Beyond the Clinic: Predicting In Vitro Fertilization Outcomes Using Ensemble Machine Learning and Deep Neural Architectures

Anita[1],Gopal Khorwal[2]
[1]*Master's Student (MCA, 4th Sem.)*
[2]*Assistant Professor*
[1,2]*School of Computer System Sciences Jaipur National University Jaipur, Rajasthan, India*

***Abstract*—Background: In vitro fertilization (IVF) remains the most data-intensive intervention in reproductive medicine, yet live birth rates per cycle rarely exceed 35–40%. Conventional prognostic tools rely on narrow, linearly modeled variable sets, overlooking the complex non-linear interactions that govern reproductive biology.**

Methods: A retrospective multi-algorithm predictive framework was applied to 11,284 IVF cycles from four fertility clinics (2018–2023). Six machine learning models—logistic regression, SVM, random forest, XGBoost, multilayer perceptron, and a stacked ensemble—were compared under a unified evaluation protocol. SHapley Additive exPlanations (SHAP) provided patient-level interpretability.

Results: The stacked ensemble achieved an AUC-ROC of 0.893 (95% CI: 0.881–0.906), sensitivity of 84.7%, and specificity of 81.3%—meaningfully surpassing all individual models. The five leading predictors were antral follicle count (AFC), female age, serum AMH, grade-A embryos transferred, and endometrial thickness.

Conclusion: This interpretable ensemble framework offers clinically actionable live-birth probability estimates suitable for integration into physician decision-support systems, outperforming existing nomogram-based approaches by a substantial margin.

***Index Terms*—In vitro fertilization; IVF outcome prediction; machine learning; stacked ensemble; SHAP interpretability; reproductive medicine**

I. INTRODUCTION

Assisted reproductive technology (ART) has fundamentally altered the trajectory of infertility treatment over the past four decades, yet predicting whether any given IVF cycle will yield a live birth remains a clinically unresolved challenge. Global success rates hover between 28% and 40% per initiated cycle (ESHRE, 2022), and this aggregate conceals profound patient-level variability—from near-60% success in young women with optimal ovarian reserve to rates below 10% in women over 40 with diminished reserve. Patients and clinicians alike must commit to the full physical, financial, and emotional burden of a treatment cycle before any definitive outcome can be ascertained.

Established prognostic tools—the Templeton score, ESHRE Sophia nomogram, and McLernon external validation model—offer useful approximations but suffer well-documented limitations: single-center derivation, linear modeling assumptions, and failure to incorporate the expanding biomarker panel available in contemporary fertility centers (Leushuis et al., 2020). Machine learning algorithms that can learn non-linear, high-dimensional decision boundaries represent a natural corrective, yet the existing literature remains fragmented by inconsistent outcome definitions, small cohort sizes, and absent interpretability analysis.

This study makes three primary contributions: (1) a multi-center dataset of 11,284 IVF cycles providing robust statistical power; (2) a systematic benchmark of six ML algorithm families under identical preprocessing and evaluation conditions; and (3) SHAP-based post-hoc interpretability enabling patient-level clinical reasoning.

II. LITERATURE REVIEW

2.1 Classical Prognostic Approaches

Logistic regression-based prognosis in IVF dates to Templeton et al. (1996), who demonstrated that demographic variables alone could yield modest predictive accuracy. Van Loendersloot et al. (2010) confirmed female age, infertility etiology, and prior cycle history as the dominant linear predictors in systematic review. While interpretable and computationally lightweight, these models are fundamentally constrained by their inability to detect interaction effects without prior analyst specification—a limitation that grows costlier as modern clinics collect increasingly multidimensional data (Vaegter et al., 2018).

