Bocconi

Bachelor of Science in Economics,
Management and Computer Science

Fairness-aware learning and subgroup modelling for chronic pelvic pain condition identification

Addressing subpopulation biases in gynaecological health data

Advisor:

Prof. Luca Saglietti

Bachelor of Science Thesis

by:

Cecilia Di Francesco

Student ID No. 3194942

Academic Year 2024-2025

Università Bocconi
MILANO

Al mio papà,
che mi ama con un'intensità che riconosco ogni giorno: negli occhi, nei gesti, nel silenzio.
Resterò sempre la tua bambina, e nei momenti difficili correrò da te.
Ti porto con me, in ogni cosa che faccio.

Alla mia mami,
per avermi insegnato a essere forte, indipendente e libera.
La mia forza, lo sai, viene da te.
So che ogni volta che parto, ti si stringe il cuore - e il mio, un po', resta con te.
Ogni passo che faccio, c'è il tuo amore dietro.

Ai miei fratellini,
che con la loro instancabile confusione rendono ogni ritorno un po' più casa.
Mi manca vedervi crescere ogni giorno, ma siete sempre con me.

Alle mie amichette,
per essere state rifugio sicuro e leggerezza pura, per le risate che hanno curato ferite invisibili e quella presenza costante che non ha mai chiesto nulla se non di esserci. Siete il dono più prezioso che porto con me da questi anni.

A Ele,
per la delicatezza, la forza condivisa e quella comprensione reciproca che non ha mai avuto bisogno di troppe parole. Sei una presenza luminosa, capace di farmi sentire accolta, sempre.

A Marco,
per avermi ricordato ogni giorno quanto valgo, e per avermi fatto scoprire una versione di me che, senza di te, non avrei conosciuto.
Grazie per avermi accompagnata, mano nella mano, anche in questo percorso.

Table of Contents

Table of Tables

Table of Figures

Table of Tables (Appendix C)

Table of Figures (Appendix D)

1. Introduction

1.1 Introduction

Chronic gynecological conditions affect a significant portion of women, yet they remain among the most underdiagnosed and poorly understood areas in modern medicine. Diseases such as Endometriosis, Adenomyosis, Polycystic Ovary Syndrome (PCOS), Vulvodynia, and Fibromyalgia typically rely on nonspecific or overlapping symptoms, which leads to years of diagnostic delay. This creates a diagnostic gap that further delays treatment and undermines the quality of life for millions of women.

In recent years, digital health platforms have provided promising new avenues to help bridge this gap: accessible, scalable, and user-centered symptom assessment tools. Such tools can be critical when combined with advancements in machine learning, as they help uncover unexpected patterns in patient-reported symptoms. Moreover, they represent an opportunity to collect large-scale, real-world datasets that reflect the real-life experiences of people seeking answers about their health.

However, there are complexities to developing predictive models in this domain: structural biases in self-reported health data, low rates of formal diagnoses followed by multiple characterizations of symptom representation and expression between user subgroups, and ultimately, the inability of most machine learning models in clinical settings to generalize from the dominant patterns in the data they are trained on to those rare patterns would benefit from the most support.

This thesis builds on the work done during an internship with Hale, a women's health platform, to investigate whether alternative modeling approaches, including *clustering*, *bias-aware subgroup analysis*, and *subpopulations-informed classification models*, can help achieve fairer, more accurate, and more clinically valuable results. This project aims to identify meaningful subpopulations and customize models according to the nuances of each group, thereby supporting the early detection of often-neglected gynecological conditions while also identifying pathways to more equitable and effective digital health solutions.

1.2 The Hale Questionnaire and Dataset

This thesis utilizes data collected through an extensive screening questionnaire developed by Hale, a digital health platform dedicated to supporting women's gynecological health and well-being. The questionnaire was developed following current clinical guidelines and has been refined over several years of interacting with patients. Its purpose is to gather symptom data that may be indicative of several chronic gynecological and comorbid conditions, many of which are commonly underdiagnosed or misdiagnosed, and help women identify the cause of their discomfort.

The questionnaire was publicly distributed via multiple digital channels, including social media, search engines, and Hale's website. It is designed to assess the likelihood of eight core conditions commonly associated with chronic pelvic pain, including: Endometriosis, Adenomyosis, Uterine Fibroids, Polycystic Ovary Syndrome (PCOS), Irritable Bowel Syndrome (IBS), Vulvodynia, Vaginismus, and Fibromyalgia.

The original goal of the Hale screening tool is threefold: to empower patients by validating their experiences and symptoms, to reduce diagnostic delays by identifying relevant symptom patterns early, and to act as a gateway to further services and medical evaluation on the Hale platform.

Each question is constructed to align with the specific symptomatology of the above conditions, based on clinical guidelines, literature, and expert opinions. The complete list of resources used in creating the screening is available in the Appendix.

The main sections include questions regarding menstrual cycle characteristics, unusual bleeding or spotting, chronic and cyclic pain patterns, intimate and intestinal symptoms, symptoms triggered by sexual activity or physical contact, systemic signs such as fatigue, hormonal imbalances, and cognitive symptoms, contraceptive use, medical history and previous diagnoses.

Participants were also asked to provide basic demographic and biometric data, including age, height, and weight. The questionnaire serves as both a clinical

assessment tool and a research instrument, providing a dataset for developing and validating machine learning models. The questionnaire is available in Appendix A: Screening Questionnaire.

1.3 Thesis Objectives

During my internship at Hale, I developed a Random Forest model to classify users' chronic pelvic pain conditions using data from the company's screening test. The model achieved improvements over baseline models in the chosen evaluation metrics, but did not achieve optimal performance. Most notably, predicted probabilities remained unusually low, even after calibration was applied, resulting in the classification of most new, unseen users as not at risk. This was due to the low prevalence of the positive class, as defined by users who reported a pre-existing medical diagnosis. Most users had not yet received a formal diagnosis, yet they still experienced symptoms and searched for solutions to their health concerns.

Moreover, there is another profound structural limitation: the model was learning to detect *who had already been diagnosed*, rather than identifying individuals *who should be diagnosed*. Since the questionnaire is often completed after users have begun some form of consultation or treatment, the symptom data is inherently *retrospective*. The questionnaire does not include questions about treatments or their duration; therefore, it is not possible to control for these factors. Treatments can reduce or mask symptoms, which introduces *spurious correlations* and distorts the true underlying clinical picture. As a result, models trained on this data risk poor generalizability, particularly for users in the early stages of symptom onset, before diagnosis or intervention has occurred.

As the dataset from the original one, on which I had built the model, has now expanded from 4,000 to 12,000 users, broader issues of data imbalance and bias have become more evident. Certain conditions are much more prevalent than others, some demographics are overrepresented, and symptom gravity varies across user subgroups. These dynamics create structural distortions in the dataset, resulting in predictive models that overfit to dominant patterns and perform poorly on atypical or underrepresented users.

Inspired by the theoretical framework presented in Saglietti et al. [1], this thesis aims to address these issues by examining how clustering, feature selection, and bias-aware subgroup modeling can improve fairness and predictive accuracy, with the overall goal of developing tools that enable the earlier and fairer identification of research-neglected gynecological conditions.

1.4 Clinical context: Chronic Pelvic Pain

Chronic pelvic pain (CPP) is a prevalent and complex condition that significantly affects women's health worldwide. It is estimated to impact up to 26% of women globally [2]. It represents a significant cause of gynecological consultations, accounting for 10% of all visits, 40% of laparoscopies, and 12% of hysterectomies in the United States alone [2]. Despite this burden, CPP is notoriously underdiagnosed and poorly understood, mainly due to its multifactorial nature, symptom overlap with many other disorders, and the absence of a single identifiable cause in many cases.

CPP is defined as non-cyclic pain localized to the pelvis or lower abdomen, lasting more than six months, and not exclusively related to menstruation or sexual intercourse [2]. However, its classification goes beyond duration or location. CPP must be approached through a biopsychosocial lens: pain perception is shaped by neurological, musculoskeletal, gastrointestinal, urological, hormonal, and psychological inputs, often co-occurring and mutually reinforcing.

Among the recognized contributors to CPP are gynecological conditions such as Endometriosis, Adenomyosis, Pelvic inflammatory disease, and Adhesions, alongside non-gynecological disorders like Irritable Bowel Syndrome (IBS), Interstitial cystitis, musculoskeletal dysfunctions, and neuropathic pain. Many of these conditions share similar symptoms (*comorbidity*) and can happen at the same time in a patient, which makes diagnosing and treating them more challenging.

Of particular interest is *central sensitization*: a heightened sensitivity of the central nervous system to pain inputs. This is crucial in the perpetuation and augmentation of pelvic pain. Women with CPP often experience multiple pain sites or syndromes, sleep disturbance, anxiety, depression, rumination, and catastrophizing, all of which

significantly impair quality of life [2]. Similarly, chronic pain becomes self-perpetuating due to *neural plasticity*, where “abnormal sensation can become independent of the original insult and may persist long after tissue healing” [3].

CPP is also closely associated with adverse life experiences. A history of sexual or physical abuse, especially in childhood, is significantly more common among women with CPP than in the general population. These experiences may lead to both physiological changes (e.g., central sensitization) and psychological conditions (e.g., PTSD), both of which are under-acknowledged in gynecological care.

Structural clinical practice and research issues contribute to these clinical and neurological difficulties. The female body has been historically underrepresented in clinical trials, resulting in a critical shortfall in the knowledge of how diseases present in women. Recent estimates indicate that women's participation in clinical trials remains below 50% [4]. Further, gender biases in clinical encounters lead to women's pain being more likely classified as emotional or psychosomatic, affecting diagnosis and care [5]. This systematic omission also reinforces the belief that women's pain is less relevant or less severe, which has serious health repercussions for women.

The literature advises against simplifying this issue. A narrow focus on gynecological anatomy risks overlooking the genuine contributors to pain, including psychosocial trauma, dysfunctional coping mechanisms, and coexisting chronic pain syndromes such as Fibromyalgia. Despite efforts to establish consensus guidelines, the diagnostic process remains complex, fragmented, and difficult to obtain. Many treatment strategies are multimodal, involving pharmacological management, physiotherapy, hormonal therapy, psychological support, and, when necessary, surgical intervention. However, high-quality evidence supporting specific interventions remains limited, especially for overlapping pain syndromes.

This diagnostic ambiguity and symptom overlap at the heart of CPP make it an ideal case study for applying machine learning (ML) tools. Yet, as this thesis explores, such applications must be calibrated to avoid perpetuating existing biases in clinical data, especially when prior diagnoses or therapeutic interventions may have altered symptoms.

1.5 Target Conditions

This thesis focuses on eight chronic conditions frequently associated with chronic pelvic pain (CPP). These disorders span gynecological, gastrointestinal, neurological, and systemic categories, and are often misdiagnosed or overlooked due to overlapping symptom profiles and gender-based biases in healthcare. Understanding their symptomatology is essential for developing fair and clinically meaningful predictive models.

1.5.1 Endometriosis

Endometriosis occurs when tissue similar to the endometrium grows outside the uterus. Common symptoms include severe *dysmenorrhea* (heavy and painful menstruation), deep *dyspareunia* (pain during intercourse), chronic pelvic pain, and painful defecation or urination, especially during menstruation. Infertility and fatigue are also prevalent. The condition affects approximately 10% of women of reproductive age and is one of the most common causes of CPP. Due to the normalization of symptoms and reliance on laparoscopic diagnosis, the average diagnostic delay ranges from 7 to 10 years [6].

1.5.2 Adenomyosis

Adenomyosis occurs when endometrial tissue infiltrates the muscular wall of the uterus, often coexisting with Endometriosis. It is typically associated with heavy menstrual bleeding (*menorrhagia*), intense cramping, and diffuse pelvic pain that can persist beyond the menstrual period [3]. Symptoms are often dismissed or misattributed to fibroids or normal menstruation. Imaging tools like MRI and transvaginal ultrasound can suggest the diagnosis, but histological confirmation is usually only possible after a hysterectomy.

1.5.3 Uterine Fibroids (Leiomyomas)

Uterine Fibroids are benign smooth muscle tumors of the uterus. While many cases are asymptomatic, fibroids can cause pelvic pressure, low back pain, heavy menstrual

bleeding, and urinary frequency. When large or submucosal, they may distort the uterus and exacerbate pain, especially when coexisting with other gynecologic conditions. Symptom severity often depends on the size, number, and location of the fibroids [7].

1.5.4 Polycystic Ovary Syndrome (PCOS)

PCOS is an endocrine-metabolic disorder characterized by *chronic anovulation* (lack or absence of ovulation), *hyperandrogenism* (excess amounts of androgens), and *polycystic ovarian morphology* (an imaging descriptor of a specific type of change in ovarian morphology). It is not a cause or condition of CPP, but it can be the cause of pelvic pain, bloating of the abdomen, and menstrual irregularity, symptoms associated with CPP. Patients also experience acne, hirsutism, weight gain, fatigue, and mood disturbances. These features make PCOS both hormonally and psychologically burdensome.

1.5.5 Irritable Bowel Syndrome (IBS)

IBS is a functional gastrointestinal disorder marked by recurrent abdominal pain, often related to bowel movements, along with bloating, constipation, diarrhea, or alternating patterns. It affects up to 80% of women with CPP, and its symptoms may mimic or exacerbate gynecological pain [8]. IBS and CPP share common mechanisms, including visceral hypersensitivity and central sensitization.

1.5.6 Vulvodynia

Vulvodynia is defined as chronic vulvar pain lasting at least three months without an identifiable cause. It can be localized (e.g., *Vestibulodynia*) or generalized, and symptoms are often described as burning, stinging, or rawness, either spontaneous or provoked (e.g., by touch, intercourse, tampon use). The pain significantly affects quality of life and sexual health. It is frequently underdiagnosed due to stigma, lack of physician training, and the absence of visible physical abnormalities during clinical examination. Many patients undergo multiple consultations before receiving a correct diagnosis, often being misdiagnosed with psychological or relational issues.

1.5.7 Vaginismus

Vaginismus is a pelvic floor dysfunction characterized by involuntary muscle spasms in response to attempted vaginal penetration. This leads to burning, tightness, or sharp pain, often described as “hitting a wall”. Vaginismus is most classically provoked by intercourse, insertion of a tampon, or gynecological exam. Vaginismus may also be secondary to trauma, anxiety, or other pain conditions such as Vulvodynia.

1.5.8 Fibromyalgia

Fibromyalgia is a chronic pain syndrome characterized by widespread musculoskeletal pain, fatigue, sleep disturbances, and cognitive difficulties (*fibro fog*). Many patients also report pelvic pain, irritable bladder symptoms, and hypersensitivity to touch. It commonly overlaps with conditions such as Irritable Bowel Syndrome (IBS) and Vulvodynia. It is often associated with central sensitization, making it a key contributor to chronic pelvic pain (CPP) presentations.

2. Fairness in Healthcare Machine Learning

2.1 Introduction

As ML models are being embedded in healthcare screening and decision-making tools, evaluating *fairness* has become a necessary process to ensure equitable outcomes for patient populations. There is evidence that, in clinical practice, inequities exist for marginalized groups concerning diagnosis, treatment, and access to care. Moreover, if left unaddressed, machine learning models that rely on historical healthcare data may recreate or amplify existing disparities. In this chapter, the foundations of fairness will be explored, especially in the context of healthcare. The analysis will examine the types of bias that can affect the data and the models, as well as the definitions and values of fairness and metrics, and how these concepts apply to the dataset used in this thesis.

2.2 Why Fairness Matters in Healthcare Machine Learning

Machine learning models have proven to be extremely valuable tools in initial screening, disease diagnosis, and risk prediction, as well as in providing recommendations for treatments. However, ML models for diagnosing disease, predicting risk, and managing triage are documented to underperform in minority groups defined by specific attributes like sex, race, and ethnicity [9]. For example, deep learning models for detecting 14 common diseases from chest X-rays have been shown to largely underdiagnose intersectional underserved subgroups, like Hispanic female patients, leading to delays in treatment if they were implemented in practice [9] [10]. In general, research highlights how the underrepresentation of women in healthcare datasets, or the absence of specific datasets for women's healthcare, can result in ML models performing poorly on female patients, mainly when the symptoms of a condition manifest differently or are less apparent in women than in their male counterparts.

Fairness is both an ethical matter and a huge challenge in developing replicable and valid models. In healthcare, where clinical algorithms assist in screening, diagnosing, and ultimately treating, trust can be eroded and compounded by systemic biases that may deepen existing inequities due to unfairness.

2.3 Types and Sources of Bias in Healthcare Datasets

Bias can arise at multiple stages of the data collection, processing, and model development. It can be categorized into three broad classes [9]: bias in the data, bias in the model, and bias during deployment. In the following sections, various biases in the data and the model are analyzed, as these are most relevant to the task.

2.3.1 Bias in the data

- *Societal bias* reflects entrenched social imbalances that are encoded in historical data. Clinical text datasets, for example, have been shown to contain biased language toward ethnic minorities [9]. In this case, it can manifest in several ways. For example, women with chronic pelvic or vulvar pain

(conditions poorly studied and discussed) may be met with dubious attitudes or delays in diagnosis due to gendered assumptions in clinical practice.

- *Selection bias* occurs when the data do not represent the population of interest. This is relevant in this case, given the reliance on a self-reported, online questionnaire. The sample is naturally skewed toward digitally literate individuals, who seek answers about their health, and who are already engaged with healthcare systems, potentially excluding those with limited access to healthcare, resources, or awareness.
- *Measurement bias* occurs when variables are collected inaccurately or inconsistently. In this case, it can stem from the subjective nature of self-symptom reporting, together with the retrospective framing of the questions. The symptoms reported by respondents may have already been affected by their diagnosis and the treatment they received.
- *Minority bias* occurs when the data lacks sufficient representation from minority groups, implying that the model struggles to learn the unique patterns of these groups effectively. This is highly pertinent to this situation because many of the target conditions for the predictive model show very low prevalence.

2.3.2 Bias in the model

- *Algorithmic Bias* can be introduced by the model architecture or the training method of the statistical model. In this case, it can occur when models overfit to prevalent conditions, thereby underperforming on other rare but equally clinically essential conditions. This is exacerbated by class imbalance and feature overlap.
- *Label Bias* refers to the use of an imperfect proxy instead of the specific outcome of interest. In this case, it could result from the retrospective approach to the diagnoses. The presence or absence of a diagnosis may reflect not the severity of symptoms but rather factors such as access to quality gynecological care, socioeconomic status, and cultural stigma associated with these issues.

2.4 Formal Fairness Definitions and Metrics

To adequately address fairness in healthcare machine learning, mathematical definitions that can be empirically evaluated and optimized are required. In this section, key fairness measures are examined in three major categories: *group fairness*, *individual fairness*, and *causal fairness*. Not only do these measures represent different notions of fairness, but they also document the trade-offs involved in clinical models, especially when there is label bias, prevalence imbalance, and a heterogeneous population.

2.4.1 Group fairness

Group fairness is defined as the constraint that machine learning models treat individuals similarly within certain identified groups (e.g., based on sex, race, and age). Group fairness is used as a fairness framework most widely in health applications due to its statistical quantification and clinical feasibility. However, employing group fairness becomes an issue of trade-off between generalization, equity, and performance.

Group fairness metrics can be categorized into three overall categories based on conditional independence relations among decisions (D), the *protected attribute* (A), and the outcome (Y) [9]. A protected attribute is typically a sensitive characteristic, such as sex, age, or race, used to assess equity across groups. In clinical settings, however, it can also include relevant subgroups, such as different symptomatic profiles, which may reflect structural differences in diagnosis or care.

Independence

According to the independence criterion, a model is considered fair if its decisions are statistically independent of the protected attribute, meaning the probability of receiving a positive prediction (e.g., being diagnosed with a condition) is similar across all groups. A widespread implementation of independence is *Standard Statistical (or demographic) Parity*, which requires that the model classifies people into the positive class at the same rate within each group.

$$P\,(D\ =\ 1\mid A=\ a_0\,)\ =\ P\,(D\ =\ 1\mid A=\ a_1\,)\,\forall a_0\,,a_1\ \in A$$ [9]

Conditional Statistical Parity weakens this constraint, as it only requires the rate of positive classifications to be the same within more discrete groups defined by the protected attribute and other relevant factors.

$$P\,(D\ =\ 1\mid Z,\ A=\ a_0\,)\ =\ P\,(D\ =\ 1\mid Z,A=\ a_1)\forall a_0,a_1\ \in A$$ [9]

Independence-based metrics, however, are commonly deemed inappropriate for application in healthcare settings, primarily because different population groups naturally have varying rates of prevalence for most conditions. Independence requires a model that not only predicts positive outcomes but does so at the same rate across all population groups, regardless of any actual differences in disease rates. This affects the meaningfulness of clinical signals and results, both in overprediction in the low-prevalence group and underprediction in the high-prevalence group, thereby reducing model accuracy and clinical usefulness.

