A Data-Driven Approach to Predicting Menstrual Irregularity in Women with PCOS and Thyroid Disorders

Tharini Kuzhali Senthilkumar[1,*], Aram Bahrini[1], and Rastin Mastali Majdabadkohne[2]
*tks8@illinois.edu

[1]*Department of Business Administration*
Gies College of Business
University of Illinois at Urbana-Champaign
Champaign, IL, USA

[2]*Primary Care*
Vithas Centro Médico Nerja
Nerja, Málaga, Spain

***Abstract*—Menstrual irregularity is a common health concern among women, frequently associated with endocrine disorders such as polycystic ovary syndrome and thyroid dysfunction. This study examines the relationship between polycystic ovary syndrome status, thyroid-stimulating hormone, luteinizing hormone, follicle-stimulating hormone, insulin resistance, and lifestyle indicators in predicting menstrual irregularity. Using a publicly available dataset of 468 women, of whom 271 had regular cycles and 197 had irregular cycles, three ensemble classifiers, including Random Forest, XGBoost, and Gradient Boosting, were evaluated via five-fold stratified cross-validation. Thyroid-stimulating hormone values were classified into clinical hyperthyroid, normal, and hypothyroid categories and encoded as binary engineered features, expanding the feature set to 16. A key methodological finding was that applying the synthetic minority oversampling technique prior to cross-validation inflated the mean cross-validation AUC from an honest 0.5449 to a misleading 0.7308, an overestimation of 0.1859. Following correction, the honest test-set AUC was 0.6208, with sensitivity of 50.0% and specificity of 64.8%. A three-tier risk stratification system produced a clinically meaningful gradient, with actual irregularity rates of 29.0%, 41.9%, and 56.2% across Low, Moderate, and High tiers, respectively. By integrating predictive modeling with hormonal, metabolic, and clinical data, this study demonstrates a risk-stratification approach for identifying women at elevated risk of menstrual irregularity and highlights the value of transparent, leakage-free methodology in clinical machine learning.**

***Index Terms*—Menstrual irregularity, polycystic ovary syndrome, thyroid dysfunction, risk stratification, clinical decision support**

I. Introduction

Menstrual irregularities are a common reproductive health concern, often associated with underlying hormonal, metabolic, or lifestyle dysfunction. It may present as missed periods, abnormal cycle length, or heavy bleeding and can affect fertility and overall well-being. Polycystic ovary syndrome (PCOS) and thyroid dysfunction are among the leading contributors, as both conditions disrupt endocrine balance and ovulation. Predicting menstrual irregularity is challenging due to the complex interaction of clinical, hormonal, metabolic, and lifestyle factors.

PCOS is a hormonal disorder characterized by elevated androgen levels, which can contribute to irregular or absent periods, lack of ovulation, acne, excess hair growth, and obesity. It affects up to 15% of women of reproductive age and is a leading cause of female infertility [1]. Insulin resistance is a key driver of PCOS, stimulating androgen production and further disrupting ovulation. Chronic low-grade inflammation and metabolic disturbances, such as elevated fasting insulin, are common and contribute to symptom variability.

Thyroid dysfunction can also interfere with menstrual health. Abnormal levels of thyroid-stimulating hormone (TSH) may disrupt hormone regulation and ovulation, resulting in irregular or absent cycles [2]. Some women with PCOS maintain normal thyroid function, but thyroid disorders may coexist with PCOS or act independently, exacerbating menstrual irregularity.

Lifestyle factors play a critical role, as diet, physical activity, sleep quality, and body weight influence hormonal balance. Obesity and high sugar intake can increase insulin resistance and androgen levels, while inadequate exercise or poor sleep may intensify hormonal disruption. Interventions such as maintaining a healthy weight, eating a balanced diet, and engaging in regular physical activity have been shown to improve menstrual regularity in women with PCOS.

Hormonal markers, including luteinizing hormone (LH), follicle-stimulating hormone (FSH), and testosterone, provide additional information about reproductive function. Elevated LH and testosterone are frequently observed in women with irregular cycles. Despite these associations, not all women with PCOS or thyroid dysfunction experience menstrual irregularity, and many without these conditions do, highlighting the multifactorial nature of menstrual health and the difficulty of identifying risk using any single factor alone.

