Context-Aware Hierarchical Bayesian Modeling of IVF Laboratory Environmental Conditions

Zahra Asghari Varzaneh [1] **Reza Khoshkangini** [1] **Pia Saldeen** [2] **Lars Johansson** [3] **Thomas Ebner** [4]

Abstract

IVF pregnancy rates are routinely modeled using patient-level variables, while high-resolution laboratory environmental data remain underutilized. We show that this is a missed opportunity. Rather than relying on raw sensor averages, we engineer 55 context-aware temporal features including rolling thermal stability, simultaneous temperature-humidity adherence, peak stress duration, and post-stress recovery speed that capture the dynamics of incubator microenvironments. On 61 weeks of data from an Asian IVF clinic, these features reduce cross-validated prediction error to 1.27%, compared to 3–5% for raw averages. We then train a hierarchical Bayesian Beta regression model that shares environmental effects across an Asian and a Northern European clinic via partial pooling, while preserving site-specific baselines. On held-out data from the Northern European clinic, the model achieves $R^2 = 0.86$ and a 64% error reduction for the 35–39 age group over a naive baseline,demonstrating that structured environmental monitoring contains clinically meaningful, transferable signal.

1. Introduction

The success of IVF depends on several interacting factors, including the patient's age, ovarian reserve, sperm quality, lifestyle, the stimulation protocol, embryo quality, etc (Younglai et al., 2005). In addition, the physical environmental conditions of the laboratory, like temperature, humidity, and air quality, play an important role. During the culture period, embryos are kept inside incubators that tightly control their internal environment (Mantikou et al., 2013; Varzaneh et al., 2025). However, these incubators do not operate in complete isolation. Each time an incubator door is opened, room air enters, and the time needed for the incubator to return to its optimal condition depends directly on how stable the surrounding laboratory environment is. Despite this, environmental monitoring in IVF laboratories is often treated primarily as a compliance exercise: sensors trigger alerts when fixed thresholds are breached, yet the rich, high-resolution sensor data are rarely analyzed statistically. This represents a missed opportunity. Prior work has focused on patient-level predictors (Kupka et al., 2016), suggesting the influence of environmental and procedural factors. Complementary controlled experiments, such as (Swain et al., 2012), demonstrate that even minor changes in culture conditions, such as oxygen levels, temperature stability, and media pH, can directly affect embryo development and clinical outcomes. However, these studies typically examine single variables under idealized conditions and do not capture the complex, time-varying fluctuations present in routine clinical practice. As a result, ignoring high-resolution environmental data may miss the cumulative effects of real-world stressors that help explain differences between laboratories.

We address this gap with two contributions:

First, we introduce *context-aware features* computed from sensor data that encode temporal patterns invisible to raw statistics: rolling thermal stability, the fraction of time temperature and humidity are simultaneously within their biological ideal ranges, the longest uninterrupted stress episode, post-stress recovery speed, and first week transfer-window summaries.

Second, we develop a *hierarchical Bayesian Beta regression* model to analyze IVF outcomes across two clinics (Asian and Northern European). The model uses partial pooling to share information about environmental effects between the two sites, while allowing each clinic to maintain its own baseline level. In this way, data from the Asian clinic helps the model learn general and stable relationships between environmental conditions and IVF outcomes, which in turn improves prediction performance on the Northern European dataset without assuming that the two clinics are identical.

This study also address three key questions:

[1]Department of Computer Science and Media Technology, Malmö University, Malmö, Sweden [2]Nordic IVF, Sweden [3]NewLifeAid-Global AB, Sweden [4]Kepler Universitätsklinikum, Linz, Austria. Correspondence to: Zahra Asghari Varzaneh <zahra.asghari-varzaneh@mau.se>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

(i) Do context-aware environmental features outperform simple aggregated sensor measures?
(ii) Which features are most important for prediction, and are these effects consistent across the two clinics?
(iii) Does sharing information between clinics through partial pooling improve performance compared to a single-site?