Separation

Separation-based fairness requires that model predictions be independent of the protected attribute, given the actual outcome. This asserts that the model will have similar error rates across groups. In a formal sense, this entails equalized odds: equal false-positive and false-negative rates.

$$P\,(D\ =\ 1\mid Y\ =\ 0,A=\ a_0\,)\ =\ P\,(D\ =\ 1\mid Y\ =\ 0,A=\ a_1\,)$$ [9]

$$P\,(D\ =\ 0\mid Y\ =\ 1,A=\ a_0\,)\ =\ P\,(D\ =\ 0\mid Y\ =\ 1,A=a_1\,)\ \forall a_0\,,a_1\ \in A$$ [9]

A relaxation of this is *Equal Opportunity*, where only the false negative rate is equalized across groups. This is preferred when false negatives have the most severe consequences:

$$P\,(D\ =\ 0\mid Y\ =\ 1,A=\ a_0\,)\ =\ P\,(D\ =\ 0\mid Y\ =\ 1,A=a_1\,)\ \forall a_0\,,a_1\ \in A$$ [9]

In healthcare, these metrics are crucial since false negatives can result in dire consequences: not diagnosing an illness means delays in treatment, and aggravated suffering. On the other hand, false positives can also lead to undue worry or unnecessary additional interventions. Achieving separation-based fairness is crucial, as it involves actively assessing whether all subgroups, regardless of their definition, are equally benefiting from the predictive performance of a model.

Sufficiency

Sufficiency-based fairness requires that, given a model's prediction, the probability of the actual outcome is independent of the protected attribute. A positive prediction should have the same probability of being correct, regardless of the individual's background or subgroup.

This is formalized as *Predictive Parity*:

$$P\,(Y\ =\ 1 \mid D\ =\ 1, A=\ a_0\ )\ =\ P\,(Y\ =\ 1 \mid D\ =\ 1, A = a_1)\ \forall a_0, a_1\ \in A$$ [9]

Sufficiency is fundamental in risk assessment and diagnostic tools because both patients and doctors rely on their results to make informed decisions.

If a model is more effective at predicting disease risk for one group than another, it may not serve underserved communities well. Maintaining fair medical assessments can be challenging, particularly when different groups exhibit symptoms in distinct ways. Identical predictions do not always lead to the same valuable results. If a model is not calibrated correctly, it may lose its effectiveness and validity in healthcare

2.4.2 Individual fairness

Individual fairness implies that similar individuals should be treated similarly by predictive models. Unlike group fairness, which compares outcomes across predefined subgroups, individual fairness operates at a more granular level, demanding that a model treat similar cases consistently.

A formal approach to individual fairness is the *Fairness Through Awareness* (FTA) [9]: a similarity metric over individuals that constrains the difference in predictions to be no greater than their measured similarity. In its strongest form, this approach implies that *causal discrimination* is avoided: if two individuals differ only in a protected attribute, their predictions should remain the same.

The key challenge in applying this concept lies in defining what it means for two individuals to be *similar*. In healthcare, patients may share similar symptoms but vary in subtle clinical characteristics, comorbidities, or treatment histories that influence their assessment outcomes. Thus, constructing a suitable similarity measure is challenging and typically requires domain expertise.

In clinical decision support systems, a failure of individual fairness might result in two patients with nearly identical symptomology being given substantially different risk scores or diagnoses. This might occur if the model is using confounding variables or is overfitting trends based on data availability rather than actual clinical significance.

Individual fairness is particularly relevant when datasets contain heterogeneous symptom presentations, as is often the case in gynecology and chronic pain conditions. Consistency in managing such heterogeneity is required to maintain patient trust and fairness of decision-making procedures.

2.4.3 Causal fairness

Causal fairness considers the effect of protected features on model predictions in accordance with *causal relations*, rather than statistical correlations. In contrast to group or individual fairness, causal fairness questions whether a model's decision would change were a sensitive feature changed, everything else being equal. This is typically captured by *counterfactual fairness*, in which a model is considered fair if every individual receives the exact prediction in both the actual and a counterfactual world where the protected attribute is altered.

Beyond counterfactual fairness, several population-level causal fairness criteria have been proposed. These include:

- *Counterfactual parity*, which evaluates the influence of a protected attribute on population predictions.
- *Conditional counterfactual fairness*, which relaxes assumptions by conditioning on observable covariates.
- *Principal fairness,* which necessitates equal outcomes for individuals similarly impacted by the decision, irrespective of their group affiliation.

This framework is especially applicable to healthcare, where numerous attributes (e.g., previous diagnosis or treatment uptake) can reflect structural inequities rather than actual clinical need. However, these definitions all require a well-defined *causal model*, often represented as a *directed acyclic graph* (DAG).

In this case, during my internship at Hale, constructing a causal graph was considered to better disentangle the effects of confounding factors when the predictive models were not working correctly. However, the idea was later rejected: it would take substantial assumptions and domain knowledge to build a trustworthy causal graph. Moreover, creating these graphs is highly challenging, particularly in gynecology, where underlying mechanisms are less well understood and symptom overlap is significant. As it stands, although causal fairness is a beneficial tool for examining bias and confounding factors, its application is still constrained by considerations of feasibility in actual clinical datasets.

3. Data Analysis and Clustering

3.1 Introduction

In this chapter, latent structures were identified in the Hale dataset using *unsupervised learning algorithms*. Clustering techniques were employed to identify symptom-based subpopulations that may not align with existing diagnostic labels but still have relevant clinical implications. The analysis aimed to move beyond existing retrospective diagnosis labels that have been influenced by factors such as treatment effects, systemic selection bias, and underrepresentation.

Before clustering, an extensive preprocessing was conducted to improve reliability and reduce bias. Demographic outliers were filtered based on plausible biological parameters. The raw symptom data collected through the questionnaire were preprocessed via *one-hot encoding*, which converts each symptom item into a 0/1 vector (with a value of 1 indicating that the user selected the corresponding symptom).

Next, the binary symptom items were grouped into profiles based on the literature for each target diagnosis (e.g., Endometriosis, PCOS, Vulvodynia). A group score was calculated for each user by averaging each user's responses on the symptoms most characteristic of each diagnostic category. These scores indicate how closely that user's symptom profile resembled the clinical symptom patterns described in the literature for each condition. They were subsequently standardized using *Z-score normalization* to ensure comparability across groups.

Clustering was performed using the *K-means algorithm* on the standardized group scores. To visualize the clusters, Principal Component Analysis (PCA) was applied for dimensionality reduction, and *silhouette scores* were calculated to assess the internal cohesion of the clusters. To better understand the clinical relevance of each cluster, the distribution of self-reported diagnoses was overlaid, and *odds ratios* were calculated across clusters. Various demographic factors and contraceptive usage statistics were also analyzed for each group.

3.2 Target Variables Prevalence

The frequency at which the target diagnoses are present in the dataset demonstrates the traditional underrepresentation of gynecological conditions within clinical datasets, which is a significant hurdle for machine learning classification.

Prevalence of Target Diagnoses	
Endometriosis	5.76%
PCOS	17.36%
Vulvodynia	3.62%

Fibromyalgia	2.12%
IBS	7.83%
Vaginismus	2.75%
Adenomyosis	5.02%
Uterine Fibroids	4.35%

Table 1: Prevalence of Target Diagnoses

While PCOS presents a relatively high prevalence, all other conditions remain under the 10% threshold. Specifically, Fibromyalgia and Vaginismus are the least commonly reported, with only 2.12% and 3.62% prevalence.

The dataset exhibits a severe class imbalance, posing significant challenges for the performance of classification models. Thus, the imbalance must be addressed, and the appropriate evaluation metrics must be chosen to enable models to predict rarer conditions and to evaluate them accurately.

3.3 Studying Correlations

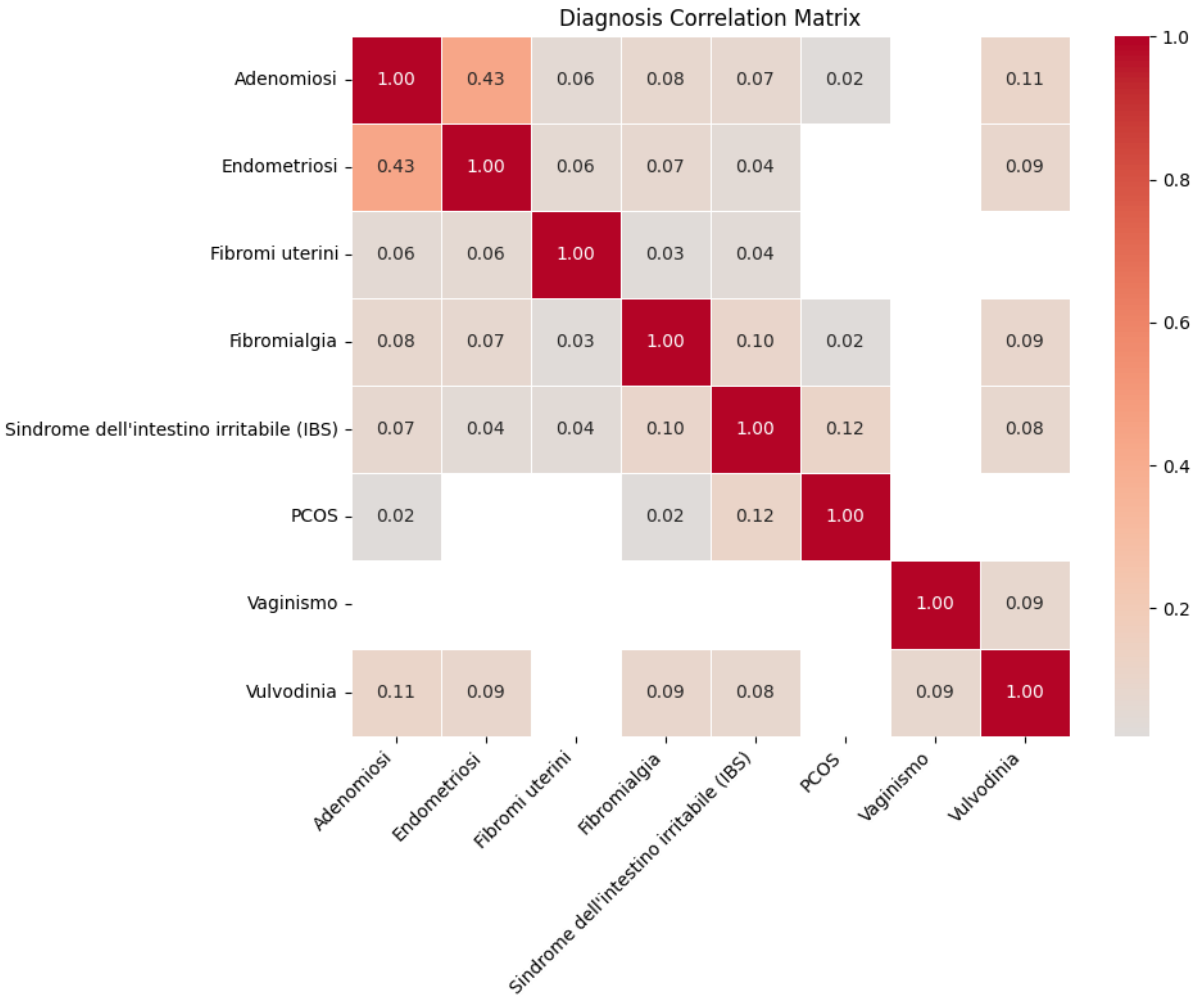

Figure 1: Diagnosis Correlation Matrix

The heatmap in Figure 1 displays the pairwise correlation coefficients for each of the target diagnoses in the dataset.

The strongest observed correlation is between Endometriosis and Adenomyosis (r = 0.43), which is clinically plausible given their shared gynecological nature and overlapping symptom profiles. Collectively, they suggest a tendency toward co-occurring diagnoses, possibly due to confounding symptomatic profiles or a diagnostic challenge due to their underlying mechanisms.

Another mild correlation is observed between PCOS and IBS (r = 0.12), which, even if weak, aligns with research that highlights the connections between hormonal imbalances and gastrointestinal symptoms.

However, beyond this, most correlations remain weak, even for conditions that are frequently comorbid in the medical literature, such as Fibromyalgia, Vulvodynia, and Vaginismus.

These weak correlations may arise because many respondents do not yet have a formal diagnosis and are using the questionnaire to better understand the cause of their symptoms.

3.4 Studying Correlations between Diagnoses and Symptoms

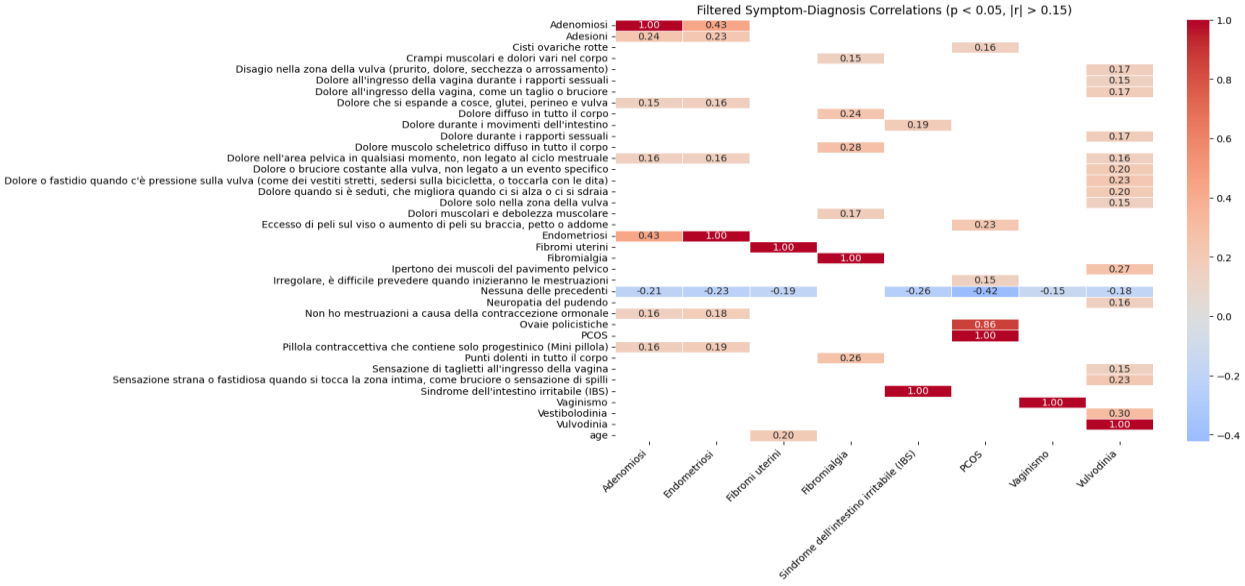

Figure 2: Correlations Diagnoses-Symptoms

The heatmap in Figure 2 displays significant correlations ($p < 0.05$, $|r| > 0.15$) between symptoms reported in the Hale questionnaire and the selected target conditions.

Vulvodynia exhibits the strongest overall correlations, particularly with localized vulvar pain features. Vestibular pain ("Vestibolodinia") correlates at r = 0.30, followed by pain or discomfort when there is pressure on the vulva ("Dolore o fastidio quando c'è pressione sulla vulva", r = 0.23), unusual or unpleasant sensation when touching the

intimate area (“Sensazione strana o fastidiosa quando si tocca la zona intima”, r = 0.23), and pain during sexual intercourse (“Dolore durante i rapporti sessuali”, r = 0.17). These values corroborate the diagnostic use of localized, provoked vulvar symptoms to identify Vulvodynia in self-reported screening situations.

Fibromyalgia, a systemic chronic pain condition, is associated with widespread musculoskeletal symptoms. It shows strong correlations with diffuse musculoskeletal pain (“Dolore muscolo scheletrico diffuso in tutto il corpo”, r = 0.28), tender points throughout the body (“Punti dolenti in tutto il corpo”, r = 0.26), and muscle pain and weakness (“Dolori muscolari e debolezza muscolare”, r = 0.17). These align with diagnostic criteria.

Both Endometriosis and Adenomyosis are most correlated with other symptoms of diffuse, persistent pelvic pain. Pelvic pain that radiates to the thighs, glutes, perineum, and vulva ("Dolore che si espande a cosce, glutei, perineo e vulva") shows correlations with Endometriosis with r = 0.16 and Adenomyosis with r = 0.15. Similarly, pelvic pain occurring at any time, unrelated to the menstrual cycle ("Dolore nell'area pelvica in qualsiasi momento, non legato al ciclo mestruale") correlates with Endometriosis of r = 0.16 and Adenomyosis of r = 0.16. These findings also illustrate the challenge of making clinically actionable distinctions between these two conditions, as highlighted by their correlation with each other (r = 0.43). Notably, progestin-only contraceptive use (“Pillola contraccettiva che contiene solo progestinico) and *amenorrhea* (lack of menstrual cycle) due to hormonal contraception (“Non ho mestruazioni a causa della contraccezione ormonale”) are also positively correlated, which reflects retrospective control of symptoms following the diagnosis.

PCOS is correlated with multiple hallmark symptoms, especially excessive body or face hair (“Eccesso di peli sul viso o aumento di peli su braccia, petto o addome”, r = 0.23) and irregular menstruation (“Irregolare, è difficile prevedere quando inizieranno le mestruazioni”, r = 0.15). These associations reflect everyday experiences of abnormal androgen excess and menstrual dysfunction reported by individuals with PCOS.

IBS is related to pain on bowel movement (“Dolore durante i movimenti dell’intestino”) with r = 0.19, as would be expected from its conceptualization as a functional

gastrointestinal disorder. Although fewer symptoms achieve the correlation criterion for this condition, the correlation confirms its relevance within the context of overlapping pelvic and intestinal pain.

Several conditions, such as Vaginismus, show minimal individual associations within this symptom matrix. This is understandable, as the prevalence of Vaginismus is extremely low in the dataset. Still, the negative correlation with symptom absence (r = -0.15) and conceptual overlap with other diagnoses (like Vulvodynia) support its development for future subgroup or interaction modeling.

3.5 Cluster Analysis

In this section, unsupervised clustering techniques are applied to the dataset to identify latent subgroups based purely on self-reported gynecological and chronic pain symptoms. Given the extensive number of symptom variables in the questionnaire, for every target condition, a symptom intensity score is calculated as the average of the symptoms most associated with such condition reported by the user. This enables the compression of individual symptom profiles into interpretable scores that reflect the burden of symptoms across conditions. A comprehensive list of the symptom items used for each condition is available in Appendix B: Clustering Symptom List.

To determine the most appropriate number of clusters, *silhouette analysis* is implemented, evaluating both the cohesion within clusters and their separation from one another. The silhouette score for a data point *i* is computed as:

$$s(i) = \frac{max(a(i), b(i))}{b(i) - a(i)}$$

a(i) = the average intra-cluster distance
b(i) = the lowest average distance to points in a different cluster

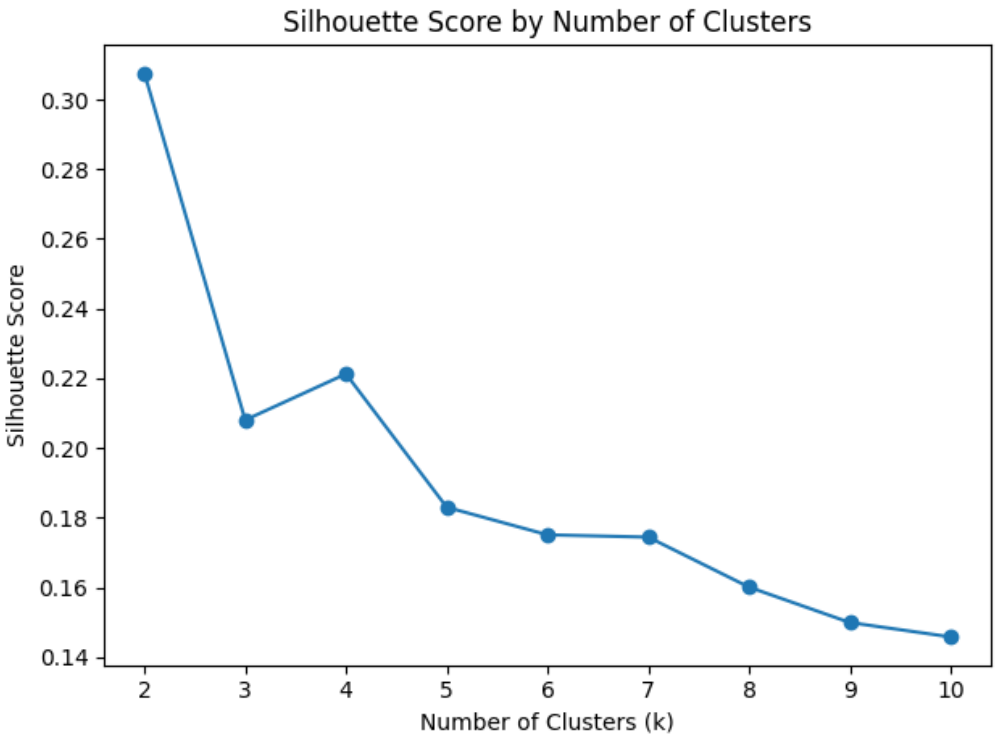

Figure 3: Silhouette Score Analysis

The optimal number of clusters is selected based on the silhouette analysis displayed in Figure 3. The silhouette score evaluates the internal consistency of clusters by balancing intra-cluster cohesion and inter-cluster separation. Although the highest verification score was for k = 2, the cluster designations were excessively coarse and did not capture sufficient clinical variability.