This study develops a machine learning framework to predict menstrual irregularity in women, with particular attention to PCOS and thyroid-related factors. The objectives are to

estimate the probability of irregular cycles, classify individuals into low, moderate, and high-risk categories, and identify key clinical, hormonal, metabolic, and lifestyle contributors. By integrating predictive modeling with clinical and lifestyle data, this research aims to support risk stratification, inform earlier clinical assessment, and demonstrate how routinely collected variables may be combined to identify women at elevated risk of menstrual irregularity.

II. Methodology

A. Dataset

The dataset was obtained from Kaggle and included 468 women patients with 52 features encompassing hormonal assays, metabolic markers, anthropometric measurements, and self-reported lifestyle information [3]. The data were complete, with no missing values or duplicate records. Menstrual irregularity was defined using established clinical criteria: a cycle length of less than 21 days or more than 35 days, cycle-to-cycle variability exceeding 7 to 9 days, or the absence of menstruation for more than 90 days [4]. In the dataset, this outcome was represented as a binary variable, where 0 indicates a regular menstrual cycle, and 1 indicates an irregular cycle meeting one or more of these criteria. One variable, `Menstrual_Cycle_Length_days`, was excluded from the analysis because it directly reflects cycle behavior and would introduce target leakage if retained. Additionally, PCOS diagnosis was recorded as a binary variable, and the specific diagnostic criteria applied to each patient were not documented in the source data. In clinical practice, PCOS is commonly diagnosed according to the Rotterdam criteria, which require the presence of at least two of the following three features: oligo-ovulation or anovulation, clinical or biochemical signs of hyperandrogenism, and polycystic ovarian morphology on ultrasound [4]. The absence of phenotypic subclassification in this dataset introduces heterogeneity into the PCOS label, as patients diagnosed under different criteria may exhibit substantially different hormonal and metabolic profiles. This variability likely contributes to the relatively modest predictive contribution of the PCOS diagnostic label compared with continuous metabolic markers such as homeostatic model assessment for insulin resistance (HOMA-IR), which capture underlying insulin resistance regardless of how PCOS was formally diagnosed.

B. Preprocessing

The following preprocessing steps were applied sequentially to ensure a leakage-free pipeline. First, a stratified 80/20 train/test split was applied (`random_state=42`), preserving the 57.9%/42.1% class ratio in both partitions and yielding 374 training and 94 test patients. Stratification was used to ensure the minority class was proportionally represented in both sets, preventing evaluation bias. Second, `StandardScaler` was fit exclusively on the training set and applied to the test set only via `transform`, preventing the test set's mean and variance from influencing feature scaling. Third, feature engineering was applied post-split using a dedicated function called separately on training and test sets, ensuring no derived features carried information across the split boundary. Finally, Synthetic Minority Oversampling Technique (SMOTE) was applied within an `imblearn` pipeline and constrained to training folds only during cross-validation [5]. Applying SMOTE before fold construction allows synthetic samples derived from real training data to appear in validation folds, violating the independence assumption and producing inflated area under the curve (AUC) estimates, as demonstrated in Section IV.

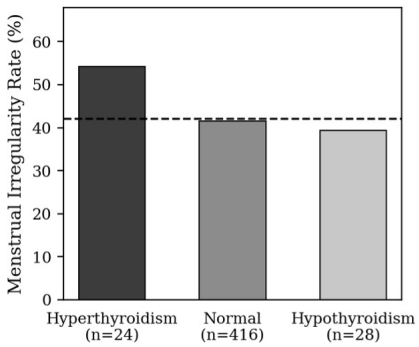

Fig. 1. Menstrual irregularity rate by clinical thyroid status category. Hyperthyroid patients show the highest rate (54.2%), exceeding the overall cohort rate of 42.1% (dashed line). Hypothyroid patients (n=28) exhibited 39.3%.