2. Data

Environmental sensor data were collected at two IVF laboratories at 10-minute intervals during 2024–2025. We refer to the Asian clinic as the **source** clinic because it provides substantially more labelled data (61 weeks) and serves as the basis for feature selection and partial-pooling priors; the Northern European clinic is the **target**, representing the site we aim to generalise to. Both stations record temperature (°C), relative humidity (%), CO_2 (ppm), and TVOC (ppb). Pregnancy rates were provided for patients under 35, aged 35–39, and aged 40 and above. Table 1 summarises the training period. The Asian record spans January 2024–October 2025, interrupted by a 196-day collection gap caused by sensor data unavailability at the clinic; this splits it into two continuous segments of 19 and 42 weeks (61 total). The Northern European clinic provides 14 months (November 2024–December 2025). The large standard deviations in the Northern European data mostly come from months with only one to three patients in an age group, where the reported rates are driven more by small numbers than by environmental conditions.

Data structure. Environmental variables are recorded every 10 minutes. Pregnancy rates were reported at different temporal granularities by the two clinics: weekly for Asia and monthly for Northern Europe, reflecting the format in which outcome data were made available to us by each laboratory. Environmental features are therefore aggregated to match each clinic's reporting interval (weekly for Asia, monthly for Northern Europe), so that features and outcomes are always aligned at the same resolution. The 196-day gap in the Asian data is handled by treating the two segments as independent time series during cross-validation, with fold boundaries never placed across the gap to avoid information leakage; the autoregressive term $\gamma\,\mathrm{logit}(y_{t-1,c})$ is reset at the start of each segment. A further challenge is that outcomes at time T reflect conditions from roughly two to six weeks earlier. Since exact embryo-level timestamps are unavailable, we approximate this delay by including lagged environmental summaries from the previous period $(T-1)$, providing a practical alignment between past conditions and observed outcomes.

3. Method

The pipeline has four steps (see Figure 1): (1) context-aware feature engineering from 10-minute readings; (2) aggregation to weekly resolution for Asia and monthly for Northern Europe; (3) SHAP-based selection of the top-16 features per age group on Asian data to keep sampling tractable; (4) joint modelling of both clinics using a hierarchical Bayesian Beta regression.

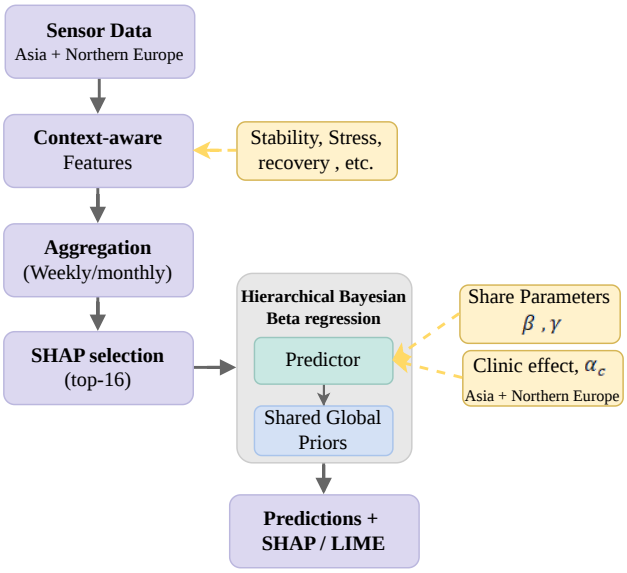

Figure 1. From sensor data to predictions. Context-aware features are extracted and aggregated over time, then fed into a hierarchical Bayesian model that shares information across clinics while preserving site-specific differences

Step 1: Context-aware features. For temperature T_t with optimum $T^* = 22.5\,°\mathrm{C}$, the stress index is $s_t^T = \max(0,\ |T_t - T^*| - 1.5)/3$; an analogous expression governs humidity ($H^* = 42.5\%$, deadband $\pm 2.5\%$). The composite comfort index $c_t = (1 - 0.5\,s_t^T - 0.5\,s_t^H)^+$ summarises environmental quality. Beyond these pointwise quantities, we compute: the rolling 1-hour temperature standard deviation σ_t^T (and its period mean and stable fraction below 0.5 °C) as a measure of thermal stability; the ideal-zone adherence $z_t = \mathbf{1}[T_t \in [21, 24]°\mathrm{C}] \cdot \mathbf{1}[H_t \in [40, 45]\%]$ capturing simultaneous optimality; the longest consecutive stress episode;

Table 1. Environmental conditions and pregnancy rates: mean±std, training period.