The score for k = 4, instead, provides a local maximum and better balances interpretability and resolution, with a relatively strong score on the silhouette score (0.22). Therefore, k = 4 is selected as the best clustering solution.

The resulting four clusters also demonstrate a meaningful distribution of users: Cluster 0 contains 3,416 users, Cluster 1 includes 1,660 users, Cluster 2 has 4,198 users, and Cluster 3 accounts for 2,290 users. This distribution shows that there is no cluster dominating the dataset: each subgroup has a substantial portion of the population.

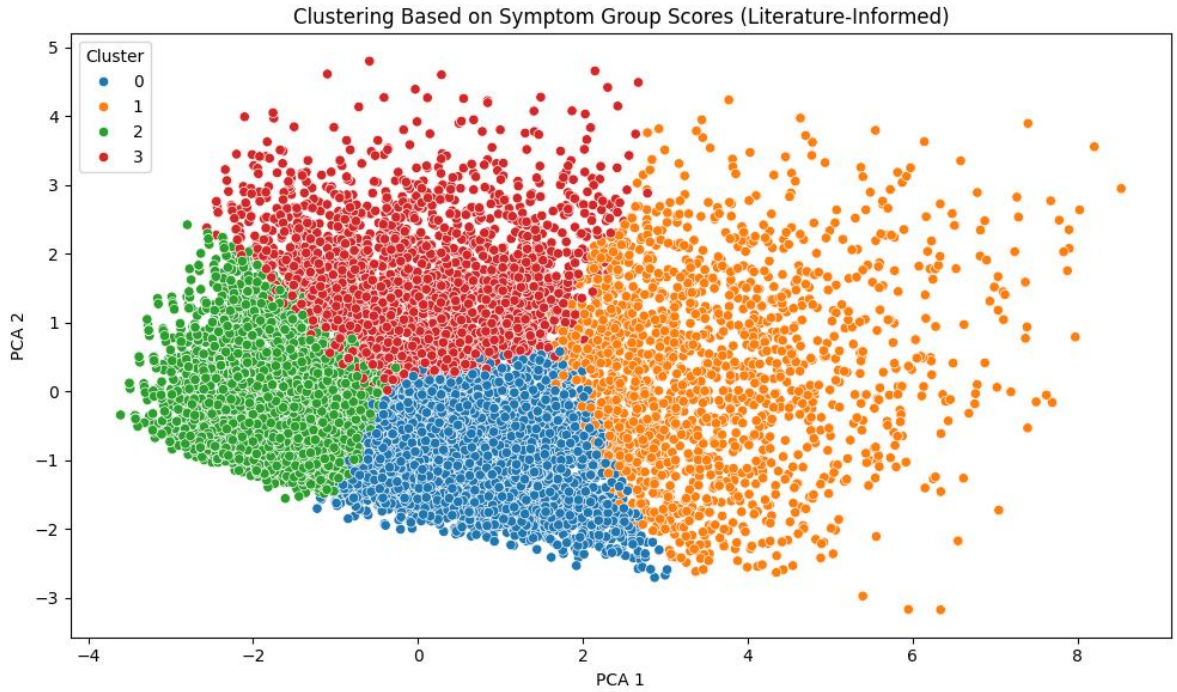

Figure 4: PCA Projection of Clusters

In Figure 4, the PCA projection of the four-cluster solution reveals that the clusters partially overlap rather than being distinctly separate. The overlap suggests that the identified subgroups may not fully capture distinct underlying pathologies; instead, they likely represent symptomatic gradients or mixed patterns of disease. Some structure is nonetheless observed: cluster centers diverge, and local clusters can be seen with relative clarity. This supports the idea that, even within overlapping clinical areas, the arrangement of symptom data can systematically highlight or hide subpopulations. This may contribute to clinical blind spots, particularly in the field of gynecological health, where the language of diagnostic criteria can be vague and subject to debate.

3.6 Diagnosis Prevalence Across Symptom-Based Clusters

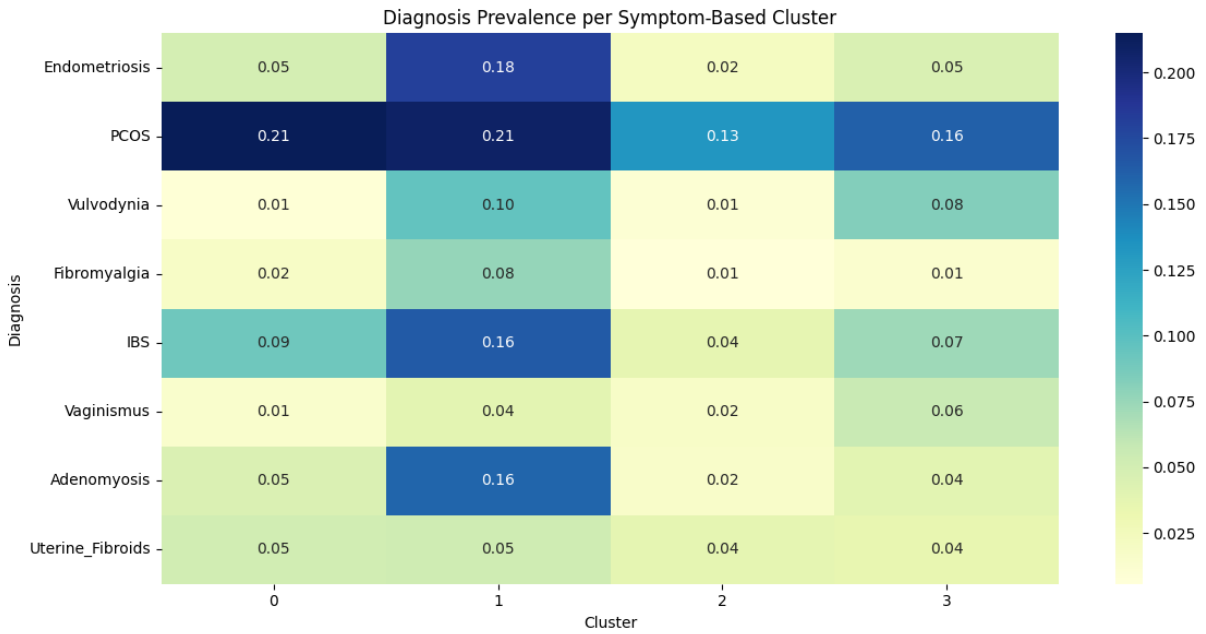

Figure 5: Diagnoses Prevalence per Cluster

In Figure 5, the distribution of the diagnoses is examined across the four symptom-based clusters found in the unsupervised clustering analysis. The degree to which symptom-based groupings correspond to traditional medical categories can be assessed by comparing the prevalence of each diagnosis within these clusters, potentially revealing underrecognized patterns.

Cluster 0 shows elevated prevalence of PCOS (21%) and IBS (9%), suggesting that this cluster captures hormonal imbalance and gastrointestinal discomfort, indicating a mix of metabolic and sensory dysfunctions.

Cluster 1 is the most diagnostically concentrated, with PCOS (21%), Endometriosis (18%), IBS (17%), and Adenomyosis (16%) as dominant pathologies. It represents the most severe form of multi-morbid chronic pain syndromes, where symptom reporting is intense and overlaps heavily with established diagnostic labels.

Cluster 2 shows a low prevalence for most conditions, except for PCOS (13%). This suggests that this cluster may represent a subclinical symptom profile or a diagnostic blind spot where symptom severity is insufficient for clinical recognition.

Cluster 3 is characterized by PCOS (16%), Vulvodynia (8%), and Vaginismus (6%), indicating that this subgroup primarily reflects sensory and neuromuscular pain, with a focus on vulvar sensitivity and penetration pain.

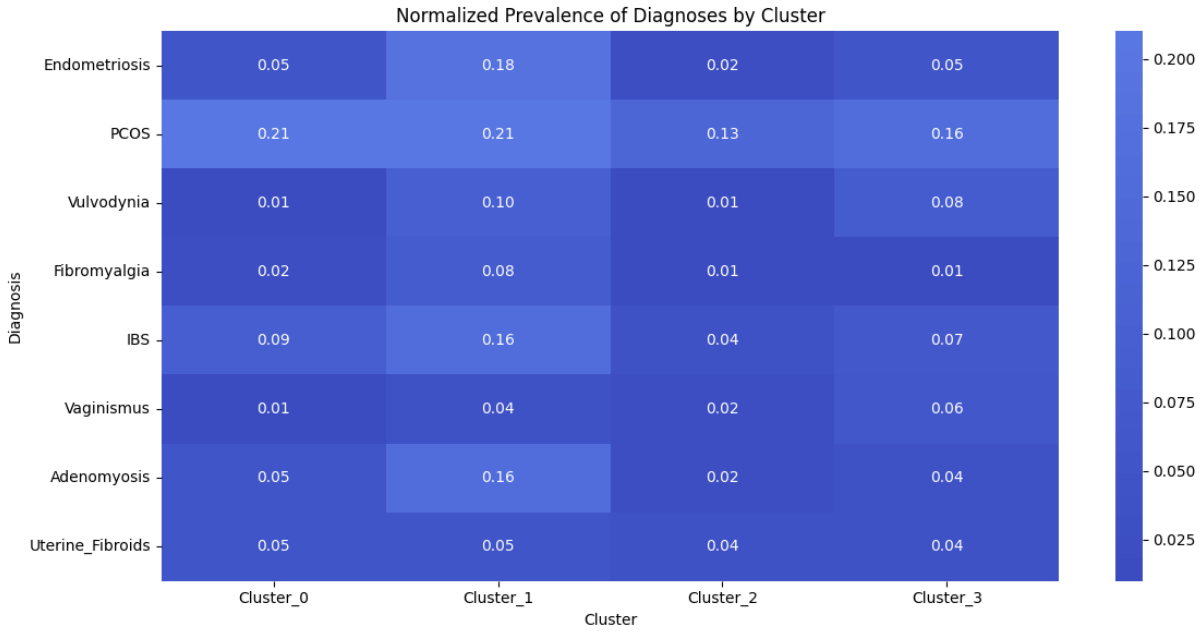

Figure 6: Normalized Diagnoses Prevalence per Cluster

The heatmap in Figure 6 shows the normalized prevalence of diagnoses, calculated as the proportion of users within each cluster who reported a given condition. The insights reaffirm the prior understanding.

Cluster 1 surfaces as the subgroup with the highest diagnostic density, characterized by a relatively high prevalence of Endometriosis (18%), Adenomyosis (16%), Vulvodynia (10%), and IBS (16%). Cluster 2 appears to have the lowest diagnosis rates overall, suggesting either a less severe symptom presentation or a population that has not yet been formally diagnosed. It is evident that although PCOS is distributed across all clusters and is highly variable, it is most prevalent in Clusters 0 and 1 (21% for both), which highlights the significant overlap of PCOS symptoms across different sub-populations.

3.7 Adjusted Odds Ratio per Diagnosis Across Clusters

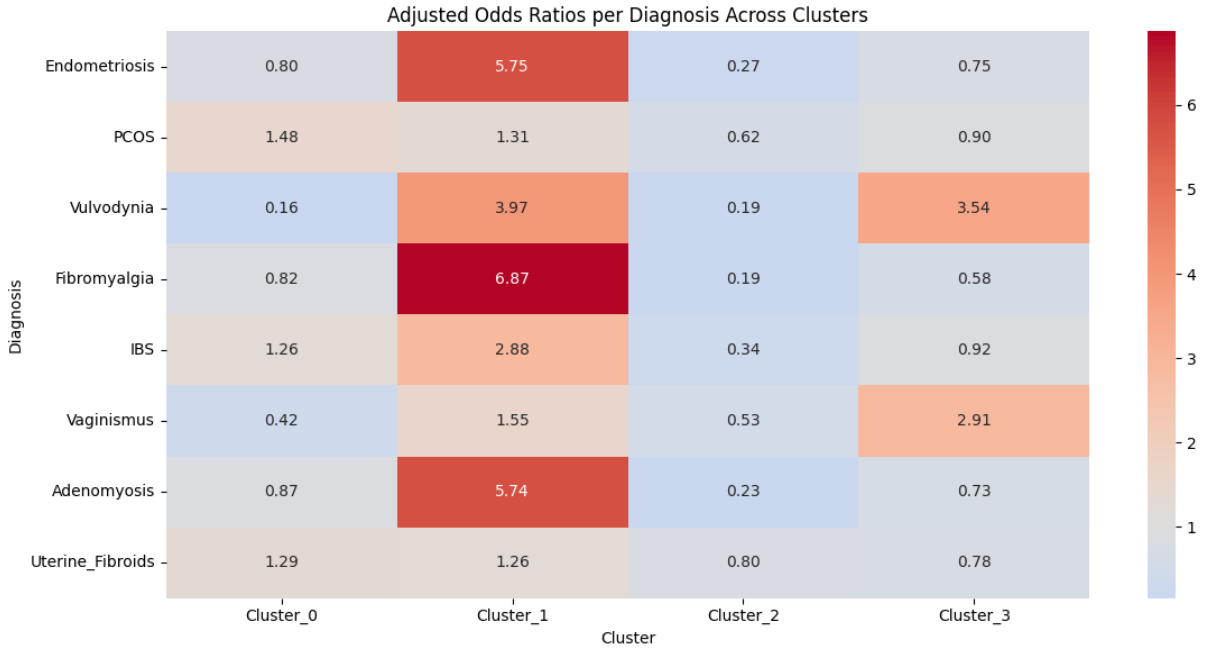

Figure 7: Adjusted Odds Ratios per Diagnosis

Figure 7 illustrates the *adjusted odds ratios* (ORs) for gynecological and chronic pain diagnoses across the four symptom-based clusters. This ratio compares the odds of a diagnosis occurring within a cluster to the odds of it occurring outside the cluster. An odds ratio (OR) above 1 indicates a higher likelihood of diagnosis in the cluster, while below 1 suggests a lower likelihood.

The odds ratio for a diagnosis *d* in cluster c is calculated as:

$$OR_{c,d} = \frac{a * d}{b * c}$$

where:

- a = number of individuals in cluster *c* with diagnosis *d*
- b = number of individuals in cluster *c* without diagnosis *d*
- c = number of individuals outside cluster *c* with diagnosis *d*
- d = number of individuals outside cluster *c* without diagnosis *d*

As expected, Cluster 1 exhibits the greatest degree of enrichment, given significant odds for most conditions, particularly for Fibromyalgia (OR = 6.87), Endometriosis (OR = 5.75), and Adenomyosis (OR = 5.74). This is consistent with the high symptom burden and visibility discussed previously.

Cluster 3, however, has a narrower profile with increased odds for Vulvodynia (OR = 3.54) and Vaginismus (OR = 2.91), revealing a localized pain type.

Cluster 2 has a very low prevalence among the diagnoses, thus strengthening the diagnostic blind spot hypothesis.

Cluster 0 has relatively moderate odds for PCOS and IBS, but with less specificity than in Clusters 1 and 3.

3.8 Most Frequent Symptoms per Cluster

The analysis of the most frequent symptoms per cluster reveals the primary drivers behind the clustering. The average intensity of each symptom was calculated by taking its mean value across each cluster, reflecting how often or intensely it was reported. Ranking symptoms based on their intensity within each cluster may explain why particular pathologies are more prevalent and help detect potential misalignments with OR and normalized prevalence maps.

3.8.1 Cluster 0: Metabolic and Pelvic Symptoms

1. Feeling tired or fatigued
2. Abdominal bloating
3. I do not use contraceptives
4. Persistent pelvic pain or cramps
5. Back or leg pain
6. Menstrual cycles that are mostly regular but occasionally early or late
7. Severe menstrual pain (dysmenorrhea)
8. Difficulty concentrating
9. Sleep disturbances
10. None of the above

This cluster is dominated by back and pelvic pain, bloating, and fatigue. The lack of contraceptive use might suggest a subpopulation of women not currently managing symptoms hormonally, potentially indicating unmanaged PCOS symptoms. The strong presence of pelvic cramps and bloating aligns with conditions like PCOS and IBS, which were shown to be the most prevalent.

3.8.2 Cluster 1: Severe Chronic Pain and Systemic Symptoms

1. Feeling tired or fatigued
2. Abdominal bloating
3. Back or leg pain
4. Persistent pelvic pain or cramps (on most days of the month)
5. Difficulty concentrating
6. Pain related to sexual activity
7. Severe menstrual pain
8. Muscle pain and weakness
9. Sleep disturbances
10. Deep pelvic pain during penetrative intercourse

This cluster is pain-dominant, with high reporting of chronic pelvic pain, severe *dysmenorrhea* (Intense menstrual cramps), *dyspareunia* (pain during sexual intercourse), and musculoskeletal symptoms, often paired with fatigue and cognitive burden.

3.8.3 Cluster 2: Non-Specific and Unlabeled

1. None of the previous situations
2. I do not use contraceptives
3. None of the above
4. None of the above
5. Menstrual cycles that are mostly regular but occasionally early or late
6. None of the above
7. None of the above
8. None of the previous answers
9. Very regular and predictable menstrual cycles
10. None of the above

Cluster 2 exhibits a high endorsement of unspecified responses ("Nessuno delle precedenti") across various key questionnaire items. People in this group usually report not having formal gynecological diagnoses and not using hormonal contraceptives.

This suggests that Cluster 2 is capturing an *underserved population*: a group of users who do not have overt or diagnosable symptoms and may be experiencing subclinical discomfort. The fact that they are still engaged with the screening tool, however, illustrates a potential diagnostic blind spot: individuals who do not meet the standard symptom criteria but still seek answers.

3.8.4 Cluster 3: Surface and Penetrative Pain

1. Pain related to sexual activity
2. Pain at the vaginal entrance during intercourse
3. Pain at the vaginal entrance, described as a cut or burning sensation
4. Feeling tired or fatigued
5. Vulvar pain triggered by touch or penetration
6. None of the previous situations
7. Pain during sexual intercourse
8. Discomfort in the vulvar area (itching, dryness, redness)
9. Abdominal bloating
10. Pain during penetrative sex, described as "hitting a wall"

Cluster 3 is predominantly characterized by surface-level pain and vulvar discomfort, especially during sexual activity. The prominence of pain during intercourse and vulvar burning is consistent with the high ORs for Vulvodynia and Vaginismus seen earlier. The presence of fatigue may indicate the presence of overlapping neuropathic pain. This symptom profile matches well with surface neuromuscular sensitivities rather than deep gynecological pathologies.

