

## A novel machine learning-based proposal for early prediction of endometriosis disease

Elena Enamorado-Díaz<sup>a,\*</sup>, Leticia Morales-Trujillo<sup>a,b,2</sup>, Julián-Alberto García-García<sup>a,3</sup>, Ana T. Marcos<sup>c,d,e,f,4</sup>, José Navarro-Pando<sup>c,d,e,f,5</sup>, María-José Escalona-Cuaresma<sup>a,6</sup>

<sup>a</sup> ES3 Group (Engineering and Science for Software Systems Group), University of Seville, Avenida Reina Mercedes, s/n., 41012 Seville, Spain

<sup>b</sup> G7innovation Company, Calle Radio Sevilla, 41001 Sevilla, Spain

<sup>c</sup> Cátedra de Reproducción y Genética Humana del Instituto para el Estudio de la Biología de la Reproducción Humana (INEBIR), Seville, Spain

<sup>d</sup> Universidad Europea del Atlántico (UNEATLANTICO), Santander, Spain

<sup>e</sup> Fundación Universitaria Iberoamericana (FUNIBER), Seville, Spain

<sup>f</sup> San Juan de Dios Hospital, Seville, Spain

### ARTICLE INFO

#### Keywords:

Endometriosis  
Predictive model  
Machine learning algorithm  
Clinical Decision Support System

### ABSTRACT

**Background:** Endometriosis is one of the causes of female infertility, with some studies estimating its prevalence at around 10 % of reproductive-age women worldwide and between 30 and 50 % in symptomatic women. However, its diagnosis is complex and often delayed, highlighting the need for more accessible and accurate diagnostic methods. The difficulty lies in its diverse etiology and the variability of symptoms among those affected.

**Methods:** This study proposes a predictive model based on supervised machine learning for the early identification of endometriosis, providing support for decision-making by healthcare professionals. For this purpose, an anonymised dataset of 5,143 female patients diagnosed with endometriosis at the private fertility clinic Inebir was used. The model integrates clinical records and genetic analysis through supervised machine learning algorithms, focusing on clinical variables and pathogenic and potentially pathogenic genetic variants.

**Results:** The developed predictive model achieves high accuracy in identifying the presence of endometriosis, highlighting the importance of combining clinical and genetic data in diagnosis. The integration of this data into the DELFOS platform, a clinical decision support system, demonstrates the utility of machine learning in improving the diagnosis of endometriosis.

**Conclusions:** The findings underscore the potential of clinical and genetic factors in the early diagnosis of endometriosis using supervised machine learning algorithms. This study contributes to the classification of clinical variables that influence endometriosis, offering a valuable tool for clinicians in making therapeutic and management decisions for their female patients.

### 1. Introduction

Endometriosis stands as a puzzling and elusive medical condition, characterized by the presence of endometrial-like tissue outside the

uterine cavity. Despite its prevalence and significant impact on women's health, the intricacies surrounding endometriosis remain largely veiled. This enigmatic nature is epitomized by the lack of a definitive etiological explanation, the vast spectrum of symptoms experienced by affected

\* Corresponding author.

E-mail addresses: [eenamorado@us.es](mailto:eenamorado@us.es) (E. Enamorado-Díaz), [lmtrujillo@us.es](mailto:lmtrujillo@us.es) (L. Morales-Trujillo), [juliangg@us.es](mailto:juliangg@us.es) (J.-A. García-García), [anateresa.marcos@inebir.com](mailto:anateresa.marcos@inebir.com) (A.T. Marcos), [jose.navarro@inebir.com](mailto:jose.navarro@inebir.com) (J. Navarro-Pando), [mjescalona@us.es](mailto:mjescalona@us.es) (M.-J. Escalona-Cuaresma).

<sup>1</sup> <https://orcid.org/0009-0002-7467-3382>.

<sup>2</sup> <https://orcid.org/0000-0001-9554-1173>.

<sup>3</sup> <https://orcid.org/0000-0003-2680-1327>.

<sup>4</sup> <https://orcid.org/0000-0002-0263-7473>.

<sup>5</sup> <https://orcid.org/0000-0003-2362-2251>.

<sup>6</sup> <https://orcid.org/0000-0002-6435-1497>.

individuals, and the perplexing delay in diagnosis that many encounter (Bullon & Manuel Navarro, 2017).

The reported prevalence of endometriosis varies widely among different regions and ethnic groups, with a general consensus suggesting that approximately 10 % (190 millions) of reproductive age women and girls may be affected (Organization, 2023). Endometriosis stands as a significant contributor to infertility, with estimates suggesting that 30–50 % of women experiencing infertility may have endometriosis (Macer & Taylor, 2012). The intricate interplay between endometriosis and fertility underscores the clinical relevance of understanding its prevalence.

While endometriosis can manifest at any age, it is predominantly diagnosed in individuals between the ages of 25 and 35 (Parasar et al., 2017). The cause of endometriosis remains a subject of intense investigation (Saunders & Horne, 2021). While theories such as retrograde menstruation and genetic predisposition have been proposed, a unifying etiological mechanism remains elusive. This complex etiology, compounded by the variability in symptom manifestation, contributes to the challenges faced by both patients and healthcare professionals. The chronic nature of endometriosis further exacerbates its impact, affecting not only physical health but also the psychological and social well-being of those affected (van Stein et al., 2023). The diagnostic journey for individuals with endometriosis is often prolonged, with many facing years of unexplained pain and infertility before receiving a conclusive diagnosis. The absence of non-invasive diagnostic tools further complicates this process, necessitating an invasive laparoscopic surgery for not always a definitive confirmation. This delay not only hampers timely intervention but also underscores the urgent need for more accessible and accurate diagnostic methods (Mak et al., 2022). Endometriosis is a complex and multifaceted condition with varied symptoms and manifestations.

Machine learning (ML) can assist in integrating a multitude of clinical and imaging data, helping physicians consider a wide range of factors for accurate diagnosis (Sivajohan et al., 2022). Physicians can benefit from the assistance of ML in the diagnosis of endometriosis for several reasons:

1. ML algorithms applied to medical imaging, such as ultrasound and MRI, have shown promise in improving the detection and characterization of endometriotic lesions (Maicas et al., 2021).
2. Supervised Machine Learning Models for Risk Prediction of endometriosis (Bendifallah et al., 2022).
3. ML has been applied to analyze molecular and genetic data to identify potential biomarkers associated with endometriosis (Zhou, 2021).
4. Natural Language Processing for Symptom Analysis has been utilized to analyze unstructured data from patient records (Koleck et al., 2019).
5. Expert systems and clinical decision support tools leveraging ML have been explored for aiding clinicians in the diagnosis of endometriosis (Nnoaham et al., 2012).

After analysing related works that address the prediction of endometriosis diagnosis using machine learning (cf., Section 5.2), it is possible to observe that most of them study a few clinical parameters related to diagnostic tests (laparoscopy and biopsies, mainly) and anamnesis history data based on interviews. These works present limitations for the diagnosis of this disease because they do not consider the complete Electronic Health Records (EHR) of the patient (e.g., genetic factors are not considered). Furthermore, although these studies provide valuable contributions, their predictive models are not integrated into technological platforms to facilitate their use by healthcare professionals. In this context, our proposal contributes from two perspectives for early detection in the diagnosis of endometriosis.

On the one hand, we have investigated which are the relevant clinical factors in the patient's EHR (including the discovery of genetic

factors) to obtain this diagnosis and, later, we have developed and trained a supervised predictive model based on machine learning algorithms. On the other hand, this predictive model has been integrated into a technological platform called DELFOS, which is an expert system that, through the analysis of electronic records, discovers genetic pathologies and provides real-time early alerts. This feature ensures healthcare professionals can make informed and rapid decisions during assisted reproduction treatments, which can be crucial in genetic risk situations. Furthermore, our platform distinguishes itself by its integration capabilities with various healthcare organisations, ensuring its transferability and application in a wide range of clinical contexts.