Variable	Asia	Northern Europe	Unit
Environmental			
Temperature	23.4 ± 0.6	24.0 ± 0.6	°C
Humidity	37.2 ± 9.8	42.8 ± 8.3	%
CO_2	168 ± 22	189 ± 18	ppm
TVOC	148 ± 420	312 ± 380	ppb
Pregnancy rates by age group			
PR (<35 yrs)	44.1 ± 3.1	38.6 ± 18.4	%
PR (35–39 yrs)	26.6 ± 2.8	34.7 ± 12.1	%
PR (40+ yrs)	13.9 ± 1.8	16.2 ± 13.7	%

the recovery score $r = \mathbb{E}[c_{t+1} \mid c_t < 0.5]$; Short-term lags (1–3 hours) capture local dynamics, while summaries from the preceding period ($T-1$) provide coarse temporal alignment with outcomes. The final feature set includes 55 context-aware variables per period.

Step 2: Temporal aggregation. Asia segments are aggregated to weekly resolution independently; the Northern European data to a monthly resolution. This yields 55 features per period.

Step 3: SHAP feature selection. Given the small number of Northern European training months, using all features would lead to an unstable Bayesian model. We therefore train an XGBoost model (Chen and Guestrin, 2016) on the Asian data and select the top-16 features based on mean absolute SHAP importance for each age group (Lundberg and Lee, 2017). This reduces dimensionality while retaining the most informative predictors.

Step 4: Hierarchical Bayesian Beta regression (Bhuwalka et al., 2023). Let $y_{t,c} \in (0,1)$ denote the pregnancy rate for clinic c at time t. Since the outcome is a proportion bounded between 0 and 1, we model it using a Beta distribution:

$$y_{t,c} \sim \text{Beta}(\mu_{t,c}\phi,\ (1-\mu_{t,c})\phi). \tag{1}$$

The expected value $\mu_{t,c}$ is linked to the predictors through:

$$\eta_{t,c} = \alpha_c + \mathbf{x}_t^\top \boldsymbol{\beta} + \gamma \,\text{logit}(y_{t-1,c}), \quad \mu_{t,c} = \sigma(\eta_{t,c}). \tag{2}$$

As shown in Eq. (2), α_c is a clinic-specific intercept that captures baseline differences in pregnancy rates between clinics. The coefficient vector $\boldsymbol{\beta}$ is shared across clinics and represents the common effect of environmental features. The autoregressive term $\gamma \,\text{logit}(y_{t-1,c})$ accounts for temporal dependence, reflecting that current outcomes are influenced by previous ones.

To enable partial pooling across clinics, we place a hierarchical prior on the intercepts:

$$\alpha_c \sim \mathcal{N}(\mu_G, \sigma_G). \tag{3}$$

where μ_G is informed by approximate age-specific clinical pregnancy rates from the general IVF literature (Appendix A), and σ_G controls how much clinics may deviate from this shared baseline. σ_G is inferred from data, yielding automatic partial pooling without a fixed pooling decision. We use weakly informative priors for the remaining parameters in Eq. (4). Posterior inference uses NUTS (Hoffman & Gelman, 2014) in PyMC (Salvatier et al., 2016) (2 chains, 1500 warm-up + 1500 draws, target acceptance 0.95).

$$\begin{aligned} \boldsymbol{\beta} &\sim \mathcal{N}(\mathbf{0}, 0.09I), \\ \gamma &\sim \mathcal{N}(0.3, 0.09), \\ \phi &\sim \text{Gamma}(3, 0.1) \end{aligned} \tag{4}$$

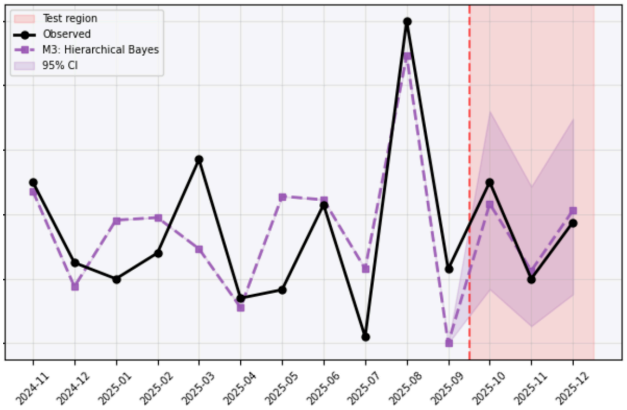

Figure 2. PR 35–39: observed (black) vs. HM posterior mean (dashed). Shading = 95% credible interval. Pink = test period.