C. Thyroid Status Classification

TSH values were classified into three clinical categories using established endocrinological thresholds: hyperthyroid (TSH < 0.4 uIU/mL), normal (TSH 0.4–4.0 uIU/mL), and hyperthyroid (TSH > 4.0 uIU/mL) [6], [7]. Exploratory analysis of the 468-patient dataset revealed 24 patients with hyperthyroid status (54.2% menstrual irregularity rate), 416 patients with normal thyroid function (41.6% irregularity rate), and 28 patients with hyperthyroid status. Notably, hypothyroid patients exhibited the highest irregularity rate in the cohort, consistent with the physiological role of thyroid dysfunction in disrupting the hypothalamic-pituitary-gonadal axis. These clinical categories were subsequently encoded as binary engineered features to capture their predictive contribution in the model.

D. Feature Engineering

Sixteen composite biomarker features were engineered across the hormonal, metabolic, lifestyle, and thyroid domains to capture interaction effects beyond those of raw clinical variables. The 14 core features spanned hormonal ratios, insulin resistance composites, and lifestyle risk indices. Two additional binary features, `Thyroid_Hypo` (TSH > 4.0) and `Thyroid_Hyper` (TSH < 0.4), were derived from the clinical thyroid classification, with Normal thyroid status serving as the reference baseline (dropped to avoid multicollinearity). Table I summarizes all engineered features.

TABLE I
ENGINEERED COMPOSITE FEATURES

Feature	Domain	Formula / Logic
Androgen_Burden	Hormonal	Testosterone / (SHBG + 1)
LH_FSH_Gap	Hormonal	LH − FSH
Free_Androgen_Index	Hormonal	(Testosterone ×100) / (SHBG + 1)
Prolactin_Estradiol_Ratio	Hormonal	Prolactin / (Estradiol + 1)
Estrogen_Progest_Ratio	Hormonal	Estradiol / (Progesterone + 1)
IR_Severity_Score	Metabolic	(HOMA-IR × Fasting Insulin) / 10
Glucose_Dysregulation	Metabolic	Fasting Glucose × HbA1c
Lipid_Risk_Ratio	Metabolic	Triglycerides / (HDL + 1)
MetSyn_Score	Metabolic	Waist Circumference × HOMA-IR / 100
BMI_IR_Index	Metabolic	HOMA-IR / (BMI + 1)
Lifestyle_Risk_Score	Lifestyle	Sum of 5 binary lifestyle risk flags
Sleep_IR_Interaction	Lifestyle	HOMA-IR / (Sleep Hours + 1)
Sugar_IR_Score	Lifestyle	Dietary Sugar × HOMA-IR
Inflammatory_Lifestyle	Lifestyle	(Smoking + Alcohol) × CRP
Thyroid_Hypo	Thyroid	1 if TSH > 4.0 uIU/mL, else 0
Thyroid_Hyper	Thyroid	1 if TSH < 0.4 uIU/mL, else 0

E. Models

Three ensemble classifiers were trained and evaluated. Random Forest was selected for its robustness to overfitting on moderate-sized clinical datasets. It operates by constructing multiple independent decision trees in parallel, each trained on a random bootstrap sample of the data, and then aggregating their predictions via majority voting. This bagging approach reduces variance without significantly increasing bias, making it well-suited for clinical datasets with noisy or correlated features. The model was configured with 700 estimators, a maximum depth of 8, balanced class weights, and `log2` maximum features.

XGBoost was chosen for its gradient-boosted regularization and efficiency on structured tabular data. Unlike Random Forest, XGBoost builds trees sequentially, with each tree correcting the residuals of the previous one. It incorporates L1 and L2 regularization to prevent overfitting and handles class imbalance effectively through its `scale_pos_weight` parameter. The model was implemented with 1,000 estimators, a learning rate of 0.01, a maximum depth of 7, a column subsampling rate of 0.7, and a row subsampling rate of 0.9.

Gradient Boosting was included as a sequential ensemble baseline. Similar in principle to XGBoost but without the advanced regularization framework, it builds trees one at a time to minimize a differentiable loss function. It was configured with 300 estimators, a learning rate of 0.05, and a maximum depth of 5. Hyperparameter optimization for Random Forest and XGBoost was conducted using `RandomizedSearchCV` with 50 candidate parameter combinations and five-fold stratified cross-validation.