With the aim of presenting a comprehensive and methodical approach, the document is structured as follows: in Section 2, describes the DELFOS Platform, which supports the predictive model and outlines how patients genetic data are processed for the model. Section 3 of the predictive model, the dataset, predictor variables, handling of missing data, class balancing, and pseudonymization of clinical information, thereby ensuring the integrity and privacy of patient data. Section 4 presents the performance of various machine learning algorithms applied to predict endometriosis, highlighting the performance of the selected model. Section 5 delves into the discussion of the study, critically evaluating its validity and comparing it with previous works through a specific case study in the context of DELFOS. This section concludes by exploring the implications of the study and highlighting potential limitations. Ending with Section 6, the document summarizes the most important discoveries and emphasizes the feasibility of the predictive model in real clinical settings, while also looking towards future research directions. This section highlights the hopeful results obtained so far and the ongoing commitment to advancing this field of study. Finally, the document ends with an acknowledgments section, where the contribution of all collaborators is valued, and the ethical approval of the study by the competent authority is verified. This gesture reinforces the team's strong ethical commitment and collaboration.

## 2. Background

This section describes briefly the background of our research. Specifically, Section 2.1 briefly describes the DELFOS technology platform, which automatically supports the predictive model proposed in this paper. One of the particular aspects during the data preprocessing stage in building the predictive model is related to the preprocessing of the patient's genetic sequence data. Section 2.2 addresses our background in this subject in detail.

### 2.1. Overview of the DELFOS platform

The need to interconnect clinical data is essentially crucial for managing and providing the best service to patients any healthcare setting. Traditionally, there has been an emphasis on including purely clinical data, but in recent years, there has been a shift towards integrating as much relevant patient information as possible in a proper manner.

In order to incorporate genetic data into the array of data stored, managed, and analyzed for patients, the DELFOS software platform was developed.

DELFO is a service-oriented technology platform containing an expert CDSS based on supervised machine learning techniques to discover genetic pathologies and provide support to healthcare professionals. For this purpose, DELFOS is mainly composed of the next modules (c.f.,

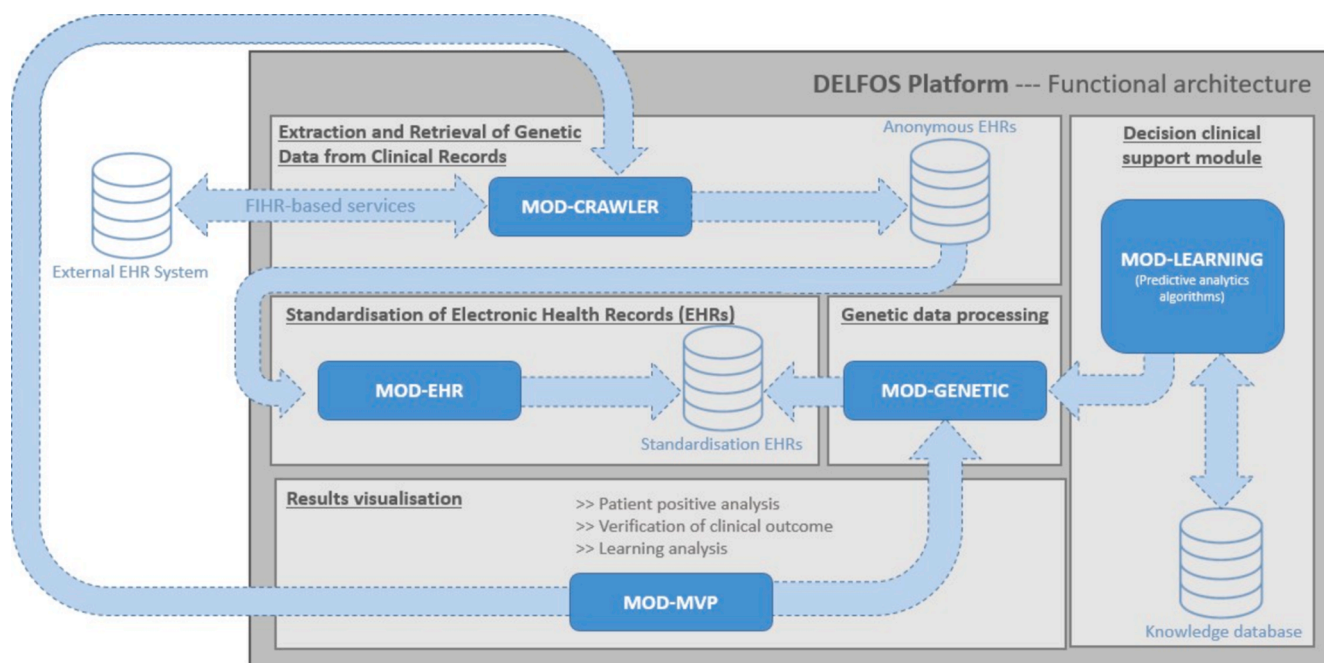


Fig. 1. Overview of the functional architecture of the DELFOS platform.

Fig. 1) whose operation is briefly explained below.

Before using this platform, it is necessary to set some configuration parameters. On the one hand, it is necessary to load the genomic variants associated with the disease to be predicted. To achieve this purpose, DELFOS is integrated with several clinical databases such as ClinVar<sup>7</sup> to automatically obtain the genomic variants that have been published in the medical scientific literature. In the context of this paper, we refer to genomic variants associated with the endometriosis disease, whose genomic variants can be consulted from (ClinVar, 2023). On the other hand, it is necessary to establish the parameters for the ideal genomic sequence of homo sapiens (Center, 2009), which will be later compared with the genetic sequence of the patient under study to identify the genetic risks associated with the preconfigured disease.

Moreover, once DELFOS is configured, first, the MOD-CRAWLER module aims to select, classify, capture and anonymise clinical records from the external EHR (Electronic Health Record) system to the MOD-EHR module, which aims to analysis and normalise clinical records within the DELFOS platform. It is important to mention that the integration between DELFOS and the external system is performed using an integration API based on FHIR (Fast Healthcare Interoperability Resource) (Ayaz et al., 2021). In the context of our research paper, DELFOS has been integrated with iMedea<sup>8</sup> (García García et al., 2022); which is the EHR system used to store clinical data of patients.

After normalizing the patient's EHRs, the MOD-GENETIC module aims to initiate the processing of the patient's enetic data (i.e., exome) related to the selected disease (specifically, endometriosis disease in the

context of this paper). This processing of the patient's genomic sequence is based on a 3-step method (cf., Section 2.2) to measure data quality, evaluate the alignment of patient genomic readings with respect to the ideal sequence, and analyze known genetic variants for the disease to be predicted (i.e., endometriosis disease). Section 2.2 describes this processing in detail.

Following this, the MOD-LEARNING module is based on the Spark technology (Salloum et al., 2016) and initiates the predictive analysis of the chosen disease. DELFOS incorporates various predictive models founded on machine learning algorithms that have been previously developed and trained. This paper, in particular, focuses on elucidating the development of the predictive model for detecting endometriosis. Ultimately, the DELFOS knowledge repository and its predictive model undergo revisions and training with the latest predictions, a process initiated upon the transmission of results to the external EHR system, after the healthcare professional has confirmed the patient's diagnosis.

## 2.2. Genetic sequence data processing

As mentioned in the previous section, DELFOS incorporates a functional module to process each patient's exome. However, it is necessary to perform a prior analysis of each exome to normalize its information before including these data in the predictive model. From a computational point of view, each exome must be provided to DELFOS in the form of compressed binary FASTQ<sup>9</sup> files.

Once the patient's FASTQ files have been acquired, DELFOS performs the following processing:

<sup>7</sup> ClinVar is a public archive with free access to reports on the relationships between human variations and phenotypes, with supporting evidence. The database includes germline and somatic variants of any size, type or genomic location. Interpretations are submitted by clinical testing laboratories, research laboratories, locus-specific databases, expert panels and practical guidelines.

<sup>8</sup> iMedea (García García et al., 2022) (innovative MEDical Engineering Assistance) is an interoperable EHR platform, which consists of more than fourteen functional modules related to clinical areas (gynaecology, andrology, assisted reproduction, cryopreservation of biological samples and diagnostic tests, among others), financial, administrative and business intelligence applied to health, among others. These modules are integrated into a single and common database of all patients' health information.

<sup>9</sup> FASTQ files are a binary format for sharing sequencing read data that stores a numeric quality score associated with each nucleotide in a sequence. It consists of four lines for each sequence record (cf., Fig. 2): (i) header line, beginning with the @ character and contains a unique tag to identify the genetic sequence; (ii) sequence line, which contains the DNA (deoxyribonucleic acid) or RNA (ribonucleic acid) sequence and its nucleotide bases in letter format (A. Adenine, G. Guanine, T. Thymine, C. Cytosine); (iii) comment line, which usually contains a "+" or may be blank; and (iv) quality line, containing a series of characters representing the quality of each corresponding nucleotide base in the sequence line.