Feature selection for the Bayesian model. With 11 training months and 55 features, directly fitting all predictors yields a non-identified posterior. We select the top-16 features per age group by mean absolute SHAP importance computed on the Asian data (via XGBoost).

4. Results

Setup. Northern European months 1–11 are used for training; months 12–14 (October–December 2025) are held out. The decision to use Asia as source and Northern Europe as target rather than a location-independent split is motivated by the asymmetry in available labelled data: the Asian clinic provides 61 weeks sufficient for cross-validation, while the Northern European clinic has only 14 months in total. A location-independent split would mix sites within folds, preventing evaluation of cross-site transfer, which is the central claim of this work. We compare the hierarchical model (HM) with a single-site XGBoost baseline (XGB) and a naive predictor (train-mean). All results use the held-out test set only. The models are evaluated using CV-MAE, MAE, and R^2.

Asian cross-validation. Table 2 reports 5-fold time-series cross-validation on the Asian data. Context-aware features achieve CV-MAE of 0.85–1.30%, compared with 1.57–1.94% when using only the four raw monthly statistics (mean temperature, humidity, CO_2, TVOC). This consistent improvement across all age groups indicates that temporal patterns provide useful predictive information beyond simple averages.

Northern European test results. Table 3 and Figure 2 report test-set metrics. HM achieves MAE = 4.30% and $R^2 = 0.86$ for the 35–39 group, a 64% improvement over naive. This group is the most reliable target because it has the largest per-month patient count at both clinics,

Table 2. Asian clinic — 5-fold time-series CV (61 weeks). Context-aware vs. raw-feature baseline (in parentheses).

Age group	Model	CV-MAE	(Raw)
PR <35 yrs	RF	1.25%	(1.94%)
PR 35–39 yrs	RF	0.85%	(1.57%)
PR 40+ yrs	RF	1.30%	(1.83%)

making the observed rate a stable estimator. We caution that with only 3 test months the reported point estimates carry substantial uncertainty; these results should be interpreted as a promising preliminary signal rather than a definitive claim. The strongly negative R^2 values for XGBoost (-5.09, -0.28, -0.33) indicate that a point-estimate model without regularisation overfits the 11 training months and collapses on the test set as a direct consequence of the limited sample size. This pathological behaviour is precisely what motivates the hierarchical Bayesian approach: the informative prior and partial pooling act as regularisation, preventing overfitting in the low-data regime. For the 40+ group HM recovers $R^2 = +0.08$ and a 17% improvement; the hierarchical prior stabilises predictions against counting noise where the single-site baseline ($+1\%$) cannot. The under-35 group is noise-dominated: monthly rates reach exactly 0% or 100% when only one or two patients are treated, and both models fall below naive, confirming that no environmental signal can be recovered at this level of aggregation.

Feature importance results from SHAP and LIME are presented in Appendix B; key findings are summarised below. Temperature and CO_2 emerge as the most consistently important variables across both clinics, while TVOC shows a stronger negative effect in Northern Europe. Site-dependent differences in CO_2 direction are discussed in Appendix B.

Limitations. The dataset is limited in size, particularly for the Northern European clinic (3 test months), which restricts statistical power and means reported metrics carry substantial uncertainty. Outcomes are aggregated proportions without patient-level information; factors such as embryo quality, stimulation protocol, and patient history are unobserved and may confound the estimated environmental effects. Similarly, clinic-month-level outcomes may correlate with staffing patterns, lab protocol changes, or seasonal effects unrelated to sensor readings.

Table 3. Northern European clinic — test-set results (3 held-out months). ΔN = improvement over naive. **Bold** = best MAE per group.