F. Evaluation

The primary evaluation metric was the receiver operating characteristic–area under the curve (ROC–AUC), selected for its robustness to class imbalance and its ability to measure model discrimination across all classification thresholds. Secondary metrics included sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV), as each captures a clinically distinct aspect of model performance. Sensitivity reflects the model's ability to correctly identify women at risk, while specificity reflects its ability to correctly rule out irregularity in unaffected women. The optimal classification threshold was determined using Youden's J statistic (J = Sensitivity + Specificity − 1), which maximizes the sum of sensitivity and specificity across all thresholds [8].

III. HYPOTHESES

During the initial analysis of the dataset, three key questions were identified regarding model performance, clinical risk patterns, and methodological integrity. Exploratory findings, including compressed class separation in hormonal markers and mild overall class imbalance, provided the foundational motivation for these hypotheses. We examine three hypotheses to explore potential statistically significant patterns:

Hypothesis 1: *XGBoost is expected to outperform Random Forest and Gradient Boosting in predicting menstrual irregularity risk, as reflected by the highest ROC-AUC score, given its regularized gradient-boosting framework and strong performance on structured clinical data.*

Hypothesis 2: *The model will assign significantly higher predicted risk of menstrual irregularity to women with elevated TSH levels, indicating hypothyroid or hyperthyroid status, and/or a confirmed PCOS diagnosis, compared to women with normal thyroid function and without PCOS.*

Hypothesis 3: *Classifying patients into clinical thyroid status categories (hypothyroid, hyperthyroid, and normal) and encoding them as binary engineered features will improve model predictive performance compared to using raw TSH values alone.*

IV. RESULTS

A. Model Comparison

Initial five-fold stratified cross-validation (with SMOTE leakage present) produced the results shown in Table II. Among the untuned models, XGBoost achieved the highest CV AUC of 0.7257 (±0.0235), followed by Random

TABLE II
INITIAL MODEL COMPARISON UNDER PRE-CORRECTION CROSS-VALIDATION CONDITIONS, INCLUDING LEAKAGE-AFFECTED CV AUC VALUES AND TUNED MODEL RESULTS.

Model	CV AUC	Std Dev
Random Forest	0.7238	±0.0488
XGBoost	0.7257	±0.0235
Gradient Boosting	0.6902	±0.0344
Random Forest (Tuned)	0.7308	—
XGBoost (Tuned)	0.7287	—

TABLE III
QUANTIFICATION OF AUC INFLATION CAUSED BY APPLYING SMOTE BEFORE, RATHER THAN WITHIN, CV FOLDS.

Condition	AUC
CV with SMOTE leakage (inflated)	0.7308
CV with SMOTE in pipeline (honest)	0.5449
Inflation due to SMOTE leakage	+0.1859
Honest test-set AUC (held-out)	0.6208

Forest at 0.7238 (±0.0488), with Gradient Boosting lowest at 0.6902 (±0.0344). Following hyperparameter tuning via `RandomizedSearchCV`, the best Random Forest AUC was 0.7308 and the best XGBoost AUC was 0.7287. Under these pre-correction conditions, the tuned Random Forest slightly exceeded the tuned XGBoost model. Contrary to Hypothesis 1, Random Forest outperformed XGBoost, although this comparison should be interpreted with caution, as the cross-validation estimates were affected by SMOTE leakage. One possible explanation is that its ensemble averaging yielded more stable generalization on this moderate-sized dataset. Hypothesis 1 is therefore rejected.

B. SMOTE Leakage — Critical Finding

A critical methodological finding was identified by comparing cross-validation AUC under two conditions: (1) SMOTE applied to the full training set before fold construction, and (2) SMOTE constrained within each training fold via an `imblearn` pipeline. When SMOTE was applied prior to cross-validation, synthetic patients derived from real training samples appeared in validation folds, violating the independence assumption and inflating AUC. The corrected honest cross-validation scores were 0.6334, 0.6481, 0.4520, 0.5342, and 0.4569 across the five folds (mean = 0.5449, SD = 0.080). Table III summarizes the inflation.