3.9 Demographic and Anthropometric Analysis of Clusters

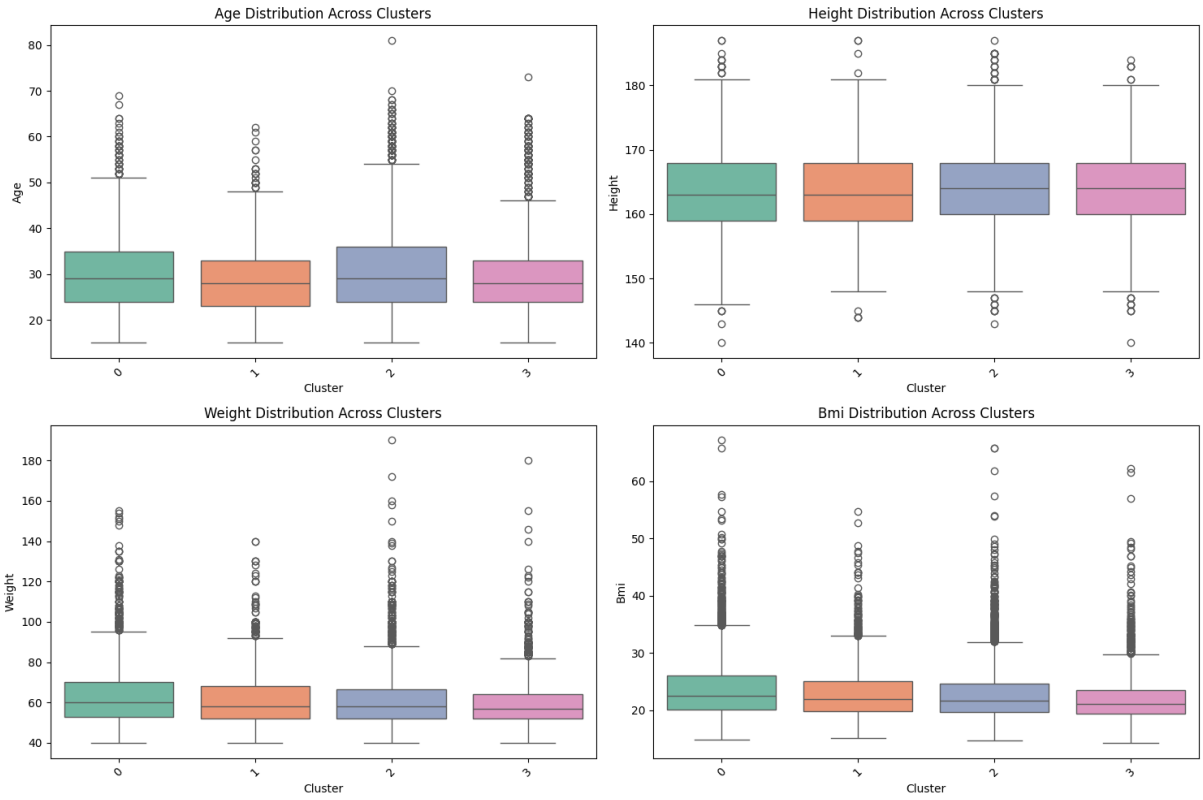

Figure 8: Demographic and Anthropometric Data per Cluster

Figure 8 shows the distributions of age, height, weight, and median body mass index (BMI) between the four symptom clusters. While there are no extreme differences, some patterns emerge that relate to the diagnostic profiles identified earlier.

As shown above, Cluster 1 comprised users with the most significant prevalence of Endometriosis, Adenomyosis and Fibromyalgia symptoms. This cluster tends to have a younger user age and a slightly lower BMI. This may reflect the earlier onset of these chronic pain conditions, which are commonly seen in women of reproductive age.

Cluster 0, characterized by a higher prevalence of PCOS, has a moderately higher BMI and weight. This is supported by the metabolic and endocrine characteristics of PCOS, such as weight gain and insulin sensitivity, which often arise in higher BMI profiles.

Cluster 2, characterized by minimal diagnostic reporting, encompasses a broader range of user ages and has the highest median age of users.

Cluster 3, enriched for Vulvodynia and Vaginismus user types, shows relatively younger users and overall lower BMI. This is consistent with conditions that often first appear early in the sexual or reproductive history and are not as associated with a systemic or metabolic condition.

3.10 Contraceptive Use Across Clusters

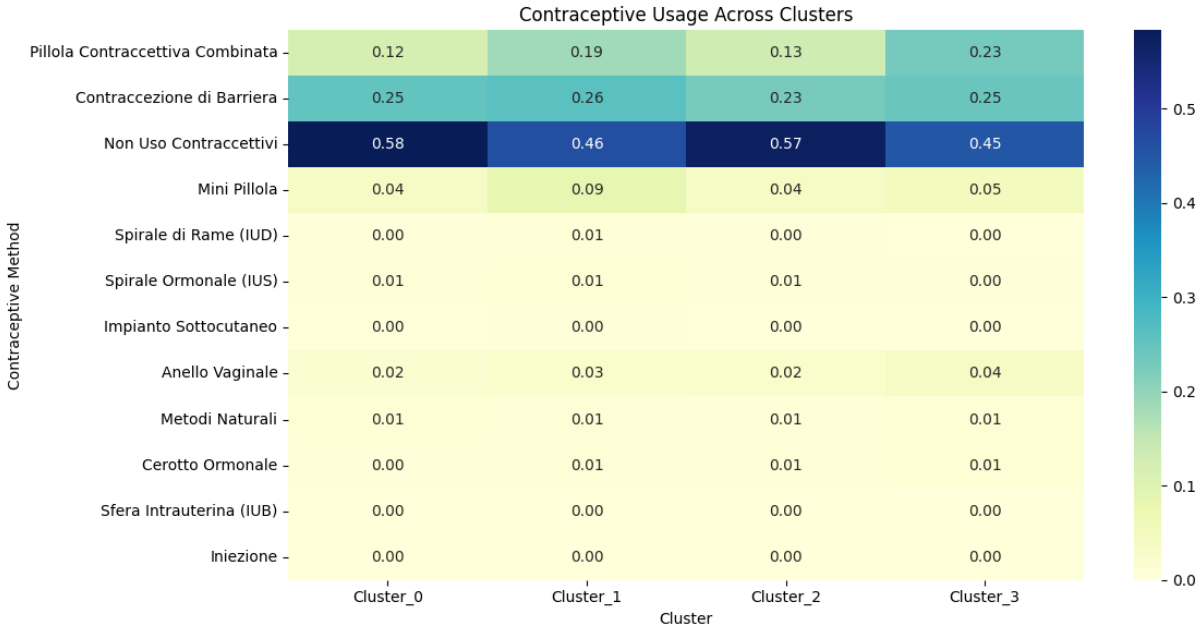

Figure 9: Contraceptive Usage Across Clusters

The heatmap in Figure 9 illustrates the distribution of contraceptive use across clusters.

Cluster 0, comprising metabolic and pelvic symptoms, illustrates a high level of no contraceptive use ("Non Uso Contraccettivi"), which denotes uncontrolled symptomatology that is typical in cases like PCOS and IBS, where control of hormones is less frequent.

Cluster 1, with severe chronic pain and systemic symptoms, also exhibits a slightly more evenly spread use of contraceptive methods, with evident use of combined contraceptive pill (“Pillola Contraccettiva Combinata”), possibly a sign of targeted hormonal treatment for symptom management.

Cluster 2, with non-specific symptoms and lack of formal diagnoses, also exhibits high non-contraceptive use, possibly a sign of the lack of medical intervention or detection.

Cluster 3, which is predominantly defined by surface and penetrative pain associated with Vulvodynia and Vaginismus, has the highest usage of the combined contraceptive pill ("Pillola Contraccettiva Combinata"), indicating attempts at hormonal management for symptom control.

4. Classification Models

4.1 Introduction

This chapter investigates whether insights gained from clustering users by symptom profiles can enhance the performance of classification models in identifying chronic gynecological conditions.

The initial stage of the analysis examines whether the distance between a user and their cluster centroid conveys clinically relevant or informative patterns. The study focuses on whether closeness to the centroid indicates the severity of a symptom profile. The relationship between a user's distance from their cluster centroid and the likelihood of having a diagnosis is explored. While it would be expected that centroids would represent "prototypical cases", the findings suggest the opposite. In most clusters, users farther away from the centroid are more likely to report a formal diagnosis. This can be better understood by reviewing how the clusters were constructed; the subsequent analysis will help clarify the underlying rationale.

The analysis continues by analyzing the results of a global classification model trained on the entire dataset. Subsequently, classification models are trained independently within each cluster for the most prevalent conditions. These models are then evaluated on the full dataset to explore their generalizability across subpopulations.

Notably, incorporating each user's distance from their cluster centroid as an additional feature improves model performance. Furthermore, *Feature Importance* and *SHAP values* analysis reveal that this distance is consistently among the most informative predictors, indicating its predictive power.

The complete tables of performance metrics and fairness evaluations for each condition and model is available in Appendix C: Model Performance Tables. The SHAP values summary plots are in Appendix D: SHAP Values.

Overall, however, the performance of the classification models remains limited. Additional techniques were also tested, including *SMOTE oversampling* to balance minority classes, *positive-unlabeled (PU) learning* to create a model capable of identifying undiagnosed cases, and extensive *hyperparameter tuning*. None of these approaches, however, led to substantial performance improvements.

4.2 Centroid Distance Analysis

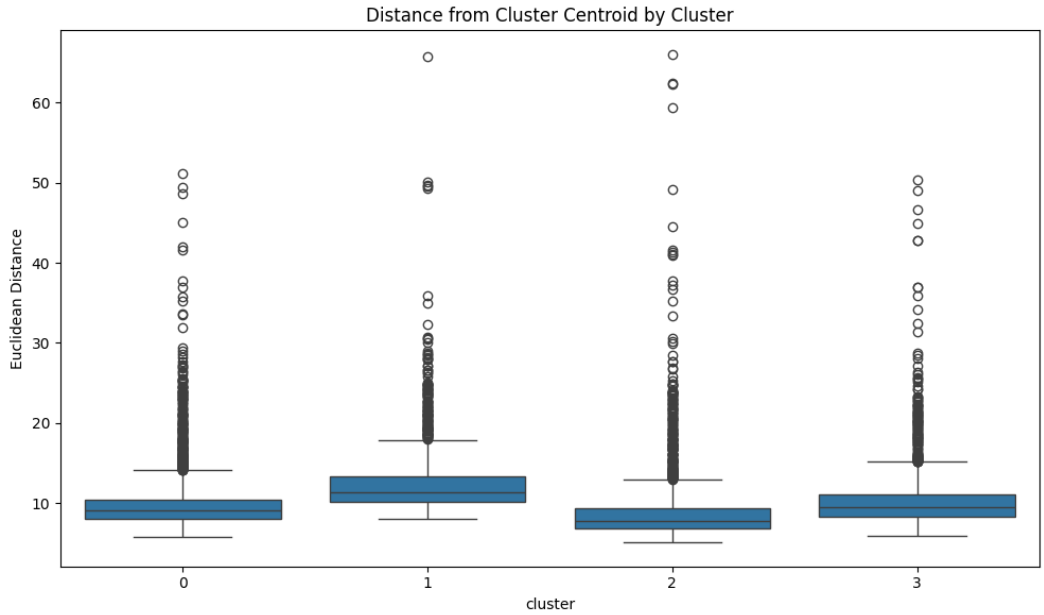

Figure 10: Distance from Centroid per Cluster

Figure 10 illustrates how Euclidean distances are distributed from the cluster centroid. All four clusters have right-skewed distributions, with a dense concentration of users around the centroids, but a relevant number of users extending into the upper tail.

These distributions reflect the outcome of the clustering procedure used. Based on the average condition scores, the centroids represent a *blended profile*: a user whose symptom burden is moderate and whose condition scores are neither extreme nor exclusive to any diagnosis. In contrast, users located far from the centroid exhibit more

sharply defined symptom patterns, with higher intensity for one specific condition and low or no presence for the others.

This represents a fascinating insight: distance from the centroid does not merely indicate *atypicality* or noise. It instead shows greater *symptom specificity*: the user deviates from the average multi-condition centroid and presents a more explicit alignment with a single diagnostic profile. This interpretation is supported by later sections, which show a positive relationship between distance from the centroid and a positive diagnosis.

4.3 Diagnoses Prevalence by Distance

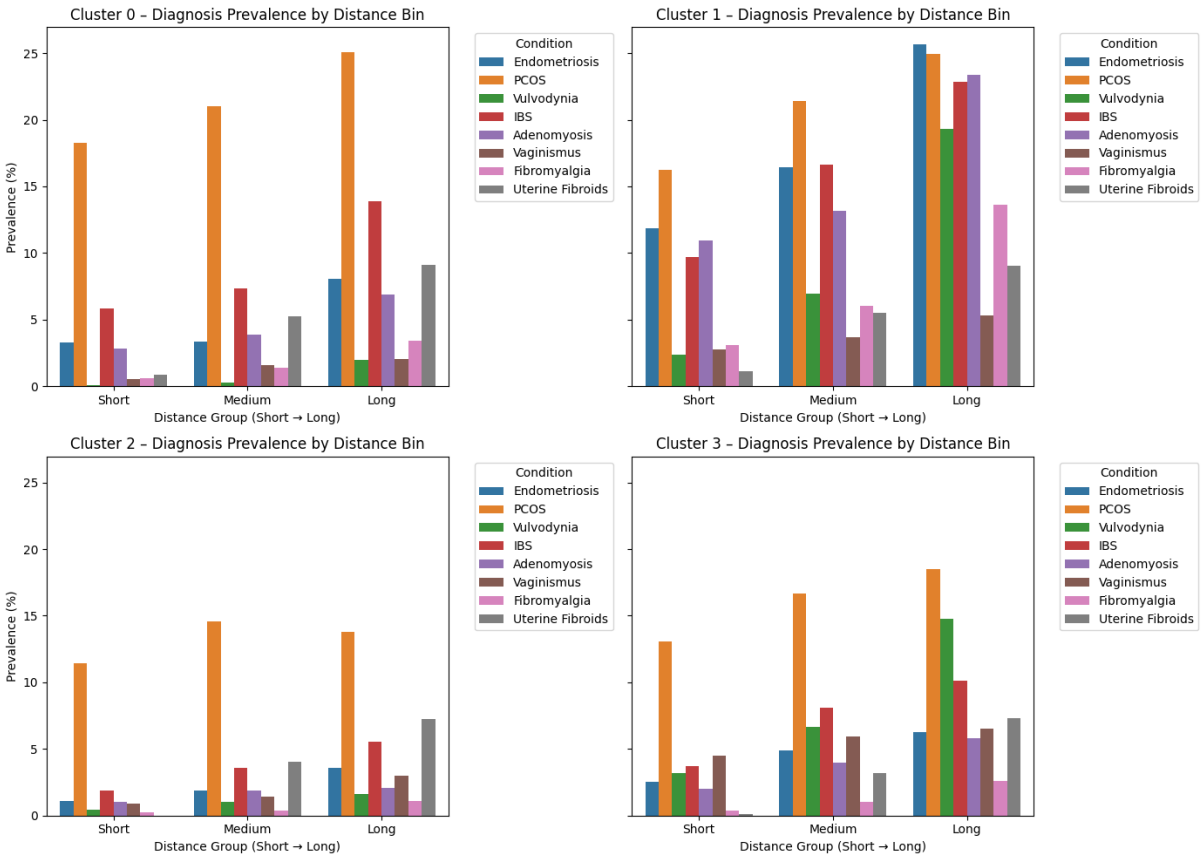

Figure 11: Diagnoses Prevalence by Distance per Cluster

In Figure 11 the prevalence of diagnoses in the clusters is examined across three distance-based user groups (short, medium, and long) for each cluster. The pattern is evident: users who are more distant from the center report diagnoses more frequently.

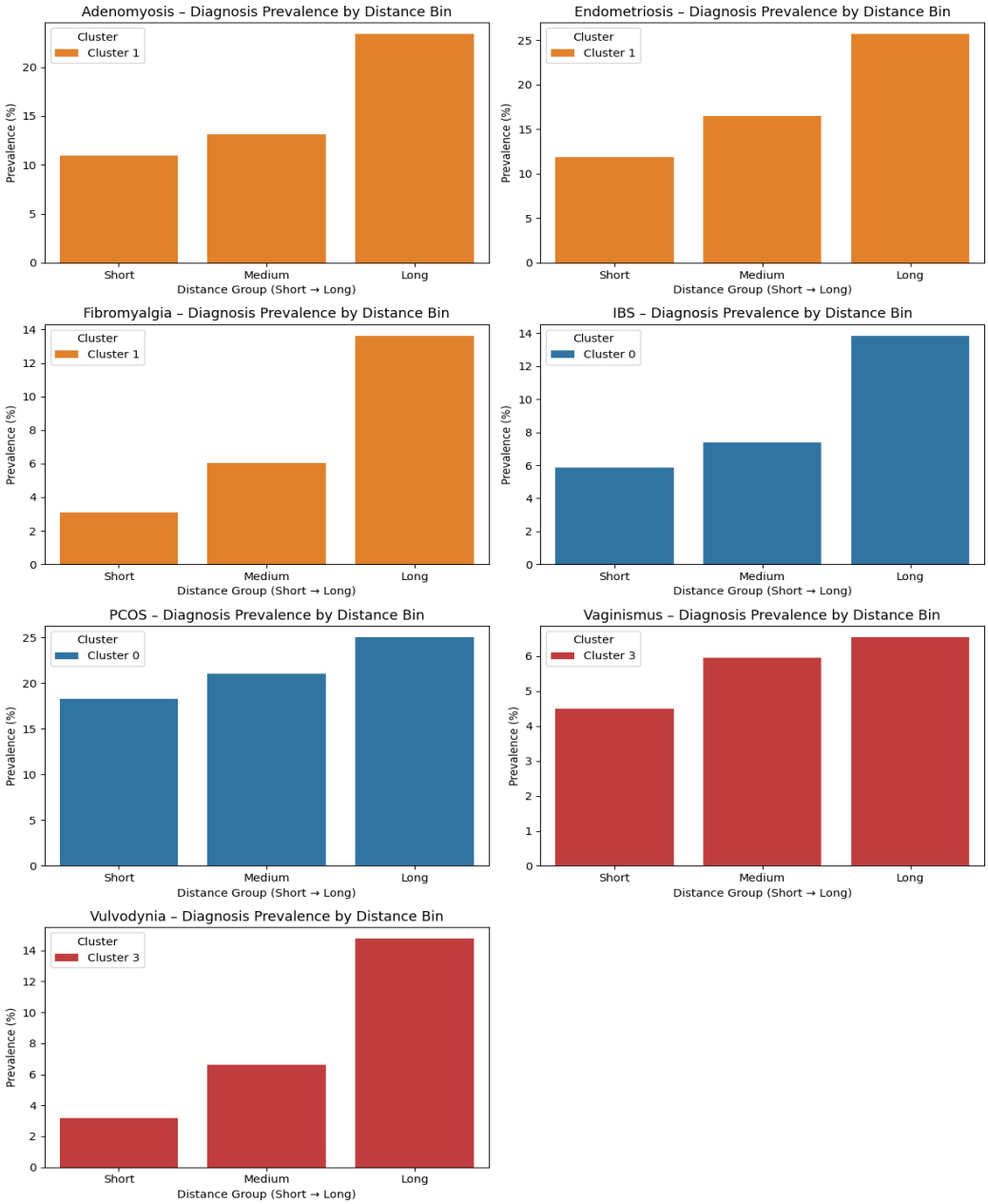

Figure 12: Focus on Most Prevalent Conditions

This finding is reinforced by Figure 12, which focuses on the most prevalent diagnoses within each cluster. Cluster 2 is excluded from this analysis due to insufficient prevalence of any single diagnosis to justify focused evaluation.

Across all clusters shown, a consistent trend emerges: diagnosis rates increase with greater distance from the cluster centroid. For instance, the likelihood of a PCOS diagnosis in Cluster 0 or Endometriosis in Cluster 1 rises steadily from the short-distance group to the long-distance group. Similar patterns are observed for Fibromyalgia, Vulvodynia, and other conditions within their respective clusters.

This confirms that distance does not simply reflect noise or deviation but rather indicates meaningful variation in how a user's symptoms align with a clinical category: it is a *clinically informative signal*.

4.4 Directional Symptom Profiles in PCA Space

The principal component analysis (PCA) projection in Figure 4 illustrates that the previously identified clusters are not symmetrical, exhibiting irregular and elongated shapes: different regions within the same cluster may represent different symptom subtypes or diagnostic expressions. To determine greater granularity in the clusters, the five most extreme users along each axis were identified within each cluster, and their shared symptoms were analyzed. The complete list of common symptoms identified is reported in Appendix B.

In Cluster 0, characterized by metabolic and pelvic symptoms, the positive end of PC1 is characterized by menstrual irregularities, pelvic cramping, nausea, and acne. This reflects PCOS or hormonal dysregulation profiles. The negative end of PC1 is characterized by fatigue, pain during penetration, and recurrent UTIs, indicating a vulvovaginal irritation pattern. The positive end of PC2 shows more widespread musculoskeletal symptoms, which align more closely with Fibromyalgia, whereas the negative end of PC2 centers on deep dyspareunia and abdominal bloating.

Cluster 1 is highly internally heterogeneous. The negative end of PC1 is dominated by dysmenorrhea, pelvic pain, and fatigue: symptoms associated with Adenomyosis and Endometriosis. The positive end of PC2 shows symptoms common in Fibromyalgia and other centrally sensitized pain conditions: cognitive problems, systemic fatigue, and muscle cramps in the whole body. The positive end of PC1 and the negative end of PC2 are enriched with vulvar symptoms: burning, pain caused by superficial touch, and deep pelvic pain, highly indicative of Vulvodynia.