Age group	Model	MAE	R^2	ΔN
PR <35 (noise-limited)	Naive	8.76%	—	—
	XGB	17.05%	-5.09	-95%
	HM	**16.55%**	-8.81	-89%
PR 35–39 **(primary)**	Naive	11.80%	—	—
	XGB	11.80%	-0.28	0%
	HM	**4.30%**	$\mathbf{+0.86}$	$\mathbf{+64\%}$
PR 40+ (few patients)	Naive	20.14%	—	—
	XGB	19.85%	-0.33	$+1\%$
	HM	**16.71%**	$\mathbf{+0.08}$	$\mathbf{+17\%}$

5. Conclusion and Future Work

Context-aware features capture patterns that are not visible in simple monthly averages, and this is supported by results on the Asian data and the transfer to Northern Europe. The hierarchical model with partial pooling reaches $R^2 = 0.86$ for the 35–39 age group, where CO_2 and temperature appear consistently important across both clinics. The 40+ group shows only small improvements, and the under-35 group is mostly dominated by noise at this level of aggregation. Overall, the results suggest that routine environmental monitoring is critical to contain useful information about laboratory conditions that goes beyond simple threshold alarms. Future work should focus on larger datasets and adding patient-level and incubator-level data to get more reliable and biologically meaningful results. In addition, counterfactual analysis could be applied in this context to explore "what-if" scenarios, enabling the assessment of how changes in environmental factors (e.g., CO_2 levels or temperature) might influence success rates. Such approaches could provide actionable insights for optimizing laboratory conditions and improving clinical outcomes.

Acknowledgements

This work was supported by Vinnova, the Swedish Governmental Agency for Innovation Systems [Grant No. 2024-01462]. The funding source had no role in the design, execution, or publication of the study.

References

Younglai, E. V., Holloway, A. C., and Foster, W. G. Environmental and occupational factors affecting fertility and IVF success. *Human Reproduction Update*, 11(1):43–57, 2005.

Mantikou, E., Youssef, M. A. F. M., van Wely, M., van der Veen, F., Al-Inany, H. G., Repping, S., and Mastenbroek, S. Embryo culture media and IVF/ICSI success rates: a systematic review. *Human Reproduction Update*, 19(3):210–220, 2013.

Varzaneh, Z. A., Wölner-Hanssen, N., and Khoshkangini, R. A Lightweight Transformer Approach for Predicting Blastocyst Formation on Limited Embryo Images. In *Proceedings of the 2025 International Conference on Visual*

Communications and Image Processing (VCIP), pp. 1–5, 2025.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

Bhuwalka, K., Choi, E., Moore, E. A., Roth, R., Kirchain, R. E., and Olivetti, E. A. A hierarchical Bayesian regression model that reduces uncertainty in material demand predictions. *Journal of Industrial Ecology*, 27(1):43–55, 2023.

Hoffman, M. D. and Gelman, A. The No-U-Turn Sampler. *JMLR*, 15:1593–1623, 2014.

Kupka, M. S. et al. Assisted reproduction in Europe, 2011. *Human Reproduction*, 31(2):233–248, 2016.

Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., and Menegaz, G. A perspective on explainable artificial intelligence methods: SHAP and LIME. *Advanced Intelligent Systems*, 7(1):2400304, 2025.

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55, 2016.

Swain, J. E. et al. Optimizing the culture environment and embryo manipulation. *Fertility and Sterility*, 103(3):571–579, 2012.

A. Prior Specification and Beta-Binomial Extension

Choice of prior means μ_0. The global mean μ_G in Eq. (3) is centred on an approximate prior pregnancy rate μ_0 for each age group. These values are not directly tabulated in any single source; they are weakly informative approximations consistent with the ranges reported across European ART registries (clinical pregnancy rates per fresh autologous cycle). Because σ_G is inferred from data and both Asia and Northern Europe contribute many observations, the posterior is not sensitive to moderate changes in μ_0.

Table 4. Weakly informative prior means μ_0 used for μ_G, consistent with European ART registry ranges (Kupka et al., 2016). These are approximate starting points; the posterior updates them substantially.

Age group	μ_0	σ_0	Rationale
$<$35 years	0.40	0.10	Approx. clinical PR per cycle, European avg
35–39 years	0.28	0.10	Lower success, wider uncertainty
$\geq$40 years	0.12	0.08	Marked decline; narrow prior, broad posterior

Beta-Binomial extension. When monthly patient counts $n_{t,c}$ are available, replace Eq. (1) with $k_{t,c} \sim \mathrm{BetaBinomial}(\mu_{t,c}\kappa, (1-\mu_{t,c})\kappa, n_{t,c})$, where $k_{t,c}$ is the number of clinical pregnancies and $\kappa \sim \mathrm{Gamma}(2, 0.1)$. This weights each month by how many patients contributed—the statistically correct treatment for small-count subgroups, and the most important modelling improvement available for future work.