- **Step 1. Quality assessment and processing of readings:** As mentioned earlier, the FASTQ file contains quality values (cf., Fig. 2), which must match the number of symbols in the sequence and its nucleotide bases in letter format (A. Adenine, G. Guanine, T. Thymine, C. Cytosine). Quality assessment helps ascertain the quality of samples, indicating the likelihood of incorrect base calls, or in other words, the accuracy of base calls. To enable the alignment of each nucleotide with its quality score, the numeric quality score is converted into a code, where each character represents the numerical quality of a specific nucleotide. The numerical value assigned to these characters depends on the sequencing platform that generated the reads. In this paper, we used the Phred metric (as defined in Equation (1) (Wagner et al., 2021), which is a logarithmic property related to the error probabilities of base calls (P), and it enables the assessment of the quality of nucleobase identification generated by DNA sequencers like Illumina<sup>10</sup>. This metric is also based on the Sanger sequencing method (Men et al., 2008), which assigns a quality score (Q-score) between 0 and 41 to each nucleotide, as shown in Fig. 3.

$$Q = -10\log_{10}P \quad (1)$$

The higher the number, the better the quality of the base. In this paper, we considered FASTQ file quality as desirable when Q-score  $\geq 15$ , indicating high reliability of base calls at all positions. This quantitative value was established based on the clinical experience of the authors of this paper and scientific clinical literature.

- **Step 2. Alignment of readings:** In this step, Delfos performs the alignment analysis of each FASTQ-based exome to determine the precise position of each gene read within the genome (Gateway, 2022). Later, the patient's exome is compared to the ideal human sequence to determine variations that may cause the genetic pathology to be predicted. Also, it's crucial to consider the version of the reference sequence assembly since it is regularly updated and improved, which can impact the alignment coordinates.
- **Step 3. Alignment analysis.** After aligning the reads, the search and detection of genetic variants are carried out. This step is crucial for understanding the genetic differences among the samples and their clinical relevance within our predictive model. In this context, DELFOS integrates the identification of genetic variants in patient's exome with published scientific findings that demonstrate hereditary pathologies (e.g., endometriosis disease in the context of our paper) utilizing medical public databases, such as ClinVar (as mentioned in the previous section). This allows us to correlate the patients' genetic data with previously documented information on genetic diseases, strengthening our understanding of the patients' health and their potential predisposition to certain hereditary conditions.

### 3. Materials and methods

This section outlines the methods and materials employed in conducting this research. Specifically, Section 3.1 explains the approach we followed to develop the predictive model for endometriosis. The TRIPOD-AI (Collins et al., 2024) standard has also been taken into account in structuring the content of this section. Later, Section 3.2 describes the study participants and dataset, including the inclusion and exclusion criteria for sample selection. Section 3.3 covers data preparation, identifying relevant clinical variables and addressing missing data. Section 3.4 defines the outcome variable (presence or absence of endometriosis) and predictor variables, which span demographic, clinical, and genetic categories. Section 3.5 details the model generation process, comparing various machine learning algorithms to select the

best performing model. Finally, Section 3.6 describes the method for pseudonymising clinical data that has been carried out to extract the dataset used in this research.

#### 3.1. Method for the predictive model development

While there is currently no definitive consensus on the best method for developing and validating predictive models, eral methodologies have been proposed, as mentioned in (Alonzo, 2009; Altman and Royston, 2000; Altman et al., 2009; Laupacis and Sekar, 1997); among others. Before delving into our specific proposal, it is crucial to acknowledge the diversity of approaches available in both general and clinical contexts. Among these options, we have chosen a methodology that aligns with our research needs and the specific study environment.

Drawing inspiration from these diverse proposals, Fig. 4 illustrates the standardized process we have followed in this paper to create our predictive model for endometriosis. This process is based on five stages:

- **Stage 1. Defining clinical need and research questions.** The purpose of this phase is to articulate the scientific necessity for building the predictive model and establish the research questions that will guide the model's construction. In the context of this paper, as explained in the previous section, endometriosis has become a topic of growing interest in the medical field in recent years, as it affects a significant number of women worldwide. Understanding this disease and its impact on reproductive health is essential for healthcare professionals treating patients with endometriosis because specialists in gynecology, genetics and other fields of medicine must work closely together to establish accurate diagnoses and develop appropriate treatment plans for each patient. In addition, the availability of automated tools that support clinical and genetic evidence-based decision making is critical to improving the care and quality of life for women affected by endometriosis.

In this context, the following research questions (RQ) guide this paper: «RQ1: What are the predictive variables related to the patient's clinical history that provide prediction of endometriosis?»; «RQ2: How is it possible to integrate genetic factors in the prediction of endometriosis?».

- **Stage 2. Identifying the dataset.** Although there is no consensus on minimum dataset size requirements (Altman et al., 2009); extensive datasets that mirror the characteristics of the target data population are ideal for developing the predictive model to ensure its reproducibility and generalizability. Section 3.2 describes in detail the study participants and the dataset which has been used in this research.
- **Stage 3. Managing variables.** This stage aims to determine the set of candidate clinical variables to be included in the predictive model after analyzing their prevalence and statistical significance in the clinical dataset. Section 3.3 and Section 3.4 describe in detail the preparation of the dataset and the selection of predictor variables, respectively
- **Stage 4. Generating the predictive model.** The purpose of this stage is to generate the predictive model considering the final predictor variables identified in the previous stage. To achieve this goal, we have evaluated and measured the performance of four supervised machine learning algorithms (i.e., random forest, logistic regression, decision tree classifier, and naive bayes) using the main performance metrics as described Section 4 in detail.
- **Stage 5. Validating the predictive model.** After the model has been generated, it needs to be validated using an independent dataset to assess the model's predictive potential. In this sense, this stage aims to carry out a retrospective and observational study in order to describe the degree of accuracy of the predictive model (using this independent dataset) versus the observable clinical diagnosis of the

<sup>10</sup> Illumina (Modi et al., 2021) is a genetic sequencer that conducts image analysis and base calling, resulting in binary compressed FASTQ files.

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*(((((***+))%%%+))(%%%%) .1***-+*''))**55CCF>>>>>CCCCCCC65
```

Fig. 2. Internal structure of the FASTQ binary files (simplified view).