Cluster 2 remains weakly structured and diagnostically vague in all PCA directions, with no consistent symptom profile.

Cluster 3 develops along a penetrative pain axis: the negative end of PC1 features users reporting muscular guarding, fear of penetration, and inability to undergo exams (symptoms related to Vaginismus); whereas, at the positive end of PC1, users suffer from tactile hypersensitivity: touch, tight clothing, clitoral pain (symptoms associated with Vulvodynia). Along PC2, both ends are marked by pain during sex, burning, and dyspareunia, suggesting a mixed presentation of inflammatory irritation.

The analysis reveals some localized tendencies (Fibromyalgia-related symptoms along PC2 or Vulvodynia and Vaginismus at opposing ends of PC1). Still, it shows high heterogeneity, diagnostic ambiguity, and how different conditions may partially overlap in this symptom space. This undermines the original hope of identifying a clearly separable area for each diagnosis. In later experiments, an attempt was made to explicitly model decision boundaries within the PCA space, using the *Support Vector Machine* algorithm. However, the classification performance remained poor, as decision boundaries failed to separate diagnosed from undiagnosed users precisely.

Spatial classification is unsuitable because of the overlapping and diffuse nature of the symptom patterns in the dataset, underscoring the need for alternative modeling strategies.

4.5 Classification Model on the Global Dataset

In this section, the results of training and testing a classification model on the global dataset (without accounting for the symptoms cluster structure) are assessed. This serves to establish a baseline for performance and fairness across the user population and to quantify the trade-offs of not tailoring the model to subgroup-specific symptom patterns.

4.5.1 Modelling Choices: Handling Confounding and Low-Signal Variables

Some features and conditions were excluded from the final classification models due to clinical irrelevance or methodological concerns.

The “Uterine fibroids” variable was excluded from the analysis due to its low prevalence (4.35% in the entire dataset) and its limited symptom-based detectability. In clinical practice, fibroids are typically diagnosed using imaging techniques (ultrasound or MRI), since they are often asymptomatic or present with nonspecific symptoms, which overlap with most other conditions. Consequently, uterine fibroids are an unreliable target for prediction within a self-reported symptoms dataset, as it would probably add noise and undermine the model's clinical validity.

Variables related to contraceptive use were also excluded from the dataset. Initially, they were retained, but the models identified them as disproportionately influential variables. The diagnoses correlated more strongly with contraceptive use than with the symptoms. This reflects the retrospective treatment mentioned above: many users may already be taking hormonal contraception, as it is commonly prescribed to manage symptoms of different conditions (such as pain during menstruation, inconsistent cycles, or pelvic discomfort). Therefore, the inclusion of those variables introduced a *spurious association* between treatment and diagnosis. The models evaluated were learning patterns such as “taking contraceptive pill = having a condition”, biasing the model toward identifying *treated* cases rather than *undiagnosed* ones.

4.5.2 Model Architecture and Evaluation Metrics

Several classification models were evaluated, such as *Logistic Regression*, *Random Forest*, *XGBoost*, *Support Vector Machines*, and *LightGBM*.

The models were trained on a dataset split into 70% for training and 30% for validation. Model performance was assessed using two key metrics: the *area under the precision-recall curve* (PR-AUC) and the *F1 score*.

A metric frequently used for evaluating ML models is the *ROC-AUC* (*Receiver Operating Characteristic - Area Under the Curve),* which measures the model's ability to differentiate between positive and negative classes across all classification thresholds. However, this metric can remain artificially high even when the model performs poorly on the minority class; in contrast, the *PR-AUC plots* precision against recall, focusing specifically on the model's performance on the positive class. Similarly, the F1 score is the *harmonic mean* of precision (the proportion of true positives among

all positive predictions) and recall (the proportion of true positives identified among all actual positives) for the positive class. Particular emphasis was placed on *recall*, an essential metric for a screening tool, whose goal is to flag individuals who may require further medical evaluation.

Across nearly all conditions, *CatBoostClassifier* outperformed the alternatives in terms of stability, recall for the positive class, and PR-AUC. The other models did not yield satisfactory results due to the complexity of symptom overlap, high dimensionality, and severe class imbalance.

4.5.2 PU Learning Attempt and Challenges

Positive-unlabeled (PU) learning was also implemented to account for the large number of users without formal diagnoses. This algorithm attempts to identify the negative class by learning from labeled positives and treating the remaining unlabeled data as a mix of positives and potential negatives. However, the model struggled to identify any meaningful negatives: the underlying data is too ambiguous, with symptom overlap making it difficult to infer which unlabeled cases truly lacked the condition. Thus, the results are not presented, since the performance was too poor.

4.5.3 CatBoost

CatBoost is a gradient boosting algorithm that builds models by combining many weak learners: *symmetric oblivious decision trees*. These are trees where each internal node at the same depth uses the same feature and threshold as a splitting condition, making the model fast, memory-efficient, and less prone to overfitting, due to the reduced variance in the tree structure.

However, the key innovation in CatBoost is its use of *ordered boosting*, which constructs training permutations and causally computes residuals by ensuring that each observation is influenced only by the examples that precede it in the permutation. This addresses the issue of subtle target leakage in traditional boosting, where residuals are computed on the same data used to fit the previous models.

The model additionally addresses class imbalance, crucial in this case, by adjusting for skewed label distributions without the need for manual reweighting.

4.5.4 Performance Evaluation

This section evaluates the performance of the global classification model and assesses fairness across the symptom-based clusters that have been identified. As explained in Chapter 2, fairness is a crucial consideration in healthcare machine learning models, especially in fields where diagnostic uncertainty and symptom diversity are inherent to clinical practice. In this case, symptom-based clusters serve as proxies for latent subpopulations with distinct clinical trajectories, symptom expressions, and diagnostic likelihoods. Each cluster represents a distinct symptom reporting pattern, likely reflecting differences in disease severity, comorbidities, or, most importantly, access to care.

As such, evaluating model performance across these clusters enables assessing whether the classifier is equally effective across clinically relevant variations within the user population.

Thus, fairness was operationalized using *group-level separation*: examining performance by cluster. This allows to determine whether the model performs consistently across clinically meaningful subgroups. However, as the results show, there are marked differences: the model performs poorly on Cluster 2, the subgroup defined by subclinical symptom profiles. On the other hand, Cluster 1 (with the most considerable symptom burden and multi-condition morbidity) produces the highest detection rates, suggesting the model is excessively drawn towards prototypical cases.

4.5.5 Results - Global Model

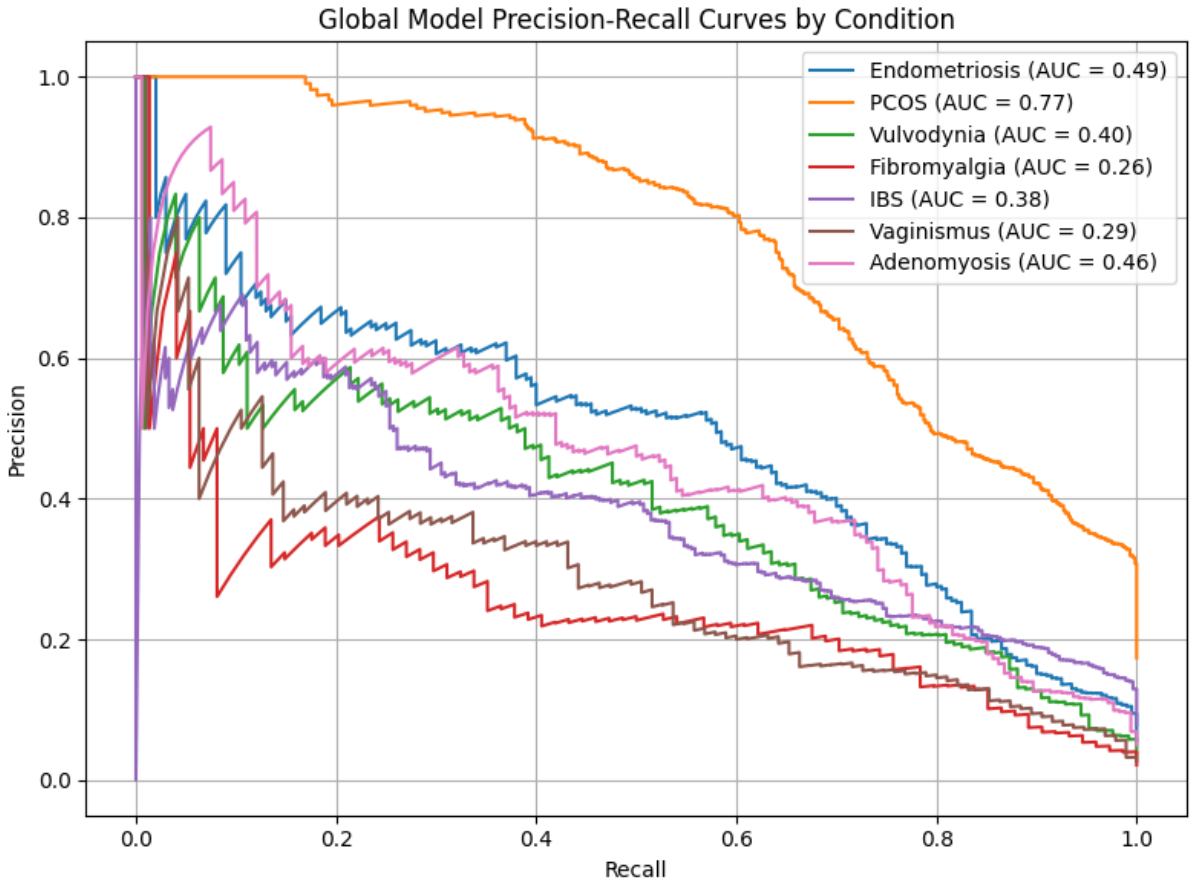

Figure 13: PR Curve Global Model

Figure 13 shows the PR curves for each condition of the global model.

As shown by the evaluation tables in Appendix C: Model Performance Tables, the model demonstrates good performance for the most common conditions, such as PCOS, which has a PR-AUC of 0.77, and Endometriosis, with a PR-AUC of 0.49. The F1 scores for the positive class are also modest, standing at 0.68 and 0.40, respectively.

However, for all other conditions, model performance drops significantly: Vulvodynia, Fibromyalgia, IBS, and Vaginismus all exhibit PR AUCs below 0.40, and their recall values for the positive class range from 0.04 (Vaginismus) to 0.17 (Vulvodynia). F1-scores remain critically low (0.26 for Vulvodynia, 0.12 for Fibromyalgia, 0.16 for IBS, and only 0.08 for Vaginismus).

Moreover, fairness across symptom clusters is demonstrably violated, as reflected by statistically significant disparities in model performance metrics, particularly recall and F1-score, across the clusters. *Equal opportunity,* which requires that the false negative rate be the same across all subgroups, is not upheld, as performance gaps between clusters consistently exceed 0.30 to 0.40 points across various diagnoses.

Across all conditions, for instance, recall for the positive class is drastically lower in Cluster 2: 0.04 for Endometriosis, 0.00 for Adenomyosis, Vulvodynia, and Vaginismus, and 0.02 for IBS. These rates contrast sharply with values in Cluster 1: 0.45 for Endometriosis, 0.43 for Adenomyosis, and 0.20 for Vulvodynia.

The disparities demonstrate structural bias in the model's interpretation and weightings of symptoms. For users in Cluster 2, there is no benefit from the model's predictions. This violates the basic principle of group fairness: all subpopulations should have an equal opportunity to be correctly identified if they represent positive cases.

Furthermore, the model did not achieve *predictive parity;* even when the model makes a positive prediction, its reliability is significantly different depending on the user's symptom cluster. For example, in the case of Vaginismus, the model achieves high precision in Cluster 3 (0.80), indicating that most of its positive predictions in this cluster are correct. However, its recall is only 0.11, and in all other clusters, it fails to identify any positive cases (recall = 0.00). The model's predictions rely on the prototypicality of symptom presentation, and it is unable to generalize to unseen, atypical symptomatology, despite being trained on the entire dataset. This highlights how ML models can worsen diagnostic blind spots, underscoring the importance of structure-aware modeling, as discussed in Section 4.6.

4.6 Classification Models using Clustering Insights

To address the limitations of a global model, the analysis continues with a *cluster-informed* training strategy. For each condition, rather than training on the entire population, a model was trained on the cluster where that condition is more prevalent, along with a small, stratified random sample (20%) of the rest of the population. The model is then evaluated on the remaining global dataset, excluding the cluster and the randomly sampled users.

This setting allows each model to learn specific cluster-specific patterns while still being minimally exposed to broader population variance.

4.6.1 Experimental Setup

Each model is trained independently for each condition following the above strategy. Below is the list of conditions trained per cluster:

- Cluster 0: PCOS, IBS
- Cluster 1: Endometriosis, Adenomyosis, Fibromyalgia
- Cluster 3: Vulvodynia, Vaginismus

No condition was trained on Cluster 2 due to the low prevalence of all diagnoses in this cluster.

4.6.2 Threshold Fine-Tuning

To improve recall and precision for the positive class, *threshold optimization* was performed after training. Instead of using the default 0.5 classification threshold, F1-scores for the positive class were tested across various thresholds (ranging from 0.1 to 0.9), and the optimal threshold was utilized to evaluate the model's performance.

4.6.3 Cluster-Aware Modelling

The cluster-aware models exhibit improvements over the global model, particularly in recall and F1 score for the positive class. The most interesting insight is that fairness increases across symptom clusters. However, this causes PR-AUC to decrease in several conditions. This implies that while a cluster-aware approach may reduce the overall discrimination power across the entire population, it improves the ability of the model to recognize underrepresented cases and subpopulations (as for Cluster 2).

The most striking result emerges in Vaginismus, the condition with the lowest prevalence (less than 3%), where the global model completely failed to identify any positive cases in Clusters 2 (recall = 0.00). The cluster-informed model increases recall in cluster 2 to 0.56, resulting in an overall F1 score of 0.27. While these results are not ideal, they make the classifier minimally functional for this very rare condition.

IBS metrics also improve: recall increases from 0.09 to 0.55, and F1 from 0.16 to 0.38, with all clusters showing positive recall shifts. For instance, in Cluster 0, recall improves from 0.10 to 0.53.

In the Endometriosis model, where global recall was 0.30, the cluster-informed model improves recall to 0.38 while preserving the same F1-score (0.41). Cluster 2 recall increases from 0.04 to 0.24, indicating that the model can now capture low-severity or subclinical presentations to a certain extent.

For Adenomyosis, the cluster-based model increases overall recall from 0.23 to 0.44, and Cluster 2 (where recall was previously 0.00) improves to 0.19.

These results demonstrate that cluster-informed models are fairer and can enhance predictive sensitivity, even in clusters where the global model performs poorly. This highlights how structure-aware modeling can address the structural biases, noise, and diagnostic blind spots that would otherwise flaw the performance of traditional classifiers.

4.6.4 Adding Distance from Centroids

Subsequently, the distance of users from all cluster centroids was incorporated into the model. Adding these variables yields improvements in classification performance, particularly in terms of recall and fairness across subgroups.

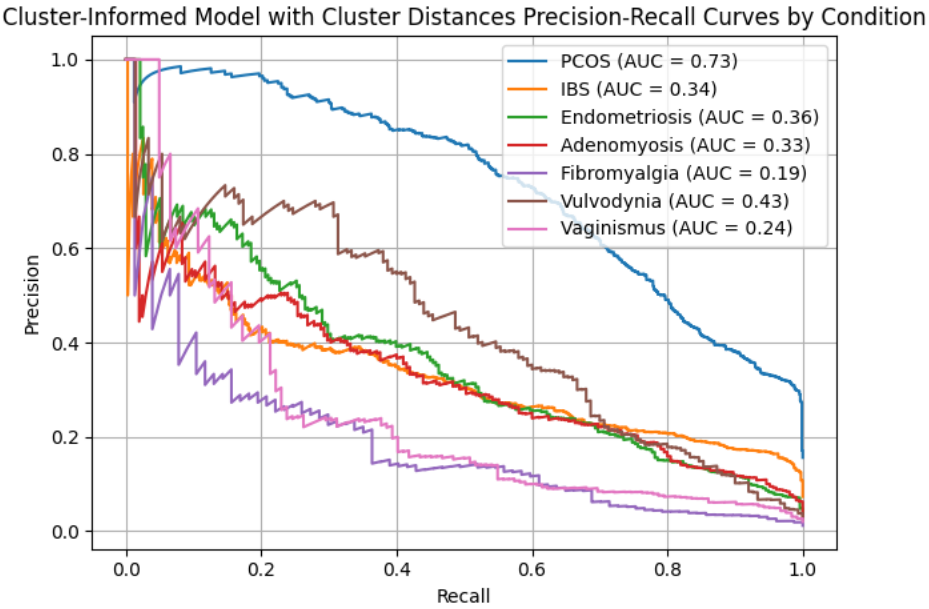

Figure 14: PR Curves Cluster-Informed Model

Figure 14 shows the final Precision-Recall (PR) curves for each condition after incorporating cluster distance features into the model. For instance, PCOS achieves an optimal PR-AUC (0.77), with consistently high precision and recall across the curve. Conditions such as Adenomyosis and Endometriosis also benefit, with PR-AUCs of 0.46 and 0.47, respectively. The performance for less common conditions remains modest, yet improvements have been observed, particularly in recall and equity among subgroups.

Additionally, the centroid distance is beneficial when calculating *Feature Importances*, which indicate how much each feature contributes to reducing prediction error across all trees in the model. In most cases, they appear within the top ten most essential predictors; the model leverages distances as an encoding of symptom pattern similarity and clinical prototypicality. While feature importances do not reveal the direction of a feature's effect (i.e., whether it increases or decreases the likelihood of a positive prediction), their consistent prominence in the rankings highlights their predictive utility.

4.6.5 SHapley Additive exPlanations values evaluation

Feature Importance scores from tree-based models (like CatBoost) show which variables are most influential on average, but they do not reveal the direction, context, or individual-level contribution of each feature. To overcome this limitation and achieve

a more interpretable and transparent model analysis, *SHAP (SHapley Additive exPlanations) values* are implemented.

SHAP values, based on cooperative game theory, highlight the extent to which each feature increases or decreases the probability of a diagnosis, both at the local level (for the individual user) and globally (for the whole dataset).

The SHAP summary plots for each target condition can be retrieved in Appendix D.

These plots show the most predictive features and their direction of impact, ranked by overall importance. Each point on a SHAP plot represents a user (i.e., an individual prediction), and the color indicates the feature value (e.g., high vs. low). The position on the x-axis shows how much that feature increases or decreases the model's predicted probability of a diagnosis.

In nearly all conditions, the distances from the cluster (e.g., dist_to_0, dist_to_3) appear among the most critical variables, substantiating their function as encodings of clinical pattern similarity.

Most importantly, SHAP values identify clinically relevant symptoms across all conditions. For example, in the Fibromyalgia model, musculoskeletal pain, fatigue, and cognitive symptoms are the most predictive features, consistent with the diagnosis. Similarly, in cases of Vaginismus and Vulvodynia, behaviors associated with surface pain and avoidance of penetration dominate the predictions. In the case of PCOS, the model focuses on classic endocrine and/or metabolic symptoms, such as facial hair, acne, and irregular periods. In IBS, the predictions mirror the gastrointestinal symptoms: bloating, constipation, and pain with bowel movements. In Endometriosis and Adenomyosis, the strong predictors include chronic pelvic pain, deep dyspareunia, and menstrual irregularities.

These plots demonstrate that the model is clinically valid, as the strong predictors are associated with known symptom patterns. This marks progress; one of the significant challenges in creating a healthcare machine learning model is obtaining *clinical explainability*. Additionally, this results in a higher level of reliable predictions by identifying exactly which symptoms influenced each decision.

4.6.6 Final Considerations

The cluster-based models have significantly contributed to enhancing recall and fairness within each identified cluster, including the previously overlooked subclinical subpopulations. This represents an achievement: the analysis was designed to identify latent subpopulations and provide more equitable evaluations, while improving the detection of true positives, rather than to create a clinically applicable model.

The performance of the classification models (specifically in terms of PR-AUC) remains limited and far from any threshold of clinical applicability. This suggests that while clustering helps mitigate unfairness, the underlying data remains affected by several structural biases outlined in Chapter 2: symptom overlap, retrospective self-reporting, treatment-related confounding, and the underrepresentation of specific subgroups. These factors introduce relevant noise, biases, and distortions in the label-symptom relationship.

To reach a balance between fairness and strong predictive performance, more expressive architectures will likely be required, as discussed in the next chapter.

5. Structure-Aware Neural Network

5.1 Introduction

This chapter introduces a new framework to enhance the predictive performance and fairness of classification models for identifying chronic pelvic pain conditions. A structure-aware neural network is implemented based on the findings from symptom-based clustering (Chapter 4).