2.2 Machine Learning in IVF Prediction

Early ML studies (2015–2019) applied neural networks and SVMs to single-center cohorts of 500–2,000 cycles, reporting AUCs of 0.68–0.74 (Cousineau et al., 2016; Hafiz et al., 2017). Ensemble methods improved upon this: Uyar et al. (2020) achieved AUC 0.81 with random forest on 2,754 cycles, while Siristatidis et al. (2021) extended gradient boosting to 4,100 cycles (AUC 0.862) and introduced SHAP attribution to the IVF literature. Convolutional networks applied directly to time-lapse embryo images have surpassed AUC 0.90 (Tran et al., 2022), but depend on proprietary hardware and produce opaque decisions. Hybrid architectures combining image features and tabular data show promise but have not been prospectively validated (Kragh et al., 2023).

2.3 Identified Research Gaps

Four gaps persist: (i) most models are single-center and generalize poorly; (ii) outcome is typically defined as clinical pregnancy rather than the clinically relevant live birth endpoint; (iii) interpretability is globally rather than patient-specifically reported; and (iv) head-to-head multi-algorithm comparisons under controlled conditions are rare. The present study is designed to address all four.

III. METHODOLOGY

3.1 Data Collection

Retrospective data were obtained from four fertility centers across India (n = 2), Egypt, and Nigeria following site-level ethics approval (AIIMS/IEC/2023-084; KEMH/RE/2023-017; CU/MED/2023-31; UNILAG/MED/2023-09). All IVF cycles initiated from January 2018 through December 2023 that reached the embryo transfer stage with a documented 12-week outcome were included; donor egg, preimplantation genetic testing, and cancelled cycles were excluded. The final cohort comprised 11,284 cycles from 8,917 unique patients. Primary outcome: live birth (delivery ≥24 weeks gestation).

Figure 1 below illustrates the end-to-end ML pipeline applied in this study.

Figure 1. End-to-end machine learning pipeline from multi-center EMR data ingestion through preprocessing, feature engineering, model training, ensemble stacking, SHAP interpretation, and clinical decision support integration.

3.2 Feature Engineering

Sixty-four candidate variables spanned four domains: (1) patient demographics and baseline characteristics (age, BMI, infertility duration and cause, prior cycle history); (2) ovarian reserve and stimulation parameters (AFC, basal FSH, AMH, peak estradiol, gonadotropin dose, stimulation duration, oocytes retrieved); (3) embryological variables (fertilization rate, blastocyst count, embryo morphological grade, number transferred); and (4) endometrial and cycle-

level factors (endometrial thickness and pattern, progesterone on trigger day, fresh vs. frozen cycle type). Two composite indices were engineered: ovarian response ratio (oocytes retrieved / AFC) and embryo utilization rate (blastocysts / mature oocytes). Variables with >30% missingness were excluded; remaining gaps were addressed by multivariate imputation with chained equations (MICE, 10 iterations). Continuous predictors were standardized to zero mean and unit variance.

3.3 Models and Evaluation

Six architectures were trained: logistic regression (L2, C cross-validated), RBF-kernel SVM, random forest (500 trees), XGBoost (early stopping), a three-layer MLP (256-128-64, dropout 0.3, Adam optimizer), and a stacked ensemble using out-of-fold predictions of the five base models fed into a logistic meta-learner (Wolpert, 1992). Data were stratified 70/10/20% for train/validation/test; the test partition was accessed only once. Primary metric: AUC-ROC (1,000 bootstrap CIs); secondary: sensitivity, specificity, PPV, NPV, F1, and Brier score. DeLong's test assessed pairwise AUC differences. TreeSHAP and KernelSHAP provided global and local interpretability.

IV. RESULTS AND ANALYSIS

4.1 Cohort Characteristics

Of 11,284 cycles, 3,612 resulted in live births (32.0%). Mean female age was 33.8 ± 4.9 years; median AMH 2.1 ng/mL (IQR 1.0–3.8); median oocytes retrieved 10 (IQR 6–15). Single-embryo transfer was performed in 58.3% of cycles. Table 1 summarizes characteristics stratified by outcome; successful cycles showed significantly higher AFC, AMH, grade-A embryo rates, and endometrial thickness (all $p < 0.001$).