```
Quality coding: !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ
Quality score: 01.....11.....21.....31.....41
```

Fig. 3. Character quality encoding of FASTQ files (simplified view).

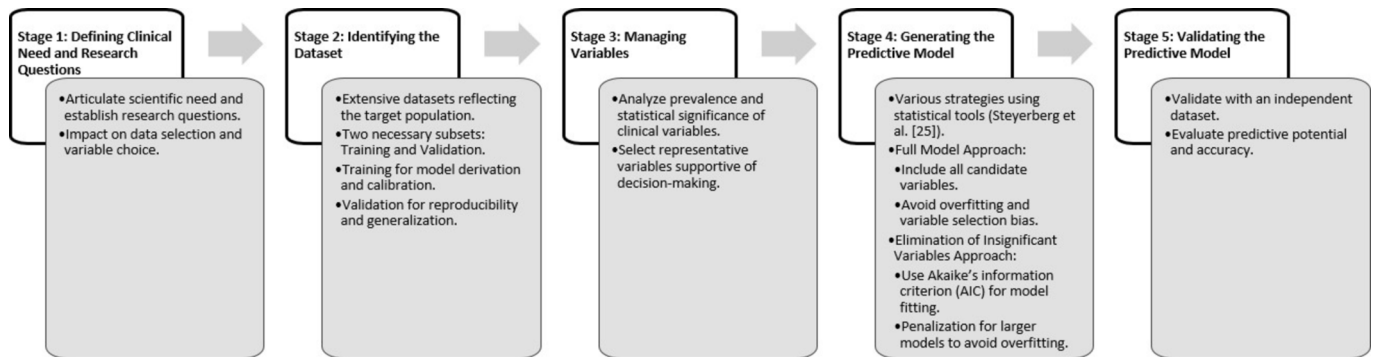


Fig. 4. Approach to Building Predictive Models.

female patient with endometriosis. In this regard, Section 5.1 evaluates in detail the results of the predictive model considering our validation dataset and clinical endometriosis values.

### 3.2. Participants and dataset

The study was carried out at the Assisted Reproduction Unit of the Spanish clinic Inebir Inc<sup>11</sup>, in which we have carried out an analytical case-control study on EHRs of female patients under follow-up at this clinic. This study was supervised by a geneticist and a gynecologist (who are experts in assisted reproduction and the endometriosis disease) and it included a study population with 5,143 female patients. Of these 5,143 patients, 270 were diagnosed with endometriosis, representing approximately 5.25 % of the total, while the remaining 94.75 % were not diagnosed with this condition.

Clinical records were obtained from iMedea, which is cloud EHR system used by Inebir to store clinical data of their patients. In this sense, the data obtained were: (a) selected according to exclusion and inclusion criteria (c.f., Section 3.3) on the basis of the judgement of healthcare professionals to avoid the inclusion of variables that are unrepresentative or meaningless for the clinical case under study; and (b) pseudonymised following the method described in Section 3.6. This data universe was divided into a training dataset and a validation dataset (cf., Section 4). The training dataset is used to derive the predictive model and calibrate its performance by evaluating the data contained in the validation dataset.

### 3.3. Data preparation

Before carrying out the study of predictor variables, first, we identify,

<sup>11</sup> Inebir is one of the leading healthcare entities in reproductive medicine in Andalusia (Spain), with more than 1200 ART cycles performed annually and more than 5500 ART cycles performed in total since 2020.

study and analyse those representative clinical data that evidence endometriosis considering the experience and judgement of healthcare professionals (who belongs to Inebir), as well as their clinical evidence obtained over the last 30 years. In this sense, clinical data related to anamnesis, genetic data and endometriosis diagnostic tests (endometrial size, uterine angle, and left and right ovarian follicles, among other aspects) are initially studied as candidate variables to be considered by the predictive model.

Once the clinical information to be considered had been established, we established criteria for selecting samples for training and validation of our predictive model. These criteria have been established considering the medical judgment and experience of the clinical investigators who have participated in this research. Specifically, we have considered female patients aged between 20 and 64 years, undergoing assisted reproductive treatments or diagnosed with infertility problems by a medical professional. In addition, we excluded female patients with age outside the specified range and with no confirmed medical diagnosis or incomplete clinical data, as well as patients with infertility-unrelated health problems that can affect the reproductive capacity.

### 3.4. Outcome and variable predictors

We considered all possible clinical diagnoses of endometriosis. This specific diagnosis is reflected through the target categorical variable named *has\_endometriosis*, whose values are detailed in Equation (2).

$$has\_endometriosis \in A = \{No, Yes\} \tag{2}$$

The original clinical data extracted from iMedea for each patient were divided into four groups of general information: (i) *patient demographic data*, which include date of birth, height, weight, ethnicity, hair and skin color; (ii) *patient anamnesis*, which includes family history, alterations, allergies, infectious diseases, surgical history, toxicities, and current medical treatments; (iii) *data associated with endometriosis diagnostic tests*, which are related to the uterine status, endometrial size,

uterine angle, uterine cavity and left and right ovarian follicles, among other aspects; and (iii) *genetic data*, which refers to the genetic sequence of the patient’s genome stored as a binary FASTQ file (cf. Section 2.2).

Taking into account these groups of clinical records, a total of 76 candidate predictor variables were included in the correlation study with the target variable (*has\_endometriosis*; c.f., Equation (2)). However, before performing the correlation study, these candidate variables were subjected to data pre-processing processes.

On the one hand, medical records with erroneous data (caused by errors in data entry into the EHR system) were eliminated from the original data set so as not to alter the prediction (e.g., records associated with patients over 2.5 m tall or 300 kg in weight), among other outliers.

On the other hand, it was necessary to transform the value of some date-type candidate variables (such as date of birth) to establish a numerical value. Then, text and categorical candidate variables (such as the target variable 2) were also transformed into numerical values. Subsequently, all null values were replaced by zero.

Finally, the patient’s FASTQ file is preprocessed following the procedure described in Section 2.2. After removing the genetic adapters from this file and aligning the patient’s sequence with the human reference sequence, the patient’s genetic variants (which are related to endometriosis) are identified. For this purpose, our preprocessing process takes into account the endometriosis-related chromosomal variants published in ClinVar (ClinVar, 2023).

In addition, once the data preprocessing process was completed, the statistical study of correlation between the candidate variables was carried out. This study is based on the construction of a data matrix with correlation coefficients, which measures the strength and direction of the relationship between two variables. In our article, we focus on the correlation coefficient between our target variable (2) and each candidate predictor variable. In this regard, Equation (3) represents the calculation of the correlation between variables  $x$  and  $y$ , where  $x_i$  and  $y_i$  are individual data points of variables  $x$  and  $y$ , respectively, and  $\bar{x}$  and  $\bar{y}$  are the mean values of variables  $x$  and  $y$ , respectively.

$$R_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}} \tag{3}$$

The result of this analysis is a correlation coefficient that can take values between  $-1$  and  $+1$ . The sign indicates the type of correlation between the two variables. A positive sign indicates that there is a positive relationship between the two variables (i.e., when the magnitude of one increases, the other also increases), while a negative sign indicates that there is a negative relationship between the two variables (i.e., when the magnitude of one increases, the values of the second variable decrease). If two variables are independent, the correlation coefficient has a null value in magnitude. In addition, we have followed recommendations from leading authors to select candidate variables: (a) excluded variables that are correlated with each other to avoid providing little useful information within the predictive model (as recommended by Harrell (Harrell et al., 2001)); (b) included some variables that are not statistically significant, but they could contribute to the fit of the predictive model according to the judgement of health professionals (as Sun et al. (Sun et al., 1996) recommend); (c) analysed categorical variables and continuous variables separately as suggested by Steyerberg et al. (Steyerberg & Vergouwe, 2014).

Finally, after considering the clinical evidence on endometriosis and the clinical knowledge of the health professionals involved in the study, the variables whose  $R > 0.2$  are appropriate to be included in our predictive models. In this regard, Table 1 shows the correlation coefficients between the target variable (*has\_endometriosis*; cf., Equation (2)) and the final predictor variables proposed to build the prediction model. In total, finally, **30 final predictor variables** were considered to build the prediction model (cf., Table 1). Fig. 5 shows the prevalence of predictor variables in relation to endometriosis outcome. In addition to this initial set, additional disease-related variants identified during alignment with the reference sequence will be incorporated. These variants, which will be integrated as an additional column of data with a unique identifier, will expand our understanding of the genetic contribution to endometriosis.