The sections below describe the model design, training process, evaluation, and comparison with the previously tested cluster-based tree models.

5.2 Model Architecture

The model is a multi-label neural network. Since users may report more than one condition, the model performs multi-label classification by outputting a separate probability score for each diagnosis, rather than assigning a user to a single class.

The neural network is composed of a *shared feature extractor*, a *cluster embedding layer,* and a *diagnosis classifier*.

The shared feature extractor consists of two hidden fully connected layers with *ReLU activations* and *dropout regularization*. These layers are shared across all users, regardless of their cluster, and operate as a nonlinear transformation of the symptom input space. They enable *feature extraction* by mapping high-dimensional symptom vectors into a lower-dimensional latent space, which captures the most informative directions of variance for classification (i.e., the symptom patterns that best distinguish between different target conditions).

The extractor, by sharing parameters across the entire dataset, captures global statistical regularities in the data: frequent symptom combinations and co-occurring structures. Dropout also allows for generalization by randomly deactivating neurons with a fixed probability. This encourages the network to learn robust features that generalize beyond individual examples or clusters.

This shared representation serves as the basis for the model's understanding of symptoms, providing a consistent and meaningful input for the subsequent cluster-aware classification layers.

Subsequently, instead of treating the cluster a user belongs to as a categorical label (i.e., a one-hot vector, such as [0, 1, 0, 0] for Cluster 1) and treating clusters as unrelated categories, the model uses a learnable *embedding layer* to represent each cluster. Each cluster gets mapped into a continuous vector of fixed size (for instance, [0.15, -0.42, 0.61, 0.08]) which is initialized randomly and optimized during training. These embeddings are trained with the rest of the network, allowing them to evolve and encode meaningful information about the cluster (i.e., the typical symptom pattern

or diagnostic tendency). The model, therefore, learns to adjust its prediction based on the user's group context, specializing its diagnostic logic for the different subpopulations in the dataset.

For each target, a separate classifier takes as input both the symptom-based representation from the shared extractor and the cluster embedding. This allows each classifier to adapt its decisions based on individual symptoms and group-level patterns: a model that retains global information but is also locally specialized. It learns how cluster context modifies the meaning of the symptoms, improving its ability to detect atypical cases that were underrepresented in previous models.

This architecture was chosen to combine feature learning (via dense neural layers) with flexibility to account for group-level variability (via trainable cluster embeddings), enabling the model to generalize across diverse symptom presentations while maintaining individualized predictions.

5.3 Training and Optimization Strategy

The dataset was split into training and validation sets using *stratified sampling* based on cluster membership, ensuring that each cluster is proportionally represented in both sets. Distance-to-centroid features were retained, as they provide a clinically meaningful signal regarding symptom specificity and atypicality and complement the cluster embeddings by adding intra-cluster variance.

Given the severe class imbalance, a *weighted sampling strategy* was employed during training: users with no diagnoses were assigned lower sampling weights, reducing their frequency in training batches, while users with at least one diagnosis had higher sampling weights. This increases the model's exposure to positive examples, specifically for rare conditions (such as Vaginismus or Fibromyalgia), which would otherwise be underrepresented.

The model was trained using the *Adam optimizer*, with a learning rate of 0.001. The training process was monitored through validation performance, with a maximum of

50 epochs. The best-performing model (based on F1-score) was selected for evaluation.

Due to the significant class imbalance, the *Focal Loss* function was selected over the *binary cross-entropy* loss function. This Loss function scales the loss assigned to well-classified users to focus more on hard-to-classify positive cases. This is particularly important in a clinical context, where rare cases carry significant importance.

5.4 Results

This section presents the results of the structure-aware neural network across the seven diagnostic targets. The model is evaluated using the same metrics presented in Chapter 4: precision, recall, F1-score, and area under the precision-recall curve (PR AUC). Results are reported both globally and stratified by cluster membership. The results tables are in Appendix C: Model Performance Tables.

5.4.1 Overall Performance of the Structure-Aware Neural Network

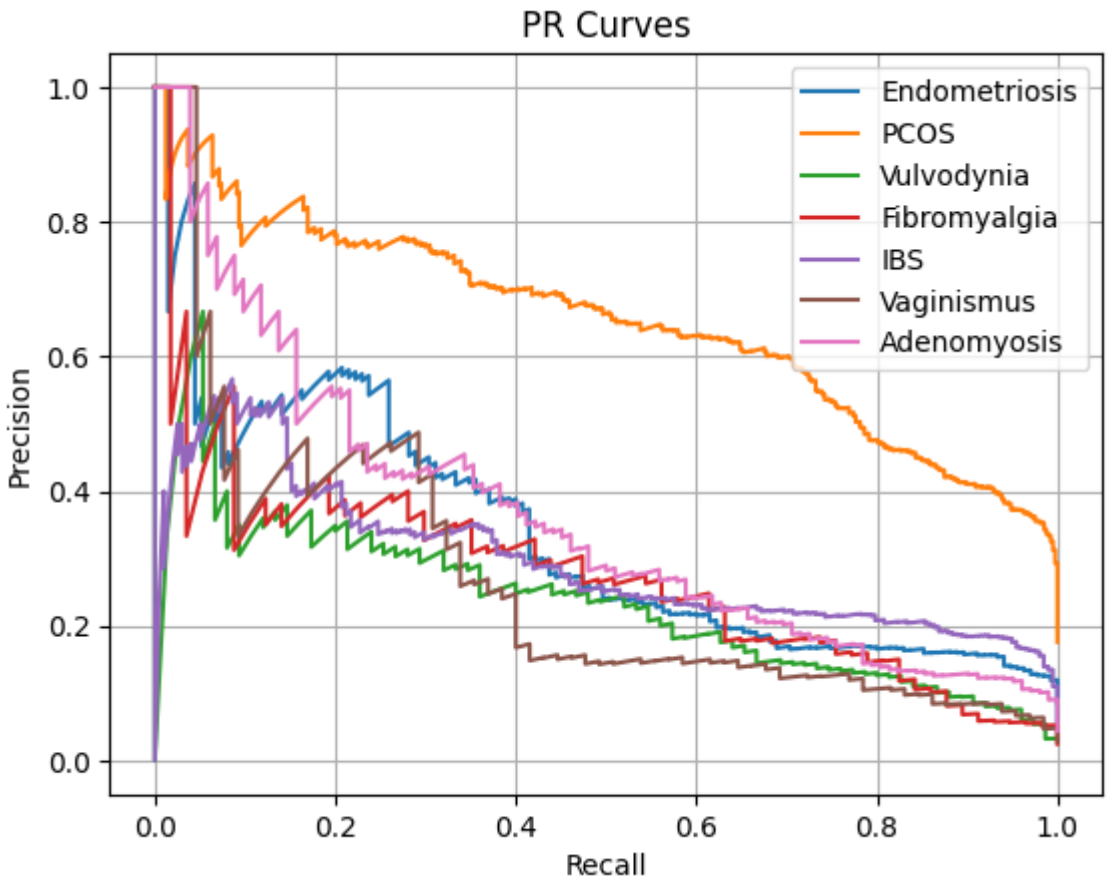

Figure 15: PR Curves Structure-Aware NN

The model demonstrates strong classification ability for prevalent conditions such as PCOS (F1 = 0.63, PR AUC = 0.65) and Endometriosis (F1 = 0.37, PR AUC = 0.33), while maintaining reasonable sensitivity for rarer targets like Vaginismus (F1 = 0.33, PR AUC = 0.27) and Fibromyalgia (F1 = 0.36, PR AUC = 0.28).

Cluster-stratified evaluation, a crucial element for fairness assessment, demonstrates that the model can adapt to the heterogeneous structures of symptoms. As expected, Cluster 1 has the overall highest recall across most conditions. Interestingly, the model performs best on Cluster 3 for conditions that correspond to its symptom structure (e.g., sensory and neuromuscular pain). Specifically, recall is 0.44 for Vaginismus and 0.54 for Vulvodynia within this subgroup.

However, Cluster 2 remains problematic. Here, recall for conditions like Fibromyalgia (0.00) and Adenomyosis (0.00) is null. This suggests that the model continues to struggle with subclinical, vague diagnostic profiles that do not conform to typical diagnostic features. Further adjustments may be required to reduce fairness gaps in this subgroup.

5.4.2 Comparison with the Cluster-Informed Gradient Boosting Model

The structure-aware neural network is compared to the top-performing model from Chapter 4: a CatBoost classifier trained on a single cluster and assessed using the remaining data, which also incorporates the distance from the cluster features.

Both models perform comparably on average, but diverge in strengths, as shown by Table 2.

Condition	**F1 (CatBoost)**	**F1 (Neural Net)**	**PR AUC (CatBoost)**	**PR AUC (Neural Net)**
PCOS	0.67	0.63	0.73	0.65
Endometriosis	0.41	0.37	0.36	0.33
Vaginismus	0.29	0.33	0.24	0.27
Fibromyalgia	0.28	0.36	0.19	0.28
Vulvodynia	0.48	0.32	0.43	0.24

Table 2: NN and CatBoost Performance Comparison

The CatBoost model generally achieves higher PR AUC values, suggesting more stable decision boundaries, particularly for prevalent and well-defined conditions like PCOS and Vulvodynia. By contrast, the NN shows improved F1 scores for conditions with more ambiguous symptom patterns, such as Fibromyalgia and Vaginismus.

At the cluster level, the neural network shows clear strengths in Cluster 3, especially for sensory and neuromuscular pain conditions: Vaginismus (Cluster 3): recall = 0.44 (NN) vs. 0.36 (CatBoost); Vulvodynia (Cluster 3): recall = 0.54 (NN) vs. 0.28 (CatBoost).

In contrast, the CatBoost model performs better in Cluster 2, which includes diagnostically vague or low-intensity cases: for instance, Vaginismus (Cluster 2): recall = 0.53 (CatBoost) vs 0.24 (NN); Adenomyosis (Cluster 2) recall = 0.20 (CatBoost) vs. 0.00 (NN).

The CatBoost model effectively identifies subtle, atypical cases. The NN captures nuanced, symptomatic patterns but struggles with low-signal users. Both models reduce disparities compared to a global baseline in complementary ways. The neural network improves sensitivity and equity in high-symptom, underserved groups (Cluster 3), while CatBoost consistently captures weakly expressed cases (Cluster 2).

5.5 Final PR Curves Comparison and Considerations

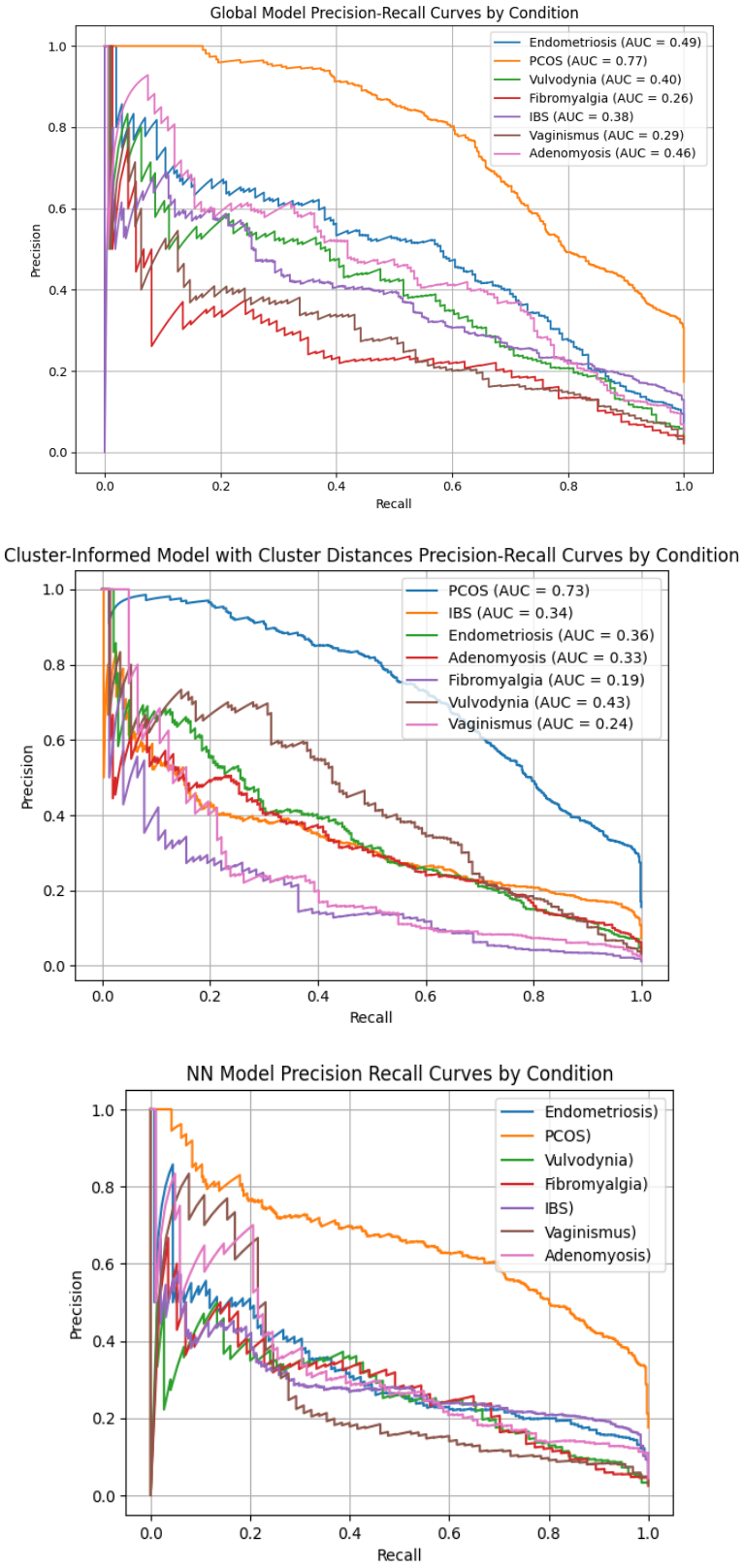

Figure 16: Comparative PR Curve Analysis Across Models

The combined visualization above enables direct comparison of the three modeling strategies across all target conditions. Several important patterns emerge:

- PCOS consistently shows the highest PR curve and widest spread across recall, with minimal variation across the three models.
- Fibromyalgia and IBS benefit from cluster-informed and neural strategies. Both conditions show flattened curves in the global model, but more balanced PR curves with improved early recall in the structure-aware approaches.
- Vulvodynia is an interesting case: the cluster-informed model achieves the highest AUC (0.43), outperforming both global and neural models.

The results presented in this chapter demonstrate that integrating structural information into predictive models, through both clustering and architectural design, can improve the detection of complex gynecological conditions. While the structure-aware neural network does not outperform the cluster-informed tree model across all metrics, it still performs better for some specific conditions, particularly in symptom clusters marked by sensory and neuromuscular pain. In any case, however, there are still limitations, especially in subgroups where low-intensity or diffuse symptoms are present, which limit model sensitivity. Future work should seek to leverage hybrid structures and additional fairness interventions to improve disparities between user groups.

Moreover, for healthcare applications, *explainability* is a critical concern. While tree-based models provide explainability solutions through frameworks like SHAP (SHapley Additive exPlanations) values, neural networks inherently operate as black-box models. They lack the transparency needed in clinical decision-making, where practitioners must understand and trust the reasoning behind predictions. This poses a significant barrier to the utilization of neural networks in healthcare.

6. Conclusions

This thesis explored the opportunities and limitations of applying machine learning (ML) algorithms to detect chronic pelvic pain conditions, with a specific focus on enhancing fairness and sensitivity in underrepresented subpopulations. Working with a large, real-world dataset of self-reported symptoms from the Hale digital health

platform, the project aimed to address the structural biases and clinical uncertainties that make predictive modeling challenging in this field.

6.1 From Objectives to Strategy

As highlighted in Chapter 1, chronic pelvic pain (CPP) syndromes such as Endometriosis, Vulvodynia, and Fibromyalgia are often underdiagnosed due to overlapping symptoms, unclear symptomologies, and gender-based medical biases.

The goal of the project was to identify latent subpopulations characterized by different symptomatologies, pain expression patterns, and diagnostic likelihoods. Given those subgroups, the analysis aimed to facilitate the early identification of these conditions while maintaining clinical validity and ensuring fairness across the entire dataset.

6.2 Understanding the Data

Chapter 2 examines the fairness issues associated with the application of machine learning in healthcare. It analyzes the impacts of structural, algorithmic, and label bias on the performance of the models and their ability to reinforce clinical divides. Fairness was examined through the notions of group-level separation and individual fairness, which are essential for creating fair clinical machine learning models.

6.3 Data-Driven Structure Discovery

In Chapter 3, symptom-based clustering was applied to the Hale dataset, revealing four latent user groups with distinct clinical characteristics. These clusters ranged from hormonally/metabolically affected users to sensory pain-dominant subgroups, and eventually, diagnostically vague profiles. The unsupervised learning algorithm helped to uncover hidden patterns in symptom expression. This indicated that many users with low diagnostic prevalence still report symptom burden, and thus, potentially identified a clinical blind spot.

6.4 Modeling with Structural Awareness

Chapters 4 and 5 demonstrated how integrating clustering insights into ML model design yields measurable effects. This integration is studied both as a training stratification method, using cluster-informed classification models, and as an architectural component, building structure-aware neural networks.

Cluster-informed tree models improved recall and fairness for rare conditions (e.g., Vaginismus, Adenomyosis) that the global model failed to detect. Moreover, adding centroid distances as features enhanced classification sensitivity across subclinical presentations, with SHAP analyses confirming that these distances encode clinically meaningful variation. The structure-aware neural network demonstrated better performance for specific symptom clusters (e.g., Cluster 3: sensory pain), although it yielded less balanced results across rare and atypical profiles.

6.5 Fairness vs. Performance Trade-Offs

An important insight is the inherent trade-off between overall predictive accuracy and fairness. While the global model achieved higher PR-AUC for the most prevalent conditions, it did so at the cost of missing rare but significant cases, especially within underdiagnosed user groups. Cluster-aware approaches, despite slightly lower PR-AUC, produced higher recall in underrepresented subpopulations, offering a fairer foundation for a clinical screening tool.

6.6 Final Reflections

This thesis demonstrates how machine learning (ML) classification models can benefit from identifying meaningful substructures within noisy, self-reported symptom data. Incorporating such latent symptom clusters into both feature engineering and model design enables the mitigation of some of the clinical blind spots that characterize gynecological care today.

Most importantly, this work demonstrates that fairness in clinical machine learning requires careful architecture and evaluation designs. Cluster-aware training and

embedding distance features show that fairness can be obtained even in highly imbalanced datasets.

However, this research also highlights some critical limitations. The clinical applicability of these models remains constrained by the retrospective and noisy nature of the data, as well as the ambiguity of chronic pain conditions. Moreover, although structure-aware neural networks show promise in modeling these complex interactions appropriately, the lack of explainability is a significant barrier to their clinical application. In contrast, tree-based models offer the possibility of interpretability, which is crucial in a field in which clinicians must be able to understand and justify each prediction.

6.7 Possible Future Work Development

Future work should strengthen both the clinical relevance, the statistical performance, and the fairness of the models developed in this thesis.

First, data quality should be improved. Adding questions on treatment (whether medication was started and when, and whether symptoms improved) would allow us to disentangle the confounding effect of potential drugs on the condition. Moreover, validating the model's predictions through follow-up visits with clinicians would assess whether the model's outputs align with actual diagnostic outcomes.

Further research should also explore fairness-aware training objectives to explicitly reduce subgroup disparities and develop hybrid architectures that retain the flexibility of neural networks while offering the interpretability required in healthcare.

Ultimately, the goal is to transition from experimental models to clinically trustworthy systems that facilitate the fair and early detection of complex gynecological conditions.

Appendix A: Screening Questionnaire

Literature used for the creation of the screening test

- Borstein, J., et al. (2016). 2015 ISSVD, ISSWSH, and IPPS Consensus Terminology and Classification of Persistent Vulvar Pain and Vulvodynia. Obstet Gynecol.
- Leslie, S. W., et al. (2022). Pudendal Neuralgia. StatPearls.
- Becker, C. M., et al. (2022). ESHRE guideline: endometriosis. Hum Reprod Open.
- Conn, A., & Hodges, K. R. (2023). Genito-Pelvic Pain/Penetration Disorder. MSD Manual Professional Edition.
- Moawad, G., et al. (2023). Adenomyosis: An Updated Review on Diagnosis and Classification. Journal of Clinical Medicine.