B. SHAP and LIME Interpretation

Figures 3 and 4 provide two complementary views of how environmental conditions relate to pregnancy rates (Salih et al., 2025). The SHAP analysis (Figure 3) shows which variables matter most and in which direction across both clinics. In the Asian data, higher CO_2 values are associated with positive SHAP contributions. In contrast, in the Northern European clinic, higher TVOC shows a clear negative effect, and both CO_2 and temperature tend to contribute negatively. The opposite directional effect of CO_2 between sites is consistent with the model structure: $\boldsymbol{\beta}$ is shared, but the posterior is weighted toward the Asian clinic, which provides far more observations. This site-dependency in CO_2 effects likely reflects genuine differences in incubator types, ventilation systems, or CO_2 calibration protocols between the two laboratories, and warrants investigation with larger datasets. Temperature shows a more consistent pattern across sites. LIME (Figure 4) provides a local view by explaining individual predictions. Averaged over Northern European months, lower CO_2 levels consistently contribute positively, while higher temperature and TVOC reduce predicted pregnancy rates. Higher humidity shows a positive contribution. LIME highlights how these effects combine within specific observations, complementing the global view from SHAP.

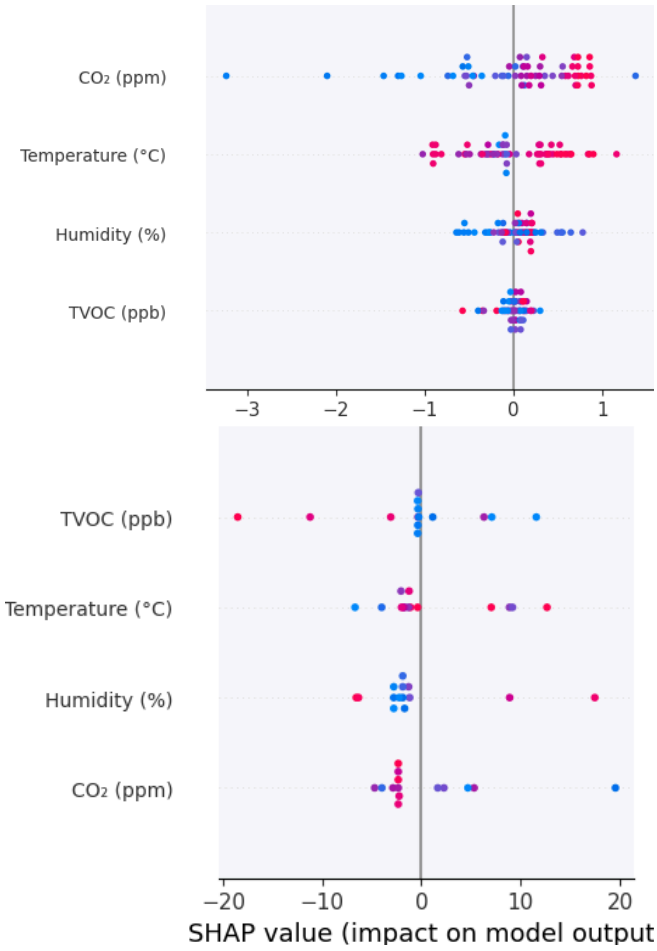

Figure 3. SHAP beeswarm for PR 35–39. *Top*: Asia (61 weeks). *Bottom*: Northern Europe (14 months). Red = high feature value; blue = low.

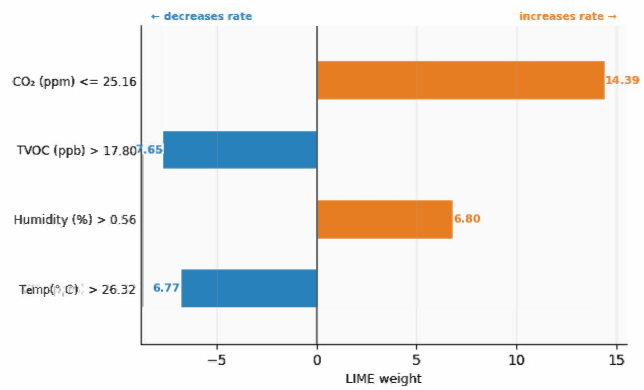

Figure 4. LIME explanation for December 2025, PR 35–39 (Northern European clinic). Orange bars increase the predicted pregnancy rate; blue bars decrease it.