**Table 1**  
Correlation coefficients between target variable (*has\_endometriosis*) and the final predictor variables proposed to build the prediction model.

Predictor variables	Target Variable <i>has_endometriosis</i> R	Predictor variables	Target Variable <i>has_endometriosis</i> R
Left ovarian follicles (Quantitative: 0 to 18)	0.07151297	Uterus shape (Qualitative: Normal, pathological, adenomyosis, anteversion, arcuato, bicornis, didelphic, double, myomatous, polypomatous, retroversion)	0.02495630
Estradiol hormone result (Quantitative: 0 to 98)	0.0132533237	Uterine cavity (Qualitative: normal, suggestive of adherent syndrome)	0.06560355
Uterine angle (Qualitative: normal, very obtuse, very sharp)	0.02430920	Current toxicity: Alcohol consumption (Qualitative: 0, 1)	0.01270785
Right ovarian follicles (Quantitative: 0 to 20)	0.05789266	Left ovary height (Qualitative: normal, distal/hydrosalpinx obstruction, atrophic obstruction)	0.02415706
Result Free T4 hormone (Quantitative: 0 to 89.5)	0.01262009	Result Androstenedione hormone (Quantitative: 0 to 21)	0.05417207
Blood group (Qualitative: 0-, 0+, A+, A-, B+, B-, AB+, AB-)	0.01201567	Right ovary height (Quantitative: 0 to 20)	0.04945191
Patient age (Qualitative: 15 to 75)	0.01932606	Alteration in the nervous system (Qualitative: 0, 1)	0.01184940
Cycle day (Quantitative: 0 to 398)	0.04537355	SHBG hormone result (Quantitative: 0 to 60)	0.01868803
Developmental alteration (Qualitative: 0, 1)	0.01138343	17OH hormone result Progesterone (Quantitative: 0 to 11)	0.04378235
Inhibin B hormone result (Quantitative: 0 to 10)	0.01830991	DHEA-S hormone result (Quantitative: 0 to 4)	0.01116521
Left ovary width (Qualitative: normal, distal/hydrosalpinx obstruction, atrophic obstruction)	0.03916578	Endometrial size (Qualitative: linear, secretory, normal, pathological, proliferative, trilaminar)	0.01688583
Eye color (Qualitative: blue, black, green, brown, amber, others)	0.01093915	FSH hormone result (Quantitative: 0 to 10)	0.03134126
Patient weight (Quantitative: 0 to 121)	0.01626366	Current toxicity: Other (Qualitative: 0, 1)	0.01090932
Hormone result Antisuprarenal antibody (Quantitative: 0 to 13)	0.02537495	Right ovarian volume (Quantitative: 0 to 21)	0.01624684
ANTI 21-Hydroxylase Hormone Result (Quantitative: 0 to 7.9)	0.01011184	Progesterone hormone result (Quantitative: 0 to 9.23)	0.02104320

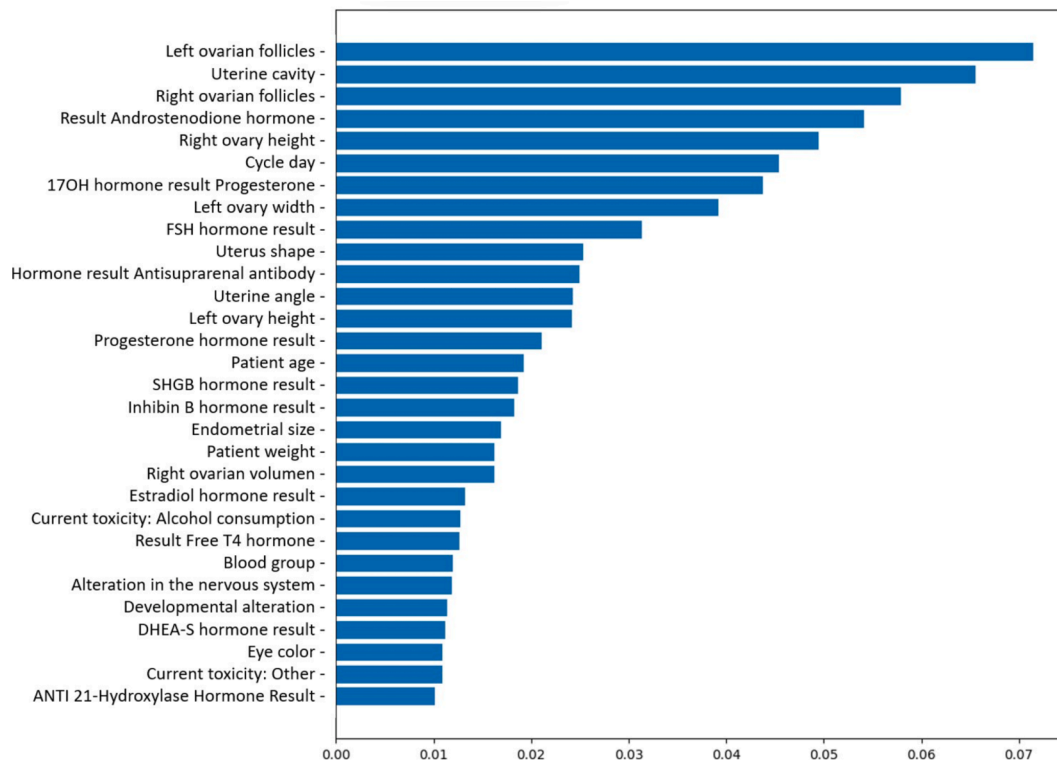


Fig. 5. Prevalence of predictor variables with respect to the target variable *has\_endometriosis*.

### 3.5. Management of missing data and unbalanced classes

On the one hand, regarding missing data, we applied a missing data imputation process. We used imputation by the median in numerical variables and by the mode in categorical variables, thus ensuring that the information represented is as close as possible to the expected clinical behavior.

On the other hand, we have applied specific macro metrics to verify the prediction of minority classes and carry the management of unbalanced classes. Specifically, metrics such as the F1 metric, sensitivity, or accuracy, among others, are considered as described below. Specifically, nine measures were used to assess the performance of each predictive model: (i) **Accuracy**, which is the fraction of accurate predictions by our model; (ii) **Sensitivity**, which represents the probability of obtaining a positive result in a test, considering that the individual is actually positive; (iii) **Precision**, which is the fraction of all relevant instances divided by the instances obtained; (iv) **AUC (Area Under the ROC Curve)**, which indicates the probability that the model classifies a random positive example higher than a random negative example; (v) **F1**, which is an alternative evaluation metric in machine learning that evaluates the predictive ability of a model by focusing on the performance per class, rather than a global evaluation such as accuracy; (vi) **False Positive Rate**, which is the probability of obtaining a test result that erroneously indicates the presence of a specific condition or attribute; (vii) **Log Loss**, which indicates how close the predicted probability is to the corresponding true value. As the predicted probability moves away from the true value, the log loss value increases; (viii) **Hamming Loss**, which is the fraction of incorrect labels relative to the total number of labels; (ix) **Time** which refers to the time in milliseconds required to make a prediction.

### 3.6. Method for pseudonymisation clinical data

In accordance with the Declaration of Helsinki (Williams, 2008) and the Data Protection Directive (Directive 95/46/EC) (Hustinx, 2013); we

have implemented a data pseudonymization method to ensure the confidentiality and security of the clinical data obtained for this research. The primary objective of this study is to gain insights into the causes, progression, and implications of the mentioned medical conditions, with the aim of enhancing preventive, diagnostic, and therapeutic interventions for patients affected by these ailments.

While there isn't a universally standardized set of guidelines for pseudonymizing clinical datasets (Chevrier et al., 2019; Hrynaszkiewicz et al., 2010); there is a degree of consensus on the techniques employed for achieving pseudonymization (Rodriguez et al., 2022). However, these techniques alone may not be sufficient to safeguard patient privacy. Hence, we have adopted a two-pronged approach by combining pseudonymization techniques with controlled access to information. In the following section, we provide a detailed description of the pseudonymization method used for extracting clinical data.

From the perspective of controlled access to information, the clinical researchers in this study (who belong to Inebir) were the only researchers with direct access to patients' clinical information (which was stored in the iMedea system, as previously mentioned). The remaining researchers only had permission to use anonymized data to carry out this research.

Once raw clinical data (exclusively, data related to anamnesis and diagnostic tests necessary for this research) were extracted, the following techniques were applied to pseudonymise:

**1. Suppression or Elimination Technique.** This method was used to remove sensitive data, including free-text fields where patients' personal information could be present, and non-essential technical data that could potentially identify the patient. Such technical data included episode identifiers, identifiers of health professionals entering information in the system, the patient's electronic medical record identifier, and the date of data registration in the system, among others.

**2. Recalculation technique.** Patients' ages were recalculated in years rather than displaying their date of birth, thus preventing the potential identification of patients through their birthdates

**3. Replacement and Obfuscation Technique.** In the iMedea

system, data is organized into multiple tables. For example, information related to ultrasound tests is distributed across various tables, such as basic test data, uterine angle, peritoneal cavity, ovarian condition and more. To ensure that these relationships could not be used to link back to a patient's medical history, the identifiers connecting these tables were replaced with new, randomly generated unique identifiers.

**4. Disaggregated Information Grouping Technique.** In conjunction with the third technique, information from various tables within the iMedea database was consolidated into a single record in the final dataset. For instance, data from each ultrasound test was unified into a single record, allowing for the elimination of connections between different database tables that could potentially lead to patient identification through inference

**5. Data Destruction technique.** Once data preprocessing was finalized, the temporary file containing the original dataset was securely destroyed, leaving only the pseudonymized post-processing dataset available for conducting the research as outlined at the beginning of this paper

## 4. Results

Once the final variable predictors have been defined, a comparison of the different prediction methods and algorithms is made. Fig. 6 shows the flow chart of the experiment to predict endometriosis outcomes. In total, medical histories of 5,143 female patients were reviewed; this dataset was divided into 80 % as a training set for model development (i. e., 4,150 records) and 20 % as a test set for the model (i.e., 993 records).