Questionnaire

Below is the full version of the screening questionnaire used by Hale to collect symptom and diagnostic data from users. The questionnaire was developed for an Italian-speaking female population and is administered exclusively in Italian. However, to ensure clarity and accessibility, an English translation was created.

Let’s begin: what is your current menstrual cycle like?

- I do not have menstruation due to hormonal contraception
- My periods are rare, but I have them occasionally (less than 8 times a year)
- Almost always regular, but sometimes start earlier or later than expected
- Very regular and predictable
- Irregular, it’s hard to predict when my period will start

- I do not have periods
- I am pregnant, so I am not menstruating

Do you experience unusual bleeding? *Select symptoms that have lasted for more than three months.*

- None of the above
- Bleeding or spotting between periods
- Very heavy menstruation
- Menstruation that lasts more than 7 days
- Bleeding after sexual intercourse
- Rectal bleeding during menstruation

Do you experience pain in these areas or situations? *Select symptoms that have lasted for more than three months, and not only during your period.*

- Back or leg pain
- Pain at the vaginal entrance during sex
- Muscle cramps and various body aches
- Persistent pelvic pain or cramps (lower abdomen)
- Pain only in the vulvar area
- None of the above

- Pain radiating to thighs, buttocks, perineum, or vulva
- Deep pain during vaginal penetration
- Diffuse pain throughout the body

Do you experience pain during these situations?

- Pain related to sexual activity
- Severe pain during or before menstruation
- Pain during ovulation
- None of the above
- Pelvic pain at any time, unrelated to the menstrual cycle
- Pain that worsens at the start of the menstrual cycle
- Pain after menstruation

Do you experience any of these symptoms in the intimate or intestinal area? *Select symptoms that have lasted for more than three months.*

- Discomfort in the vulvar area (itching, pain, dryness, or redness)
- Vaginal discharge or unusual odor
- Pain during bowel movements
- Burning in the bladder or urethra when urinating

- Frequent need to urinate or difficulty emptying the bladder
- Abdominal pain or cramps that improve after a bowel movement
- Abdominal pain or cramps that worsen after eating
- Painful defecation or difficulty emptying the rectum
- Diarrhea
- Sensation of pressure, lumps, or bulges in the abdominal/pelvic area
- Frequent urinary tract infections or vaginitis in the last year
- None of the above
- Nausea or discomfort, especially during menstruation
- Abdominal bloating
- Flatulence
- Feeling of not emptying the bladder or urinary leakage
- Constipation
- Burning sensation while urinating due to vulvar irritation

Do you experience any of these symptoms? *Select symptoms that have lasted for more than three months.*

- Stinging sensation at the vaginal entrance

- Itching and irritation in the vulvar area
- Pain when sitting that improves when standing or lying down
- None of the above
- Constant vulvar pain or burning not related to a specific event
- Tender points throughout the body
- Strange or unpleasant sensation when touching the intimate area (e.g., burning, prickling)
- Pain or discomfort from vulvar pressure (e.g., tight clothing, sitting on a bike, touching)
- Pain only during penetration attempts (e.g., intercourse, gynecological exams)
- Swelling of the labia (inner or outer)
- Unable to insert tampons or undergo gynecological exams
- Widespread musculoskeletal pain

Do you experience any of these symptoms related to sexual activity?

- Pain at the vaginal entrance, like a cut or burning
- I tense up and contract my abdominal muscles during penetration
- None of the above
- Pain during intercourse

- Pain during deep vaginal penetration
- Fear before intercourse, when thinking about it
- Pain during penetration, as if there were a wall
- Pain when the clitoris or vulva is touched externally

Do you significantly experience any of these symptoms?

- Thinning hair or hair loss
- Muscle pain and weakness
- Fatigue or tiredness
- Excess facial or body hair (arms, chest, abdomen)
- Acne or oily skin
- Difficulty concentrating
- Sleep disturbances
- None of the above
- Weight gain

Do you use any form of contraception?

- Combined birth control pill (most common pill)
- Barrier method (e.g., male condom)

- I do not use contraception
- Progestin-only birth control pill (mini pill)
- Copper IUD
- Hormonal IUD
- Subdermal implant
- Vaginal ring
- Natural methods (e.g., basal temperature)
- Hormonal patch
- Intrauterine ball (IUB)
- Injection

Have you ever had any of the following conditions?

- Vaginitis (itching, burning, and discharge in the vaginal area)
- Recurrent yeast infections (>3 per year)
- Recurrent cystitis (>3 per year)
- Urinary tract infections (UTIs)
- None of the above
- Human papillomavirus (HPV)

- Gonorrhea
- Bacterial vaginosis (discharge with bad odor but little itching)
- Genital herpes
- Chlamydia
- Trichomoniasis
- HIV
- Ureaplasma
- Mycoplasma genitalium

Have you ever received an official diagnosis for any of the following (only gynecological conditions)?

- Adhesions
- Ovarian cysts
- Polycystic ovaries
- PCOS – Polycystic Ovary Syndrome
- None of the above
- Hypothyroidism
- Adenomyosis

- Ovulation pain (Mittelschmerz)
- Irritable Bowel Syndrome (IBS)
- Uterine fibroids
- Vulvodynia
- Uterine prolapse
- Primary dysmenorrhea
- Pelvic floor muscle hypertonia
- Cervical cancer
- Hyperthyroidism
- Primary amenorrhea
- Pelvic varicocele
- Vestibulodynia
- Pudendal neuropathy
- Endometrial anomaly
- Perimenopause
- Menopause
- Diverticulitis

- Vaginismus
- Fibromyalgia
- Pelvic inflammatory disease
- Interstitial cystitis (painful bladder syndrome)
- Endometritis
- Premature ovarian insufficiency
- Endometrial cancer
- Ovarian cancer
- Ovarian torsion
- Asherman’s syndrome

Appendix B: Clustering Symptom List

The list below summarizes the symptom items used to calculate group scores for each diagnosis category included in the clustering analysis. These symptom profiles were defined using clinical literature and expert input and reflect the core symptoms that characterize each diagnosis.

Endometriosis: Persistent pelvic pain or cramps, Severe pain during or before menstruation, Pain during ovulation, Pain after menstruation, Pain during penetrative intercourse, Pain worsening at the start of the cycle, Abdominal bloating, Fatigue or tiredness, Pain related to sexual activity, Back or leg pain, Pelvic pain not related to the cycle, Radiating pain to thighs, buttocks, perineum, vulva, Pain during bowel movements, Painful defecation or difficulty emptying rectum

PCOS: Acne or oily skin, Excess hair on face/arms/chest/abdomen, Thinning hair or hair loss, Weight gain, Infrequent menstruation, Irregular menstruation, Abdominal bloating, Fatigue or tiredness, Absence of menstruation

Vulvodynia: Pain at the vaginal entrance, Burning or pins-and-needles sensation, Pain from external pressure, Constant vulvar pain, Pain when touching the clitoris, Pain related to sexual activity, Stinging sensation at the vaginal entrance, Pain only in the vulvar area, Pain when sitting

Fibromyalgia: Widespread musculoskeletal pain, Muscle pain and weakness, Tender points on the body, Fatigue or tiredness, Sleep disturbances, Difficulty concentrating, Muscle cramps and various pains, Generalized body pain

IBS: Abdominal bloating, Constipation, Diarrhea, Cramps relieved after defecation, Flatulence, Nausea or discomfort during menstruation, Pain during bowel movements, Fatigue or tiredness, Cramps worsen after eating, Painful defecation, Frequent urination or bladder difficulties

Vaginismus: Pain at the vaginal entrance, Tightening during penetration, Inability to use tampons or undergo exams, Fear before intercourse, Sensation of "wall" during penetration, Pain related to sexual activity

Adenomyosis: Very heavy menstruation, Menstruation longer than 7 days, Severe menstrual pain, Persistent pelvic cramps, Pain during penetrative intercourse, Pain related to sexual activity, Chronic pelvic pain, Pain during bowel movements, Back or leg pain, Painful defecation

Uterine Fibroids: Very heavy menstruation, Sensation of pressure in the abdomen/pelvis, Menstruation longer than 7 days, Abdominal bloating, Back or leg pain, Chronic pelvic pain, Vaginal discharge or unusual odor

Results of Cluster Analysis

Cluster 0

Extreme users' symptoms along PC1 Negative:

- Painful defecation or difficulty emptying the rectum
- Pain during penetrative intercourse, deep inside the vagina
- Frequent urinary tract infections or vaginitis in the last year
- Urinary tract infections (UTIs)
- Feeling tired or fatigued

Extreme users' symptoms along PC1 Positive:

- Acne or oily skin
- Pain that worsens at the start of the menstrual cycle
- Severe pain during or before menstruation
- Persistent pain or cramps in the pelvis (lower abdomen)
- Very heavy menstruation
- Nausea and feeling unwell, especially during menstruation
- None of the above
- None of the above

- None of the above
- None of the above
- Feeling tired or fatigued

Extreme users' symptoms along PC2 Positive:

- Weight gain
- Muscle cramps and various body aches
- Difficulty concentrating
- Back or leg pain
- Diffuse pain throughout the body
- Widespread musculoskeletal pain
- Muscle pain and weakness
- Abdominal bloating
- None of the previous situations
- None of the above
- Feeling tired or fatigued

Extreme users' symptoms along PC2 Negative:

- Pain related to sexual activity
- Pain during penetrative intercourse, deep inside the vagina

- Deep pain during vaginal penetration
- Abdominal bloating

Cluster 1

Extreme users' symptoms along PC1 Negative:

- Severe pain during or before menstruation
- Persistent pain or cramps in the pelvis (lower abdomen)
- Abdominal bloating
- Very heavy menstruation
- None of the above
- Feeling tired or fatigued

Extreme users' symptoms along PC1 Positive:

- Discomfort in the vulvar area (itching, pain, dryness, or redness)
- Pain at the entrance of the vagina, like a cut or burning
- Pain that radiates to thighs, buttocks, perineum, and vulva
- Widespread musculoskeletal pain
- Pelvic pain at any time, unrelated to the menstrual cycle
- Pain when sitting that improves when standing or lying down
- Muscle pain and weakness

- Abdominal bloating
- Itching and irritation in the vulvar area
- Feeling tired or fatigued
- Strange or unpleasant sensation when touching the intimate area, like burning or pins-and-needles

Extreme users' symptoms along PC2 Positive:

- Muscle cramps and various body aches
- Difficulty concentrating
- Back or leg pain
- Pain that radiates to thighs, buttocks, perineum, and vulva
- Diffuse pain throughout the body
- Pain after menstruation
- Pain during ovulation
- Persistent pain or cramps in the pelvis (lower abdomen)
- Muscle pain and weakness
- Abdominal bloating
- Nausea and feeling unwell, especially during menstruation
- Sensation of pressure, lumps, or bulges in the abdominal/pelvic area

- Feeling tired or fatigued

Extreme users' symptoms along PC2 Negative:

- Discomfort in the vulvar area (itching, pain, dryness, or redness)
- Pain at the entrance of the vagina during sexual intercourse
- Pain related to sexual activity
- Pelvic pain at any time, unrelated to the menstrual cycle
- Feeling of not having emptied the bladder or leaking drops of urine

Cluster 2

Extreme users' symptoms along PC1 Negative:

- Cystitis (bladder infection)

Extreme users' symptoms along PC1 Positive:

- None of the above
- None of the previous answers
- None of the previous answers
- None of the above
- None of the above

Extreme users' symptoms along PC2 Positive:

- Pain at the entrance of the vagina during sexual intercourse
- Pain related to sexual activity

- Menopause
- None of the above
- None of the above
- None of the previous situations
- None of the above
- I do not have menstruation

Extreme users' symptoms along PC2 Negative:

- Discomfort in the vulvar area (itching, pain, dryness, or redness)
- Vaginitis

Cluster 3

Extreme users' symptoms along PC1 Negative:

- Pain during penetrative sex, as if there were a wall
- None of the above
- None of the above
- I am completely unable to insert tampons or undergo gynecological exams
- Fear before sexual intercourse, when I think about it
- I feel pain in the vulvar area only during penetration attempts (e.g., intercourse or gynecological exam)
- If penetration is attempted, I tense up and contract my abdominal muscles

Extreme users' symptoms along PC1 Positive:

- Pain or discomfort when there is pressure on the vulva (such as tight clothing, sitting on a bicycle, or touching with fingers)
- Pain when the clitoris or vulva are touched externally
- Pain only in the vulvar area
- Strange or unpleasant sensation when touching the intimate area, like burning or pins-and-needles
- Vaginitis

Extreme users' symptoms along PC2 Positive:

- Pain at the entrance of the vagina during sexual intercourse
- Pain at the entrance of the vagina, like a cut or burning
- Pain related to sexual activity
- Menopause
- I do not have menstruation

Extreme users' symptoms along PC2 Negative:

- Pain at the entrance of the vagina during sexual intercourse
- Pain at the entrance of the vagina, like a cut or burning
- Pain during sexual intercourse
- Severe pain during or before menstruation

- Deep pain during vaginal penetration
- I am completely unable to insert tampons or undergo gynecological exams
- Fear before sexual intercourse, when I think about it

Appendix C: Model Performance Tables

Endometriosis Global Model				
	precision	**recall**	**f1-score**	**support**
0	0,96	0,99	0,97	3270
1	0,62	0,30	0,40	200
accuracy	0,95	0,95	0,95	
macro avg	0,79	0,64	0,69	
weighted avg	0,94	0,95	0,94	
PR AUC: 0.49				
cluster	**precision**	**recall**	**f1_score**	**support**
0	0,65	0,22	0,32	1027
1	0,61	0,45	0,52	512
2	0,33	0,04	0,07	1245
3	0,73	0,23	0,35	686

Adenomyosis Global Model				
	precision	**recall**	**f1-score**	**support**
0	0,96	0,99	0,98	3296
1	0,61	0,23	0,33	174
accuracy	0,95	0,95	0,95	
macro avg	0,78	0,61	0,65	
weighted avg	0,94	0,95	0,94	
PR AUC: 0.46				
cluster	**precision**	**recall**	**f1_score**	**support**
0	0,40	0,09	0,15	1008
1	0,64	0,43	0,52	514
2	0,00	0,00	0,00	1232
3	0,80	0,12	0,21	716

Table C 1: Endometriosis and Adenomyosis – Global Model

IBS Global Model				
	precision	**recall**	**f1-score**	**support**
0	0,93	1,00	0,96	3198
1	0,65	0,09	0,16	272
accuracy	0,92	0,92	0,92	
macro avg	0,79	0,54	0,56	
weighted avg	0,91	0,92	0,90	
PR AUC: 0.38				
cluster	**precision**	**recall**	**f1_score**	**support**
0	0,62	0,10	0,16	1007
1	0,68	0,15	0,25	515
2	0,50	0,02	0,04	1245
3	0,67	0,04	0,07	703

Fibromyalgia Global Model				
	precision	**recall**	**f1-score**	**support**
0	0,98	1,00	0,99	3396
1	0,45	0,07	0,12	74
accuracy	0,98	0,98	0,98	
macro avg	0,72	0,53	0,55	
weighted avg	0,97	0,98	0,97	
PR AUC: 0.26				
cluster	**precision**	**recall**	**f1_score**	**support**
0	0,00	0,00	0,00	1000
1	0,38	0,08	0,14	527
2	1,00	0,11	0,20	1277
3	1,00	0,11	0,20	666

Table C 2: IBS and Fibromyalgia – Global Model

PCOS Global Model				
	precision	**recall**	**f1-score**	**support**
0	0,92	0,97	0,94	2868
1	0,81	0,59	0,68	602
accuracy	0,91	0,91	0,91	
macro avg	0,86	0,78	0,81	
weighted avg	0,90	0,91	0,90	
PR AUC: 0.77				
cluster	**precision**	**recall**	**f1_score**	**support**
0	0,78	0,60	0,68	1058
1	0,75	0,50	0,60	481
2	0,85	0,67	0,75	1246
3	0,86	0,53	0,66	685

Vaginismus Global Model				
	precision	**recall**	**f1-score**	**support**
0	0,97	1,00	0,99	3375
1	0,80	0,04	0,08	95
accuracy	0,97	0,97	0,97	
macro avg	0,89	0,52	0,53	
weighted avg	0,97	0,97	0,96	
PR AUC: 0.29				
cluster	**precision**	**recall**	**f1_score**	**support**
0	0,00	0,00	0,00	1022
1	0,00	0,00	0,00	485
2	0,00	0,00	0,00	1266
3	0,80	0,11	0,19	697

Table C 3: PCOS and Vaginismus – Global Model

Vulvodynia Global Model				
	precision	**recall**	**f1-score**	**support**
0	0,97	0,99	0,98	3344
1	0,54	0,17	0,26	126
accuracy	0,96	0,96	0,96	
macro avg	0,75	0,58	0,62	
weighted avg	0,95	0,96	0,96	
PR AUC: 0.40				
cluster	**precision**	**recall**	**f1_score**	**support**
0	0,00	0,00	0,00	981
1	0,65	0,20	0,31	504
2	0,00	0,00	0,00	1275
3	0,46	0,20	0,28	710

Table C 4: Vulvodynia – Global Model

Endometriosis Cluster-Informed Model; Best Threshold: 0.56				
	precision	**recall**	**f1-score**	**support**
0	0.98	0.98	0.98	6677
1	0.44	0.38	0.41	256
accuracy	0.96	0.96	0.96	
macro avg	0.71	0.68	0.69	
weighted avg	0.96	0.96	0.96	
PR AUC: 0.35				
cluster	**precision**	**recall**	**f1_score**	**support**
0	0.44	0.40	0.42	2406
2	0.44	0.24	0.31	2956
3	0.42	0.50	0.46	1571

Adenomyosis Cluster-Informed Model; Best Threshold: 0.39				
	precision	**recall**	**f1-score**	**support**
0	0.98	0.97	0.98	6712
1	0.33	0.44	0.38	221
accuracy	0.95	0.95	0.95	
macro avg	0.65	0.71	0.68	
weighted avg	0.96	0.95	0.96	
PR AUC: 0.34				
cluster	**precision**	**recall**	**f1_score**	**support**
0	0.35	0.51	0.42	2397
2	0.17	0.19	0.18	2960
3	0.37	0.51	0.43	1576

Table C 5: Endometriosis and Adenomyosis – Cluster-Informed Model

IBS Cluster-Informed Model; Best Threshold: 0.41				
	precision	recall	f1-score	support
0	0.96	0.90	0.93	5286
1	0.29	0.55	0.38	417
accuracy	0.87	0.87	0.87	
macro avg	0.63	0.72	0.66	
weighted avg	0.91	0.87	0.89	
PR AUC: 0.34				
cluster	**precision**	**recall**	**f1_score**	**support**
1	0.33	0.61	0.43	1158
2	0.26	0.54	0.35	2929
3	0.27	0.47	0.34	1616

Fibromyalgia Cluster-Informed Model; Best Threshold: 0.21				
	precision	recall	f1-score	support
0	0.99	0.98	0.99	6850
1	0.22	0.40	0.29	83
accuracy	0.98	0.98	0.98	
macro avg	0.61	0.69	0.64	
weighted avg	0.98	0.98	0.98	
PR AUC: 0.18				
cluster	**precision**	**recall**	**f1_score**	**support**
0	0.23	0.52	0.32	2358
2	0.13	0.11	0.12	2987
3	0.24	0.38	0.30	1588

Table C 6: IBS and Fibromyalgia – Cluster-Informed Model

PCOS Cluster-Informed Model; Best Threshold: 0.64				
	precision	recall	f1-score	support
0	0,93	0,94	0,94	4813
1	0.68	0.64	0,66	891
accuracy	0,91	0,91	0,91	
macro avg	0,86	0,78	0,81	
weighted avg	0,90	0,91	0,90	
PR AUC: 0.72				
cluster	**precision**	**recall**	**f1_score**	**support**
1	0.61	0.52	0.56	1130
2	0.69	0.73	0.71	2964
3	0.71	0.59	0.65	1610

Vaginismus Cluster-Informed Model; Best Threshold: 0.29				
	precision	recall	f1-score	support
0	0.99	0.98	0.98	6360
1	0.23	0.35	0.27	132
accuracy	0.96	0.96	0.96	
macro avg	0.61	0.66	0.63	
weighted avg	0.97	0.96	0.97	
PR AUC: 0.22				
cluster	**precision**	**recall**	**f1_score**	**support**
0	0.12	0.15	0.13	2418
1	0.27	0.24	0.26	1155
2	0.25	0.56	0.34	2919