C. Honest Test-Set Performance

The corrected Random Forest pipeline, incorporating binary thyroid-status features, was evaluated on the 94-patient held-out test set. All metrics reflect performance on real, unseen patients only. The optimal classification threshold of 0.429 was determined by Youden's J statistic. Table IV presents the full confusion matrix metrics. Sensitivity was 50.0%, indicating the model correctly identified 20 of 40 truly irregular patients. Specificity was 64.8%, indicating that the model correctly classified 35 of 54 regular patients. Importantly, the addition

TABLE IV
HONEST TEST-SET PERFORMANCE METRICS

Metric	Value
ROC-AUC	0.6208
Optimal Threshold (Youden's J)	0.429
Sensitivity (Recall)	50.0%
Specificity	64.8%
PPV (Precision)	51.3%
NPV	63.6%
True Positives	20
True Negatives	35
False Positives	19
False Negatives	20

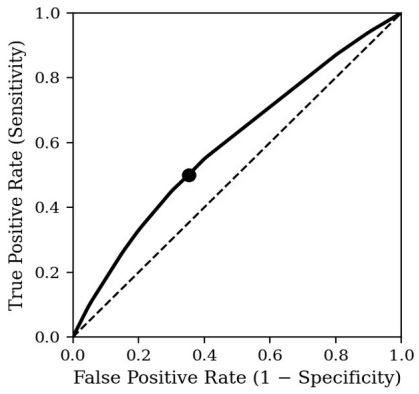

Fig. 2. ROC curve for the corrected Random Forest classifier on the 94-patient held-out test set (ROC–AUC = 0.6208). The operating point at threshold 0.429, selected by Youden's J statistic, is indicated.

of `Thyroid_Hypo` and `Thyroid_Hyper` binary features improved test-set AUC from 0.6088 to 0.6208, a gain of 0.012, supporting Hypothesis 3 and suggesting that clinical thyroid status classification contributes additional predictive signal beyond raw TSH values alone.

D. Risk Stratification

The model's predicted probabilities ranged from 0.234 to 0.679 (mean = 0.4624), reflecting a moderately compressed distribution characteristic of clinical risk models on heterogeneous populations. Compared to the pre-thyroid-feature model (range: 0.249–0.634), the widened probability range suggests greater separation in predicted risk after incorporating thyroid status. Percentile-based thresholds, Low: $p < 0.429$ (33rd percentile), Moderate: $0.429 \leq p < 0.515$, High: $p \geq 0.515$ (66th percentile), were adopted to ensure balanced tier assignment. Table V presents the stratification results.

Hypothesis 2 received partial support. PCOS diagnosis and TSH appeared among the top predictive features in Random Forest feature importances, consistent with their plausible clinical relevance. Moreover, hyperthyroid patients demonstrated the highest cohort irregularity rate (54.2%), exceeding even PCOS-positive patients, highlighting the importance of thyroid function as a clinical predictor. However, the separation in irregularity rate between PCOS-positive (43.3%) and PCOS-negative (41.2%) women was only 2.1 percentage points in

TABLE V
OBSERVED IRREGULARITY RATES ACROSS MODEL-DERIVED RISK TIERS IN THE HELD-OUT TEST SET.

Tier	n	% Test Set	% Truly Irregular
Low	31	33.0%	29.0%
Moderate	31	33.0%	41.9%
High	32	34.0%	56.2%

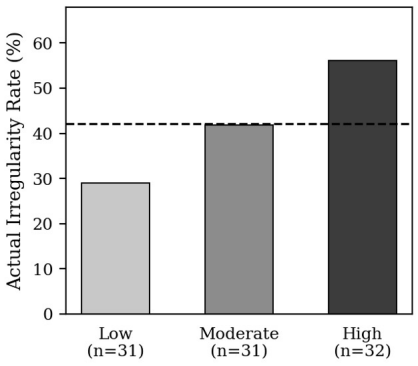

Fig. 3. Observed menstrual irregularity rates across the three risk tiers (n = 94 test patients). The dashed line indicates the test-set base rate of 42.1%. Percentile-based thresholds were defined as Low ($p < 0.429$), Moderate ($0.429 \leq p < 0.515$), and High ($p \geq 0.515$).

the exploratory data analysis (EDA), suggesting that metabolic markers, specifically HOMA-IR, IR_Severity_Score, and Androgen_Burden, may contribute more strongly to model prediction than the PCOS diagnostic label alone.

The 27.2 percentage point spread from Low (29.0%) to High (56.2%) confirms a clinically meaningful risk gradient. A high-risk tier assignment corresponds to an observed irregularity rate of 56.2%, compared with the overall base rate of 42.1%. The top-ranked features by importance were HOMA-IR, `IR_Severity_Score`, `Androgen_Burden`, `Free_Androgen_Index`, `PCOS_Diagnosis`, `TSH_uIU_mL`, BMI, and `LH_FSH_Gap`.