After splitting the dataset into two subsections (training and test), a predictive model was generated for each of the classification algorithms supported by Spark. Four classifier models were tested to construct the predictive model for endometriosis: random forest, decision tree classifier, and naive Bayes. In addition, the Receiver Operating Characteristic (ROC) curve graphs was calculated to determine the best endometriosis prediction model. The ROC curve graphically represents the sensitivity versus specificity of a classifier model. To know whether the ROC curve shows a correct prediction, the shape of the curve should be observed. The farther above the central line, the better the predictive model, whereas, if it coincides with the central line, the diagnostic value will be null.

Finally, Table 2 reflects that the *Random Forest* model outperformed the *Decision Tree Classifier*, *Naive Bayes* and *One-Versus-All Classifier* models in terms of precision, sensitivity, accuracy, AUC, F1, false positive rate, log loss, Hamming loss and time. Once the associated ROC curves are obtained, which can be seen in Fig. 7, we observe that the *Decision Tree Classifier* algorithm shows the highest prediction accuracy with the shortest time, so this model was chosen to predict the results of endometriosis in our study. Additionally, we have included in the table the standard deviation (SD) of the data for each algorithm in order to compare the performance of training and testing datasets. Standard deviation is a measure used to quantify the variation or dispersion of a

set of numerical data is calculated as follows in Equation (4), where  $x_i$  is a value from a data set,  $\mu$  is the mean of the data set and  $n$  is the sample size, in our case 2, corresponding to the training and test sets.

$$\sigma = \sqrt{\frac{\sum_{i=1}^2 (x_i - \mu)^2}{n - 1}} \quad (4)$$

## 5. Discussion

### 5.1. Case study

Endometriosis is a complex and challenging condition to diagnose, as it generally requires laparoscopic surgery followed by histological examination, which are considered the gold standard. The difficulty and cost associated with these methods mean that many women suspected of having this condition do not receive a definitive confirmation of their status. As a result, efforts have been intensified to identify biomarkers and more accessible and reliable diagnostic methods. Recently, a significant correlation between genetic alterations and the manifestation of endometriosis has been observed, particularly highlighting the ARID1A gene, although other genes have also been implicated. In this context, machine learning emerges as a promising tool for linking clinical data from gynecological patients with their genomic information, such as clinical exome data, thus enabling reliable disease predictions.

To address these challenges, the development of the model to predict the possible presence of endometriosis was based on an innovative approach that integrates both detailed clinical data and patients' genomic information. This multidisciplinary approach, described in previous sections of the article, leverages the potential of AI to analyze and correlate extensive clinical data sets with specific genetic variants, including those associated with the ARID1A gene and other genes potentially involved in endometriosis. Using a dataset of anonymized data from 5,143 female patients diagnosed with endometriosis at the Inebir private fertility clinic, the model employs machine learning algorithms to identify key patterns and variables that predict the presence of the disease with high reliability.

Once the model for predicting the possible presence of endometriosis was prepared, a retrospective and observational patient study was conducted. The proposal was fed with clinical and exome data from 14 patients, divided into two groups to assess the impact of genetic information on the model's accuracy. The first group included 7 patients without a previous genetic study but with a clinical history of endometriosis. The second group consisted of 7 patients who had genetic data, of which 4 showed no clinical signs of endometriosis and 3 were diagnosed by laparoscopy.

The results of the predictive model (Table 3 highlighted the relevance of integrating genetic data for the accurate identification of endometriosis. Of the 7 patients with genetic data, the model achieved a successful clinical prediction in all cases. On the other hand, in the group

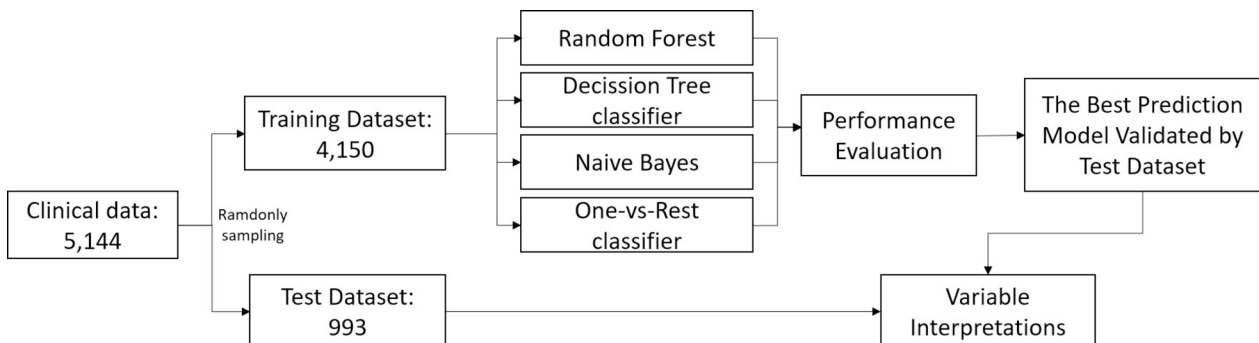
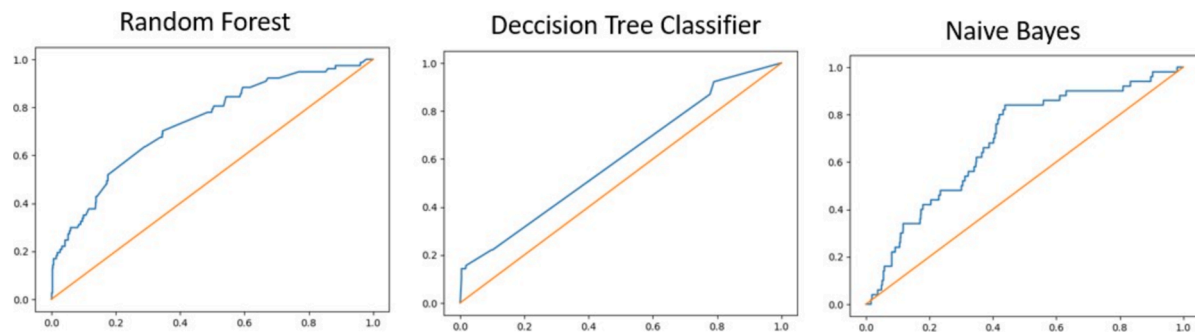


Fig. 6. Experiment flowchart for predicting endometriosis outcomes.

**Table 2**

Predictive performance of different machine learning algorithms for the training and test datasets. Abbreviations: Random forest (RF), Decision Tree classifier (DTc), Naive Bayes (NB), One-vs-Rest classifier (ORc), Standard Deviation (SD). Training dataset: 4150 records. Test dataset: 993 records.

Alghm.	Dataset	Accuracy	Sensitivity	Precision	AUC	F1	False Pos. Rate	Log Loss	Hamming Loss	Time (
RF	Training	0.947662	0.947662	0.950402	0.947662	0.922436	0.943053	0.176545	0.05233	135.17
	Test	0.957576	0.957576	0.916951	0.957576	0.936823	0.957576	0.161371	0.04242	60.72
	SD	0.007010	0.007010	0.023908	0.007010	0.010173	0.010269	0.010730	0.007007	52.64
DTc	Training	0.955900	0.955900	0.952419	0.955900	0.942171	0.768495	0.170549	0.04410	92.58
	Test	0.956566	0.956566	0.944079	0.956566	0.947646	0.775583	0.392673	0.04343	38.18
	SD	0.000471	0.000471	0.005897	0.000471	0.003871	0.005012	0.157065	0.000474	38.46
NB	Training	0.712624	0.712624	0.912124	0.712624	0.791275	0.555664	1.084868	0.28737	107.86
	Test	0.662626	0.662626	0.929589	0.662626	0.762919	0.492795	1.349928	0.33737	57.35
	SD	0.035354	0.035354	0.008107	0.035354	0.020263	0.044455	0.187426	0.777817	49.85
ORc	Training	0.947904	0.947904	0.931520	0.947904	0.925630	0.791275	–	0.05209	130.55
	Test	0.956566	0.956566	0.931939	0.956566	0.762919	0.934866	–	0.04343	102.33
	SD	0.006125	0.006125	0.000296	0.006125	0.115054	0.101534	–	0.006124	19.95



**Fig. 7.** Receiver operating characteristic (ROC) curves for endometriosis predictions based on the random forest, decision tree classifier, and naive bayes models for the test dataset.

**Table 3**

Predictions of the machine learning model with data from patients with and without endometriosis included to determine the reliability of the tool.

n	Endometriosis	Genetic data (exome)	Correct model prediction
7	Yes	No	3/7 (42 %)
4	No	Yes	4/4 (100 %)
3	Yes	Yes	3/3 (100 %)

without genetic data, the prediction was accurate in only 3 of them. This finding emphasizes the importance of exome sequencing data in the model’s prediction, significantly improving diagnostic accuracy in cases where genetic data were included.

This study, although conducted with a limited sample and the absence of genetic data in some patients, strongly suggests that the inclusion of exome sequencing information is key to improving the accuracy of predictive models for endometriosis. By integrating clinical and genetic data, the predictive model developed and tested in this study demonstrates significant potential for early and non-invasive diagnosis of endometriosis, offering a valuable tool for clinical decision-making and management of this complex disease.

**5.2. Related works**

In the context of addressing the complexity of infertility and reproductive assistance, several studies have explored approaches to assess and enhance the endometriosis diagnosis. In this section, we examine related works addressing the prediction of endometriosis using machine learning, highlighting their contributions and limitations. We underline the differences with our proposal in this field.

Wölfler et al. (Wölfler et al., 2005) presented a logistic regression model to predict endometriosis disease by considering clinical ameters from endometrial biopsies and laparoscopic testing of symptomatic

women for the presence of the omatase enzyme. The authors’ dataset contained clinical data of 48 consecutive symptomatic and eligible patients, 25 (52.1 %) had endometriosis and 23 (47.9 %) were free of disease. The authors’ model only considers these key variables, revealing that 95.5 % of patients whose eutopic endometrium tested positive for the aromatase enzyme (and who also had dysmenorrhoea) had endometriosis at laparoscopy. Authors showed that screening for eutopic endometrial aromatase could have discriminative value in predicting the endometriosis.