Table C 7: PCOS and Vaginismus – Cluster-Informed Model

Vulvodynia Cluster-Informed Model; Best Threshold: 0.5				
	precision	**recall**	**f1-score**	**support**
0	0.99	0.99	0.99	6331
1	0.49	0.41	0.45	161
accuracy	0.97	0.97	0.97	
macro avg	0.74	0.70	0.72	
weighted avg	0.97	0.97	0.97	
PR AUC: 0.40				
cluster	**precision**	**recall**	**f1_score**	**support**
0	0.40	0.11	0.17	2415
1	0.57	0.54	0.56	1186
2	0.10	0.07	0.08	2891

Table C 8: Vulvodynia – Cluster-Informed Model

Endometriosis Cluster-Informed Model; Optimized Threshold: 0.45				
	precision	**recall**	**f1-score**	**support**
0	0.98	0.97	0.98	6200
1	0.39	0.43	0.41	238
accuracy			0.95	
macro avg	0.68	0.70	0.69	
weighted avg	0.96	0.95	0.96	
PR AUC: 0.36				
cluster	**precision**	**recall**	**f1_score**	**support**
0	0.44	0.49	0.46	2238
2	0.38	0.25	0.30	2737
3	0.33	0.52	0.41	1463

Adenomyosis Cluster-Informed Model; Optimized Threshold: 0.42				
	precision	**recall**	**f1-score**	**support**
0	0.98	0.98	0.98	6232
1	0.37	0.40	0.39	205
accuracy			0.96	
macro avg	0.67	0.69	0.68	
weighted avg	0.96	0.96	0.96	
PR AUC: 0.33				
cluster	**precision**	**recall**	**f1_score**	**support**
0	0.36	0.47	0.41	2217
2	0.24	0.20	0.22	2743
3	0.47	0.44	0.46	1477

Table C 9: Endometriosis and Adenomyosis – Cluster-Informed Model with Distances

IBS Cluster-Informed Model; Optimized Threshold: 0.43				
	precision	recall	f1-score	support
0	0.96	0.92	0.94	4909
1	0.32	0.49	0.38	387
accuracy			0.89	
macro avg	0.64	0.70	0.66	
weighted avg	0.91	0.89	0.90	
PR AUC: 0.34				
cluster	**precision**	**recall**	**f1_score**	**support**
1	0.33	0.57	0.42	1080
2	0.30	0.47	0.36	2700
3	0.31	0.37	0.33	1516

Fibromyalgia Cluster-Informed Model; Optimized Threshold: 0.27				
	precision	recall	f1-score	support
0	0.99	0.99	0.99	6361
1	0.23	0.34	0.28	77
accuracy			0.98	
macro avg	0.61	0.66	0.63	
weighted avg	0.98	0.98	0.98	
PR AUC: 0.19				
cluster	**precision**	**recall**	**f1_score**	**support**
0	0.24	0.48	0.32	2185
2	0.11	0.06	0.08	2774
3	0.26	0.30	0.28	1479

Table C 10: IBS and Fibromyalgia – Cluster-Informed Model with Distances

PCOS Cluster-Informed Model; Optimized Threshold: 0.61				
	precision	**recall**	**f1-score**	**support**
0	0.93	0.95	0.94	4469
1	0.69	0.64	0.67	827
accuracy			0.90	
macro avg	0.81	0.79	0.80	
weighted avg	0.90	0.90	0.90	
PR AUC: 0.734				
cluster	**precision**	**recall**	**f1_score**	**support**
1	0.62	0.58	0.60	1051
2	0.73	0.71	0.72	2749
3	0.70	0.59	0.64	1496

Vaginismus Cluster-Informed Model; Optimized Threshold: 0.26				
	precision	**recall**	**f1-score**	**support**
0	0.99	0.97	0.98	5906
1	0.23	0.38	0.29	122
accuracy			0.96	6028
macro avg	0.61	0.68	0.64	6028
weighted avg	0.97	0.96	0.97	6028
PR AUC: 0.24				
cluster	**precision**	**recall**	**f1_score**	**support**
0	0.14	0.20	0.16	2257
1	0.24	0.31	0.27	1067
2	0.27	0.53	0.36	2704

Table C 11: PCOS and Vaginismus – Cluster-Informed Model with Distances

Vulvodynia Cluster-Informed Model: Optimized Threshold: 0.47				
	precision	**recall**	**f1-score**	**support**
0	0.99	0.99	0.99	5879
1	0.46	0.49	0.48	150
accuracy			0.97	
macro avg	0.73	0.74	0.73	
weighted avg	0.97	0.97	0.97	
PR AUC: 0.43				
cluster	**precision**	**recall**	**f1_score**	**support**
0	0.11	0.06	0.08	2247
1	0.51	0.64	0.57	1095
2	0.23	0.12	0.15	2687

Table C 12: Vulvodynia – Cluster-Informed Model with Distances

Endometriosis NN Model; Optimized Threshold: 0.55				
	precision	**recall**	**f1-score**	**support**
0	0.96	0.97	0.97	2178
1	0.42	0.33	0.37	135
accuracy			0.93	
macro avg	0.69	0.65	0.67	2313
weighted avg	0.93	0.93	0.93	2313
PR AUC: 0.33				
cluster	**precision**	**recall**	**f1-score**	**support**
0	0.60	0.30	0.40	748
1	0.38	0.42	0.40	520
2	0.40	0.24	0.30	612
3	0.37	0.38	0.38	433

Adenomyosis NN Model; Optimized Threshold: 0.55				
	precision	**recall**	**f1-score**	**support**
0	0.97	0.97	0.97	2211
1	0.38	0.38	0.38	102
accuracy			0.95	
macro avg	0.68	0.68	0.68	2313
weighted avg	0.95	0.95	0.95	2313
PR AUC: 0.36				
cluster	**precision**	**recall**	**f1-score**	**support**
0	0.41	0.32	0.36	683
1	0.39	0.53	0.45	332
2	0.00	0.00	0.00	840
3	0.35	0.43	0.39	458

Table C 13: Endometriosis and Adenomyosis – NN Model

IBS NN Model; Optimized Threshold: 0.50				
	precision	recall	f1-score	support
0	0.94	0.91	0.93	2115
1	0.29	0.41	0.34	198
accuracy			0.87	
macro avg	0.62	0.66	0.63	2313
weighted avg	0.89	0.87	0.88	2313
PR AUC: 0.30				
cluster	**precision**	**recall**	**f1-score**	**support**
0	0.31	0.43	0.36	683
1	0.29	0.54	0.38	332
2	0.41	0.26	0.32	840
3	0.21	0.29	0.24	458

Fibromyalgia NN Model; Optimized Threshold: 0.50				
	precision	recall	f1-score	support
0	0.99	0.98	0.98	2256
1	0.31	0.42	0.36	57
accuracy			0.96	
macro avg	0.65	0.70	0.67	2313
weighted avg	0.97	0.96	0.97	2313
PR AUC: 0.28				
cluster	**precision**	**recall**	**f1-score**	**support**
0	0.24	0.44	0.31	683
1	0.35	0.64	0.46	332
2	0.00	0.00	0.00	840
3	0.33	0.14	0.20	458

Table C 14: IBS and Fibromyalgia – NN Model

PCOS NN Model; Optimized Threshold: 0.50				
	precision	recall	f1-score	support
0	0.94	0.87	0.90	1906
1	0.54	0.74	0.63	407
accuracy			0.84	
macro avg	0.74	0.80	0.76	2313
weighted avg	0.87	0.84	0.85	2313
PR AUC: 0.65				
cluster	**precision**	**recall**	**f1-score**	**support**
0	0.62	0.77	0.69	683
1	0.46	0.55	0.50	332
2	0.52	0.84	0.64	840
3	0.50	0.75	0.60	458

Vaginismus NN Model; Optimized Threshold: 0.50				
	precision	recall	f1-score	support
0	0.98	0.98	0.98	2248
1	0.34	0.32	0.33	65
accuracy			0.96	
macro avg	0.66	0.65	0.66	2313
weighted avg	0.96	0.96	0.96	2313
PR AUC: 0.27				
cluster	**precision**	**recall**	**f1_score**	**support**
0	0.00	0.00	0.00	683
1	0.00	0.00	0.00	332
2	0.44	0.24	0.31	840
3	0.52	0.44	0.48	458

Table C 15: PCOS and Vaginismus - NN Model

Vulvodynia NN Model; Optimized Threshold: 0.45				
	precision	**recall**	**f1-score**	**support**
0	0.98	0.94	0.96	2238
1	0.23	0.52	0.32	75
accuracy			0.93	
macro avg	0.61	0.73	0.64	2313
weighted avg	0.96	0.93	0.94	2313
PR AUC: 0.24				
cluster	**precision**	**recall**	**f1-score**	**support**
0	0.20	0.20	0.20	683
1	0.31	0.65	0.42	332
2	0.27	0.31	0.29	840
3	0.17	0.54	0.26	458

Table C 16: Vulvodynia - NN Model

Appendix D: SHAP Summary Plots

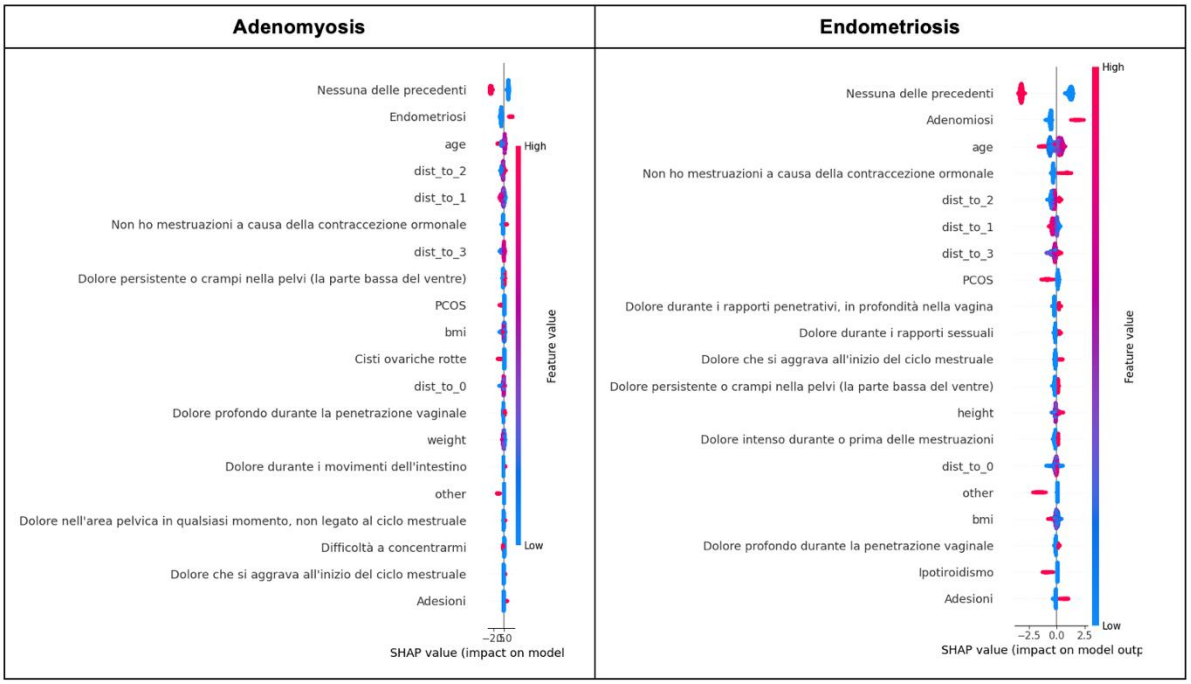

Figure D 1: Adenomyosis and Endometriosis – SHAP values

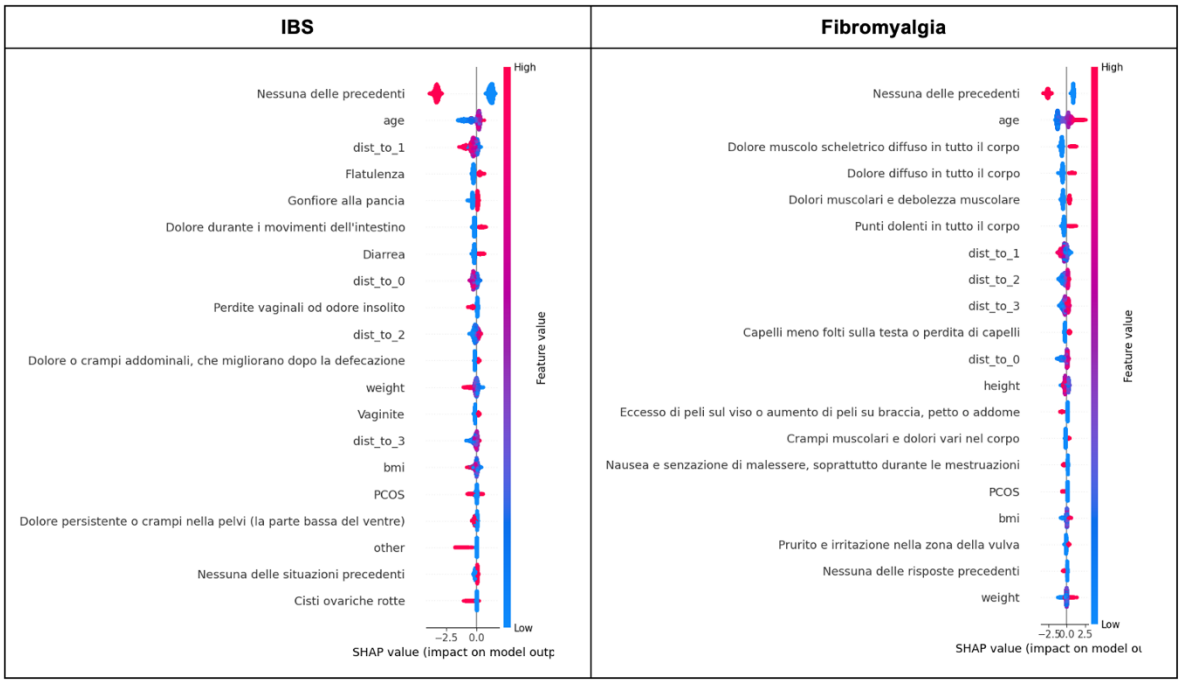

Figure D 2: IBS and Fibromyalgia – SHAP values

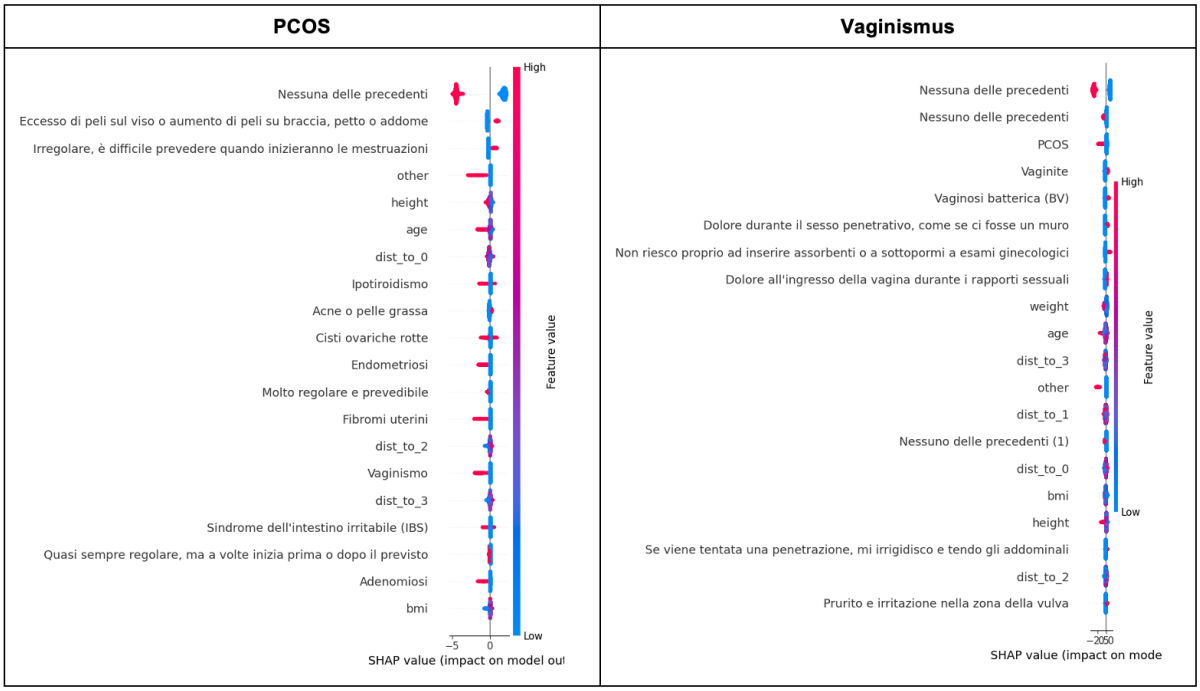

Figure D 3: PCOS and Vaginismus– SHAP values

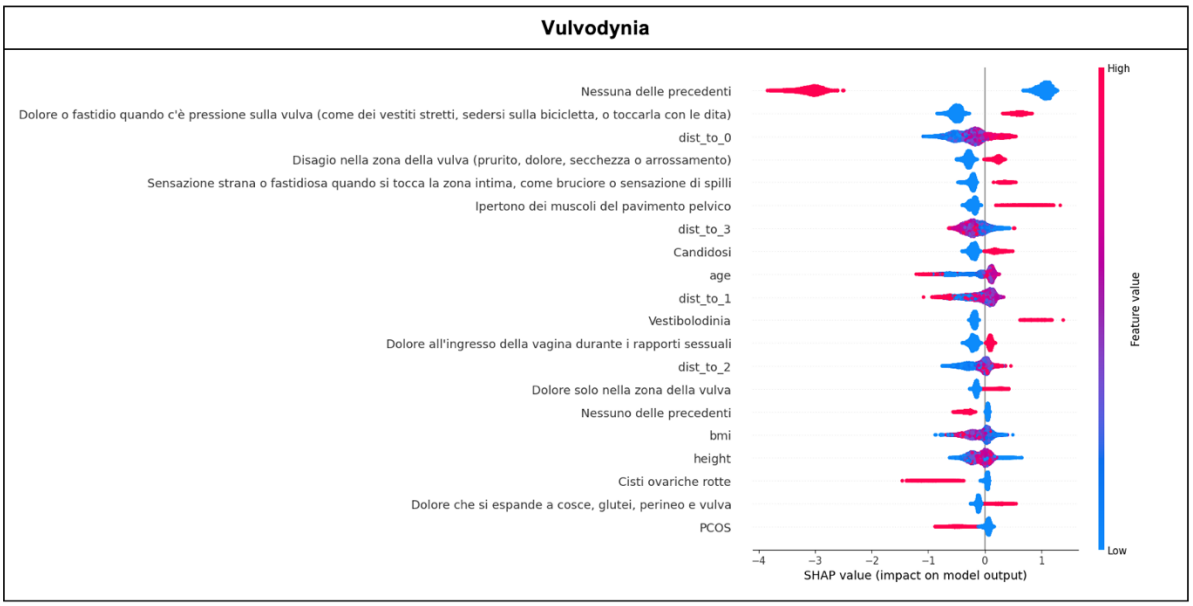

Figure D 4: Vulvodynia – SHAP values

Bibliography

[1] F. G. N. R. L. S. Stefano Sarao Mannelli, "Bias-inducing geometries: exactly solvable data model with fairness implications," in *Geometry-grounded Representation Learning and Generative Modeling at 41 st International Conference on Machine Learning*, Vienna, Austria, 2024.

[2] J. C. C. O. A. R. Georgine Lamvu, "Chronic Pelvic Pain in Women; A Review," *JAMA,* 2021.

[3] S. K. Jane Moore, "Causes of chronic pelvic pain," *WOMEN'S HEALTH MEDICINE,* vol. 2, no. 1, 2006.

[4] Bridget Balch, "Assosication of American Medical Colleges," March 2024. [Online]. Available: https://www.aamc.org/news/why-we-know-so-little-about-women-s-health.

[5] Lindsey Bever, "The Washington Post," 2022. [Online]. Available: https://www.washingtonpost.com/wellness/interactive/2022/women-pain-gender-bias-doctors/.

[6] "World Health Organization," March 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/endometriosis.

[7] J. K. M. S. Kyle Barjon, "Uterine Leiomyomata," *StatPearls Publishing,* 2025.

[8] A. G. A. T. A.-H. M. S. A. G. C. M. A. B. Ali Yosef, "Chronic pelvic pain: Pathogenesis and validated assessment," *Middle East Fertility Society Journal,* 2016.

[9] B. C. Z. R. M. H. T. P. V. C. H. J. G. Jianhui Gao, "What is Fair? Defining Fairness in Machine Learning for Health," 2024.

[10] N. Z. M. D. X. H. Qizhang Feng, "Fair Machine Learning in Healthcare: A Survey," *IEEE Transactions on Artificial Intelligence,* 2020.