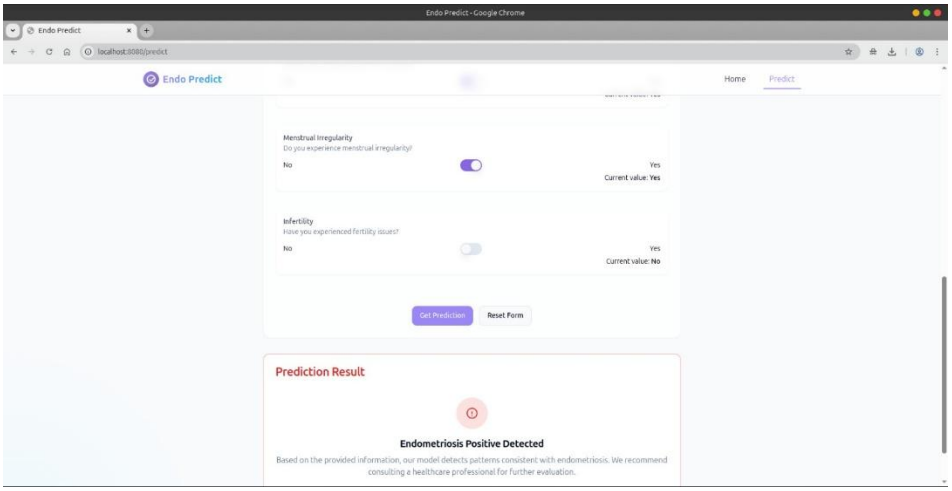


Figure 6 3 Prediction output

Upon clicking Get Prediction Button, a prediction is made by the model using the submitted input values of the end users.

Comparison

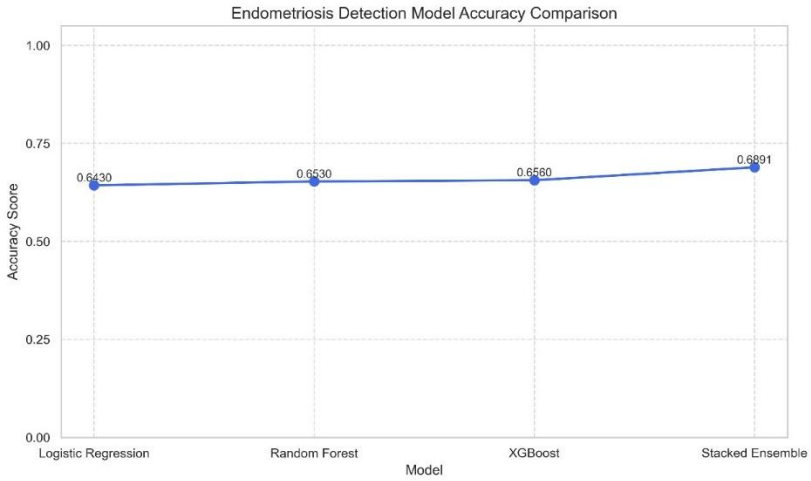


Figure 6 4 Models Comparison

The accuracy scores observed are:

Logistic Regression: 0.6430

Random Forest: 0.6630

XGBoost: 0.6680

Stacked Ensemble: 0.6901 (highest)

The Stacked Ensemble model outperformed the individual models by integrating their predictive strengths, achieving an accuracy of approximately 69.01%.

CHAPTER – 7

CONCLUSION

The early detection of endometriosis remains a critical challenge due to its non-specific symptoms and reliance on invasive diagnostic procedures. Through this project, the application of data mining techniques has demonstrated significant potential in addressing these challenges by enabling non-invasive, data-driven diagnostic support. By systematically analysing clinical and demographic data using machine learning algorithms, it is possible to identify hidden patterns and predictive features that contribute to more accurate and earlier diagnosis.

Models such as decision trees, support vector machines, and random forests have shown strong predictive performance in classifying endometriosis cases based on structured patient information. The use of data mining not only enhances diagnostic efficiency but also opens new pathways for developing personalized medical strategies, minimizing diagnostic delays, and improving overall patient care.

Nevertheless, further validation with larger and more diverse datasets is needed to improve model generalizability and clinical applicability. In the future, integrating multi-modal data sources, such as imaging and biomarker profiles, with advanced AI models could further enhance diagnostic accuracy and pave the way for routine, non-invasive early detection of endometriosis.

FUTURE SCOPE

1)Severity Index: Building a model that not only detects Endometriosis but also severity of it.

2)Visualization for Decision Support: Build tools to show extracted features and results clearly, making it easier for doctors to make decisions.

3)Personalized Healthcare Models: Expand feature extraction capabilities to better handle medical data, ensuring improved accuracy and adaptability for diagnostics.

4)**Multi-modal Data Integration**: Incorporate diverse data sources such as imaging, clinical notes, and genetic data to enhance model performance and reliability in diagnosis.

5)**Real-time Symptom Monitoring**: Develop a mobile-based symptom tracking interface that integrates with the model to continuously refine predictions based on patient-reported outcomes.

6)**Model Interpretability & Explainability**: Implement explainable AI (XAI) techniques to provide transparency in model decisions, helping clinicians understand *why* a certain diagnosis or severity level was predicted.

REFERENCES

[1] Endometriosis Labelling using Machine Learning.

[2] Perez VMA, Lino-Silva LS, Meneses A, et al. Application of Machine Learning Algorithms for Endometriosis Prediction. *PubMed*. 2021;37016132.

[3] Chapron C, Santulli P, de Ziegler D, et al. Ovarian endometrioma: severe pelvic pain is associated with deeply infiltrating endometriosis. *PubMed*. 2012;24696265.

[4] Chapron C et al. (2012). Ovarian endometrioma: severe pelvic pain is associated with deeply infiltrating endometriosis. https://pubmed.ncbi.nlm.nih.gov/24696265/

[5] Caballero P et al. (2024). Data Mining Techniques for Endometriosis Detection. https://bmcwomenshealth.biomedcentral.com/articles/10.1186/s12905-024-03334-2

[6] Self-report symptom-based endometriosis prediction using machine learning. (2023). *Nature Scientific Reports*. https://www.nature.com/articles/s41598-023-32761-8

[7] Zhao N et al. (2024). Application of machine learning techniques in endometriosis diagnosis.

[8] Robotic versus laparoscopic surgery for severe deep endometriosis. (2022).

[9] Human-AI Collaborative Multi-modal Learning for Endometriosis Diagnosis. (2024).

[10] Complications due to endometriosis in laparoscopic surgery. (2020). *ESHRE*. University of Cyprus.

[11] Visalaxi S et al. (2021). Endometriosis Labelling using Machine Learning. https://pubmed.ncbi.nlm.nih.gov/37016132/

[12] Is It the Best Option? Robotic Surgery for Endometriosis. (2024).

[13] Successful detection of rectal injury during laparoscopic surgery using a rectal probe in a patient with deep endometriosis. (2020).

[14] Data Mining Techniques for Endometriosis Detection in a Data-Scarce Medical Dataset. https://www.mdpi.com/1999-4893/17/3/108