Ashrafi et al. (Ashrafi et al., 2016) describe a logistic regression model to predict the risk of diagnosis of endometriosis in infertile omen and identify the correlation between some relevant factors and symptoms. The authors’ dataset contained clinical data of 341 infertile women with endometriosis (cases) and 332 infertile women with a normal pelvis (comparison group). The main predictor variables used by the authors are related to patient anamnesis. Specifically, the authors’ model revealed that fatigue, diarrhoea, constipation, dysmenorrhoea, dyspareunia, pelvic pain and premenstrual spotting were more significant among patients with late-stage endometriosis than in those with early- stage endometriosis, and more prevalent among patients with endometriosis than among those in the comparison group. In addition, pregnancy, family history of endometriosis, history of galactorrhoea, history of pelvic surgery, dysmenorrhoea, pelvic pain, dispareunia, premenstrual spotting, fatigue and diarrhoea were significantly associated with endometriosis. However, the number of pregnancies was negatively associated with endometriosis.

Verket et al. (Verket et al., 2019) identified predictors of endometriosis disease among a few factors (based on interviews) commonly associated endometriosis, which are mainly used in primary care (these predictors are: age at menarche; severe dysmenorrhoea in adolescence; school absenteeism for dysmenorrhoea; analgesic use for dysmenorrhoea in adolescence; oral contraceptive use for dysmenorrhoea in adolescence; family history of endometriosis). These predictors were

combined to develop a logistic regression prediction model for early identification of endometriosis risk in at-risk women. The authors' research considered 18–45 years of age and surgically confirmed diagnosis as inclusion criteria. In total, 157 women with endometriosis and 156 women without endometriosis were included in this study. The authors acknowledge weaknesses such as that they did not use medical records. Therefore, the severity of endometriosis could not be assessed. In addition, the authors did not exclude the possibility of recall bias. Women with endometriosis may be more likely to recall symptoms suggestive of endometriosis experienced in adolescence compared to women without endometriosis. A third weakness is the low response rate from the general population, following a general international trend of declining response rates to postal surveys.

Compared to previous works, the DELFOS project stands out as a multidisciplinary initiative addressing genetic counseling in assisted reproduction through machine learning technologies. Unlike related works, our primary contribution is the creation of an expert system that, through the analysis of electronic health records, discovers genetic pathologies and provides real-time early alerts. This feature ensures healthcare professionals can make informed and rapid decisions during assisted reproduction treatments, which can be crucial in genetic risk situations. Furthermore, our platform distinguishes itself by its integration capabilities with various healthcare organizations, ensuring its transferability and application in a wide range of clinical contexts.

### 5.3. Highlights and threats to validation

While in the final this study has made significant progress and contributions, it is important to acknowledge and address the inherent limitations that could impact the reliability, generalizability, and applicability of the findings. These limitations not only strengthen the scientific rigor of the study, but they also provide a solid foundation for future research and enhancements to the proposed method. The key limitations that need to be considered when interpreting the results and extrapolating the conclusions are presented below:

On the one hand, it is possible to identify some limitations related to the dataset. Although we have included several important variables in our predictive model, it is possible that there are other factors or aspects that have not been taken into account and could affect the predictions (*unconsidered variables*). This study was based on the specific variables that were available in the dataset used, but variables that were not considered in this study could influence the observed connections. The inclusion of additional variables could provide a more comprehensive understanding of the results. Investigating new variables could improve the accuracy and generalizability of the predictive model. In addition, the *quality of the dataset used* to train and validate our model is essential and critical to increase the reliability of the model. Therefore, erroneous, incomplete or mislabelled data could affect the accuracy of the predictions. Another factor related to the dataset is its size. The *sample size* used to train and evaluate the model can affect its performance because a small sample size could lead to inaccurate results or a lack of generalizability. Obtaining a larger and more varied dataset could help improve the model's ability to make accurate predictions.

On the other hand, it is possible to identify some limitations related to the predictive model. Our model has been created and tested on a particular dataset. The model's ability to apply to different populations, clinical settings, or demographic groups may be restricted. Validation on several datasets (*generalization*) could help demonstrate its reliability in different situations. In addition, predictive models frequently rely on certain *assumptions* about the connection between variables. These assumptions may not always be accurate. Examining the validity of these assumptions and investigating more adaptable methods could improve the model's accuracy.

Finally, it is important to consider the *interpretation of results* as another possible limitation. While our model can make accurate forecasts, the interpretation of the results and the making of clinical

decisions based on those forecasts require the judgment and expertise of healthcare experts. This risk has been mitigated thanks to the participation of researchers and health experts in the field of genetics and human reproduction in this research.

## 6. Conclusions and future works.

The utilization of predictive models stands as a versatile resource capable of contributing to various sectors of society, with health being a particularly crucial domain. In this paper, we delve into the benefits of employing such technology, particularly within the clinical context of endometriosis detection.

Commencing with a comprehensive overview of endometriosis, we establish the research hypothesis proposing the utilization of predictive models for early detection, leveraging clinical data. Subsequently, we elucidate the materials and methods employed, detailing their integration within the DELFOS platform. DELFOS, a real-world system utilizing both clinical and genetic data, is instrumental in predicting endometriosis.

Following this, we present a thorough analysis of the results obtained. Concluding the paper, we highlight significant findings and underscore the empirical application of the predictive model within the authentic clinical setting of DELFOS.

Looking forward, the prospects for future work within this context appear promising. While our current data may be limited, preliminary results underscore the viability of employing such technology for endometriosis detection. Moreover, despite the constraints posed by a limited dataset, the machine learning protocols integrated within DELFOS offer avenues for enhancing results. Furthermore, while the genetic data available in this study is somewhat restricted, our commitment to ongoing research in this domain remains steadfast. Preliminary findings serve as a catalyst, motivating us to pursue further investigations and gather additional insights within this realm.

In future work, we aim to enhance the model by exploring deep learning techniques. Although shallow models were prioritized in this study for their interpretability and computational efficiency, the potential of deep learning approaches remains significant. Advanced architectures, such as convolutional neural networks and deep neural networks, offer enhanced capabilities for capturing complex interactions within clinical and genetic data. Additionally, in order to enhance model calibration, we plan to incorporate the Brier score and calibration plots in future iterations of the model. This metric will provide a valuable layer of assessment regarding the alignment between predicted probabilities and actual clinical outcomes, thus enhancing the model's applicability and reliability in clinical settings.