E. Clinical Dashboard and Patient Reporting

A four-panel Plotly dashboard was developed to support interpretation of model outputs, including the distribution of risk tiers across the 94 test patients, a predicted probability histogram with threshold markers at 0.429 and 0.515 defining tier boundaries, a ranked bar chart of the top 15 feature importances, and a comparison of actual irregularity outcomes across risk tiers, allowing clinicians to quickly assess both overall model behavior and individual risk patterns. In addition, an interactive `ipywidgets`-based patient explorer was created to support dual-audience use. In educational mode, it provides plain-language explanations of a patient's predicted risk level along with personalized, actionable recommendations covering weight management, dietary changes, physical activity, and sleep hygiene, adapted from the 2023 International PCOS Guidelines [9]. In clinical mode, it displays the patient's predicted probability, assigned risk tier, key clinical variables including TSH, HOMA-IR, BMI, and PCOS status, as well as the true outcome label for comparison. Both modes update dynamically in real time when a patient is selected via dropdown or slider, making the tool suitable for both research and clinical exploration.

To improve clinical interpretability, the interface linked each risk tier to an illustrative management pathway rather than a prescriptive recommendation set. Patients in the low-risk group were mapped to routine follow-up, standard lifestyle counselling, and annual cycle monitoring; those in the moderate-risk group were mapped to targeted lifestyle interventions such as weight management, dietary modification, and regular physical activity, along with hormonal screening at the next clinical visit and follow-up within three to six months if symptoms persisted; and patients in the high-risk group were mapped to endocrinology referral, a comprehensive hormonal panel including TSH, LH, FSH, testosterone, and prolactin, and further evaluation for possible PCOS or thyroid dysfunction under specialist supervision. These pathways were intended as example decision-support pathways rather than validated care recommendations, and they should not replace physician judgment or be used as a standalone basis for clinical decision-making.

V. DISCUSSION

A. Limitations

The dataset used in this study lacks longitudinal menstrual cycle data, which represents a significant limitation given that menstrual irregularity is inherently dynamic. A cross-sectional snapshot does not capture cycle trajectories, seasonal variations, or hormonal trends over time. In addition, lifestyle variables such as sleep duration, dietary sugar intake, physical activity, smoking, and alcohol consumption were self-reported, introducing the potential for recall and social desirability biases. The moderate sample size ($n = 468$) further limits generalizability across diverse clinical populations, particularly within the thyroid subgroups, which included only 24 hyperthyroid and 28 hypothyroid patients, rendering subgroup analyses statistically underpowered. Moreover, thyroid status was classified solely using TSH values, as free thyroxine (FT4) and free triiodothyronine (FT3) measurements were unavailable in the dataset. Although TSH is widely used as a first-line screening marker, the absence of these hormones reduces diagnostic precision, especially in subclinical cases where TSH may fall within the normal range despite underlying dysfunction. Consequently, the thyroid classifications in this study represent a screening-level assessment rather than a definitive clinical diagnosis, and future research should incorporate a comprehensive thyroid panel to enhance classification accuracy. Finally, the absence of external validation means that test-set performance on a truly independent cohort remains unknown, and the percentile-based risk thresholds were internally calibrated to this dataset, thereby requiring prospective clinical validation before real-world deployment.

B. Future Work

The next steps for this research include collecting longitudinal hormonal and cycle-tracking data to better understand how menstrual cycles change over time, using approaches such as long short-term memory (LSTM) networks. This could allow future models not only to model cycle dynamics but also to identify when menstrual irregularity is likely to occur. Alongside cycle dates, tracking menstrual flow, including intensity, duration, and variability, would provide a more complete picture of menstrual health. Expanding to a larger, multi-site dataset, ideally with more than 2,000 participants, would improve generalizability and allow analysis of different PCOS phenotypes, thyroid disorder severity, and overlapping conditions such as endometriosis, which frequently coexists with PCOS and thyroid disorders, particularly hypothyroidism and autoimmune thyroiditis. Existing lifestyle factors in the dataset, including nutrition, exercise, stress, and sleep, could also be further explored to uncover patterns contributing to irregularity. Application programming interfaces (APIs) could be used to retrieve hormonal, cycle, or lifestyle data directly from tracking applications or wearable devices, enabling more automated and near-real-time data collection. While TSH is commonly used to assess thyroid function, the absence of FT3 and FT4 measurements in our dataset limits our ability to detect subclinical thyroid dysfunction, which may be identified only through more comprehensive thyroid testing. Making the dashboard more interactive through a Streamlit or Dash web application would help clinicians and patients explore risk patterns. Finally, testing the system in real clinical settings is essential before it can be deployed as a reliable decision-support tool.