These planned enhancements are expected to further evolve the DELFOS platform, enabling it to become an increasingly valuable tool for clinicians in the early detection and management of endometriosis. By incorporating these improvements, we aim to provide more accurate, reliable, and actionable insights for clinical decision-making, thereby advancing the standard of care for patients affected by this complex disease.

### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### Acknowledgments

We are grateful for the outstanding collaboration of software engineers Reyes Blasco and José Antonio Domínguez during the development of the Delfos platform. They belong to the G7innovation company and the ES3 Group (Engineering and Science for Software Systems group) of the University of Seville (Spain), respectively.

This research was supported by the (i) EQUAVEL project PID2022-137646OB-C31, funded by MICIU/AEI/10.13039/501100011033 and by ERDF, EU.; the grant PREP2022-000332 funded by MICIU/AEI/10.13039/501100011033 and by ESF+; (ii) the DELFOS project (021/C005/00151010) of the red.es Spanish organisation within its 2021 call to fund R&D projects related to artificial intelligence and other digital technologies integrated in value chains (C005/21-ED).

### Ethic Committee

It is important to note that this study has been approved by the Ethics Committee of the European University of the Atlantic as recorded in Act number 78, under registration number CEI-40/2023, dated 05 October 2023. In addition to the approval from the Ethics Committee, informed consent was obtained from all participants involved in the study. This ensures compliance with legal and ethical standards regarding patient data.

### Data availability

Data will be made available on request.

### References

- Alonzo, T.A., 2009. Clinical prediction models: a practical approach to development, validation, and updating: by ewout w. steyerberg.
- Altman, D. G., & Royston, P. (2000). What do we mean by validating a prognostic model? *Statistics in Medicine*, 19, 453–473.
- Altman, D. G., Vergouwe, Y., Royston, P., & Moons, K. G. (2009). Prognosis and prognostic research: Validating a prognostic model. *BMJ*, 338.
- Ashrafi, M., Sadatmahalleh, S. J., Akhond, M. R., & Talebi, M. (2016). Evaluation of risk factors associated with endometriosis in infertile women. *International Journal of Fertility & Sterility*, 10, 11.
- Ayaz, M., Pasha, M. F., Alzahrani, M. Y., Budiarto, R., & Stiawan, D. (2021). The fast health interoperability resources (fhir) standard: Systematic literature review of implementations, applications, challenges and opportunities. *JMIR Medical Informatics*, 9, Article e21929.
- Bendifallah, S., Puchar, A., Suisse, S., Delbos, L., Poilblanc, M., Descamps, P., Golfier, F., Touboul, C., Dabi, Y., & Daraï, E. (2022). Machine learning algorithms as new screening approach for patients with endometriosis. *Scientific Reports*, 12, 639.
- Bullon, P., & Manuel Navarro, J. (2017). Inflammation as a key pathogenic mechanism in endometriosis. *Current Drug Targets*, 18, 997–1002.
- Chevrier, R., Foufi, V., Gaudet-Blavignac, C., Robert, A., & Lovis, C. (2019). Use and understanding of anonymization and de-identification in the biomedical literature: Scoping review. *Journal of medical Internet research*, 21, Article e13484.
- ClinVar, 2023. Chromosomal variants related to endometriosis. URL: <https://www.ncbi.nlm.nih.gov/clinvar/?term=endometriosis>.
- Collins, G. S., Moons, K. G., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., Van Smeden, M., et al. (2024). Tripod+ ai statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385.
- García García, J. A., Ramírez de Verger, C., & Sánchez Gómez, N. (2022). La calidad del software como mecanismo de éxito en proyectos multidisciplinares: Proyecto imedeia y meet2care. *Calidad y Sostenibilidad de Sistemas de Información en la Práctica*.
- Gateway, G.B., 2022. Human genome browser - hg38 assembly. URL: <https://genome.ucsc.edu/cgi-bin/hgGateway?redirect=manual&source=genome.ucsc.edu>. [Sequencing/Assembly provider ID: Genome Reference Consortium Human GRCh38.p14 (GCA\_000001405.29); Last access: September, 2023].
- Harrell, F.E., et al., 2001. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. volume 608. Springer.
- Hrynaskiewicz, I., Norton, M. L., Vickers, A. J., & Altman, D. G. (2010). Preparing raw clinical data for publication: Guidance for journal editors, authors, and peer reviewers. *BMJ*, 340.
- Hustinx, P. 2013. Eu data protection law: The review of directive 95/46/ec and the proposed general data protection regulation. University of Tartu. Data Protection Inspectorate, Tallinn.
- Koleck, T. A., Dreisbach, C., Bourne, P. E., & Bakken, S. (2019). Natural language processing of symptoms documented in free-text narratives of electronic health records: A systematic review. *Journal of the American Medical Informatics Association*, 26, 364–379.
- Laupacis, A., Sekar, N., et al. (1997). Clinical prediction rules: A review and suggested modifications of methodological standards. *JAMA*, 277, 488–494.
- Macer, M. L., & Taylor, H. S. (2012). Endometriosis and infertility: A review of the pathogenesis and treatment of endometriosis-associated infertility. *Obstetrics and Gynecology Clinics*, 39, 535–549.
- Maicas, G., Leonardi, M., Avery, J., Panuccio, C., Carneiro, G., Hull, M. L., & Condous, G. (2021). Deep learning to diagnose pouch of douglas obliteration with ultrasound sliding sign. *Reproduction and Fertility*, 2, 236–243.
- Mak, J., Leonardi, M., & Condous, G. (2022). ‘Seeing is believing’: Arguing for diagnostic laparoscopy as a diagnostic test for endometriosis. *Reproduction and Fertility*, 3, C23–C28.
- Men, A.E., Wilson, P., Siemering, K., Forrest, S., 2008. Sanger DNA sequencing. Next Generation Genome Sequencing: Towards Personalized Medicine, 1–11.
- Modi, A., Vai, S., Caramelli, D., & Lari, M. (2021). The illumina sequencing protocol and the novaseq 6000 system. *Bacterial Pangenomics: Methods and Protocols*. Springer, 15–42.
- National Center for Biotechnology Information, NCBI, 2009. Homo sapiens genome assembly GRCh37. URL: [https://www.ncbi.nlm.nih.gov/datasets/genome/GCF\\_000001405.13/](https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.13/).
- Nnoaham, K.E., Hummelshoj, L., Kennedy, S.H., Jenkinson, C., Zondervan, K.T., Consortium, W.E.R.F.W.H.S.S., et al., 2012. Developing symptom-based predictive models of endometriosis as a clinical screening tool: results from a multicenter study. *Fertility and Sterility* 98, 692–701.
- Organization, W.H., 2023. Endometriosis. URL: <https://www.who.int/news-room/fact-sheets/detail/endometriosis/>.
- Parasar, P., Ozcan, P., & Terry, K. L. (2017). Endometriosis: Epidemiology, diagnosis and clinical management. *Current Obstetrics and Gynecology Reports*, 6, 34–41.
- Rodriguez, A., Tuck, C., Dozier, M. F., Lewis, S. C., Eldridge, S., Jackson, T., Murray, A., & Weir, C. J. (2022). Current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: A scoping review. *Clinical Trials*, 19, 452–463.
- Salloum, S., Dautov, R., Chen, X., Peng, P. X., & Huang, J. Z. (2016). Big data analytics on apache spark. *International Journal of Data Science and Analytics*, 1, 145–164.
- Saunders, P. T., & Horne, A. W. (2021). Endometriosis: Etiology, pathobiology, and therapeutic prospects. *Cell*, 184, 2807–2824.
- Sivajohan, B., Elgendi, M., Menon, C., Allaire, C., Yong, P., & Bedaiwy, M. A. (2022). Clinical use of artificial intelligence in endometriosis: A scoping review. *NPJ Digital Medicine*, 5, 109.
- van Stein, K., Schubert, K., Ditzen, B., & Weise, C. (2023). Understanding psychological symptoms of endometriosis from a research domain criteria perspective. *Journal of Clinical Medicine*, 12, 4056.
- Steyerberg, E. W., & Vergouwe, Y. (2014). Towards better clinical prediction models: Seven steps for development and an abcd for validation. *European Heart Journal*, 35, 1925–1931.
- Sun, G. W., Shook, T. L., & Kay, G. L. (1996). Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *Journal of Clinical Epidemiology*, 49, 907–916.
- Verket, N. J., Falk, R. S., Qvigstad, E., Tanbo, T. G., & Sandvik, L. (2019). Development of a prediction model to aid primary care physicians in early identification of women at high risk of developing endometriosis: Cross-sectional study. *BMJ Open*, 9, Article e030346.
- Wagner, D. D., Carleton, H. A., Trees, E., & Katz, L. S. (2021). Evaluating whole-genome sequencing quality metrics for enteric pathogen outbreaks. *PeerJ*, 9, Article e12446.
- Williams, J. R. (2008). The declaration of helsinki and public health. *Bulletin of the World Health Organization*, 86, 650–652.
- Wölfler, M., Nagele, F., Kolbus, A., Seidl, S., Schneider, B., Huber, J., & Tschugguel, W. (2005). A predictive model for endometriosis. *Human Reproduction*, 20, 1702–1708.
- Zhou, Z. H. (2021). *Machine learning*. Springer Nature.