C. Conclusion

This study makes three primary contributions. From a methodological perspective, it demonstrates that applying SMOTE prior to cross-validation can inflate the AUC by 0.1859, substantially overestimating model performance and potentially leading to misleading clinical conclusions if not properly addressed. From a clinical standpoint, the inclusion of TSH-derived thyroid status features improved the test-set AUC from 0.6088 to 0.6208 and produced a clearer risk gradient, supporting the relevance of thyroid function in predicting menstrual irregularity. The proposed three-tier risk stratification system further highlights the model's potential practical value, with a 27.2 percentage point increase in irregularity observed across risk tiers.

It is important to interpret the reported ROC-AUC of 0.6208 within the appropriate clinical context; while this value reflects modest discriminative ability at the individual patient level, the model is not intended to function as a standalone diagnostic tool but rather as a population-level risk stratification instrument designed to identify women who may benefit from closer monitoring or earlier clinical evaluation. At the individual level, a moderate AUC implies the likelihood of both false positives and false negatives, underscoring the importance of clinical judgment in decision-making. The strength of the model lies in its ability to systematically identify higher-risk subgroups using routinely collected clinical data, thereby enabling more efficient and targeted allocation of healthcare resources.

In addition, the prominence of insulin resistance-related features among the top predictors reinforces the well-established link between metabolic dysfunction and anovulatory cycles in PCOS. Overall, this study shows that machine learning can identify interpretable risk patterns in routinely collected clinical data, while also highlighting the importance of leakage-free evaluation and careful clinical interpretation in the development of decision-support tools.

References

[1] Cleveland Clinic, "Polycystic Ovary Syndrome (PCOS)," Cleveland Clinic. Accessed: Apr. 2026. [Online]. Available: https://my.clevelandclinic.org/health/diseases/8316-polycystic-ovary-syndrome-pcos

[2] National Library of Medicine, "Menstrual Irregularities and Thyroid Dysfunction," PMC11259460. Accessed: Apr. 2026. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11259460/

[3] Sy, Michael Mendiola. "PCOS Clinical Dataset." Kaggle, December 27, 2025. [Online]. Available: https://www.kaggle.com/datasets/michaelmendiolasy/pcos-clinical-dataset.

[4] Christ, John P., and Marcelle I. Cedars. "Current Guidelines for Diagnosing PCOS." Diagnostics 13, no. 6 (2023): 1113.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-Sampling Technique, "Journal of Artificial Intelligence Research", vol. 16, pp. 321–357, 2002.

[6] UCLA Health, "Normal Thyroid Hormone Levels, " Endocrine Surgery. Accessed: Apr. 2026. [Online]. Available: https://www.uclahealth.org/medical-services/surgery/endocrine-surgery/conditions-treated/thyroid/normal-thyroid-hormone-levels

[7] Numan Diagnostics, "Thyroid Test Results Chart. " Accessed: Apr. 2026. [Online]. Available: https://www.numan.com/diagnostics/hormone-health/thyroid-test-results-chart

[8] N. Smits, "A Note on Youden's J and Its Cost Ratio," *BMC Medical Research Methodology*, vol. 10, p. 89, 2010.

[9] H. J. Teede *et al.*, "Recommendations from the 2023 International Evidence-based Guideline for the Assessment and Management of Polycystic Ovary Syndrome," *Human Reproduction*, vol. 38, no. 9, pp. 1655–1679, 2023.

[10] National Guideline Centre (UK), *Thyroid Disease: Assessment and Management*. London, U.K.: National Institute for Health and Care Excellence (NICE), 2019.