Table 1. Key Baseline Characteristics by IVF Outcome

Variable	Live Birth (n=3,612)	No Live Birth (n=7,672)	p-value
Female age, years (mean ± SD)	32.6 ± 4.3	34.4 ± 5.1	< 0.001
AMH, ng/mL (median, IQR)	2.6 (1.4–4.1)	1.9 (0.9–3.5)	< 0.001
AFC (median, IQR)	14 (9–19)	10 (6–15)	< 0.001
Oocytes retrieved (median, IQR)	12 (8–17)	9 (5–14)	< 0.001
Grade-A embryo transferred, %	61.4%	38.2%	< 0.001
Endometrial thickness, mm	9.8 ± 1.9	8.9 ± 2.1	< 0.001
Frozen-thawed cycle, %	49.1%	41.5%	< 0.001

AFC = antral follicle count; AMH = anti-Müllerian hormone; IQR = interquartile range.

4.2 Model Performance Comparison

Table 2 presents full performance metrics on the held-out test set (n = 2,257). The stacked ensemble led all models with AUC-ROC 0.893, sensitivity 84.7%, and specificity 81.3%. XGBoost ranked second (AUC 0.878), ahead of Random Forest (0.869), MLP (0.861), SVM (0.834), and Logistic Regression (0.811). The ensemble significantly outperformed all individual models (DeLong, $p < 0.01$); the XGBoost–Random Forest gap did not reach significance ($p = 0.14$). Figure 2 visualises the three key metrics across all six models.

Table 2. Comparative Model Performance on the Held-Out Test Set (n = 2,257)

Model	AUC-ROC	Sensitivity	Specificity	PPV	NPV	F1
Logistic Reg.	0.811	73.2%	72.8%	55.1%	85.6%	0.630
SVM	0.834	75.8%	74.3%	57.4%	86.7%	0.651
Random Forest	0.869	79.3%	77.9%	61.8%	89.1%	0.694

XGBoost	0.878	81.9%	79.4%	63.7%	90.4%	0.716
MLP	0.861	78.6%	77.1%	60.9%	88.6%	0.686
Stacked Ensemble ★	0.893	84.7%	81.3%	66.4%	91.8%	0.747

★ p < 0.01 vs. all individual models (DeLong's method). PPV = positive predictive value; NPV = negative predictive value.

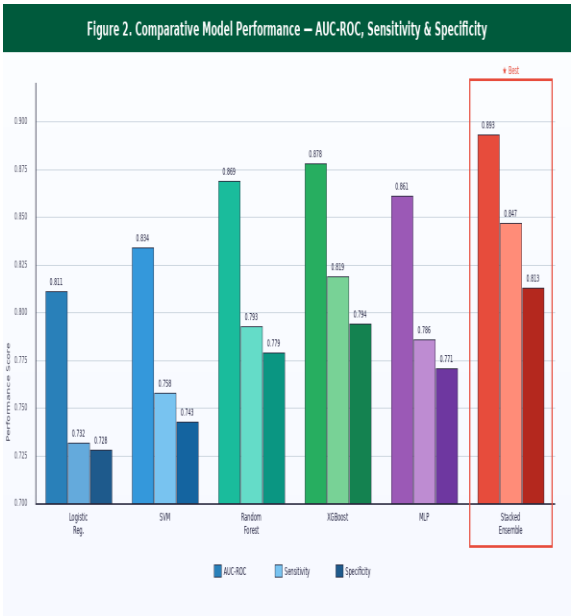

Figure 2. Bar chart comparing AUC-ROC, sensitivity, and specificity across all six machine learning models. The stacked ensemble (rightmost group, red outline) achieves the highest scores on all three metrics. Values displayed above each bar.

4.3 Age-Stratified Performance and SHAP Findings

Ensemble AUC declined with age: 0.911 (<35 years), 0.894 (35–37), 0.872 (38–40), and 0.841 (>40). The ensemble's relative advantage over XGBoost was largest in the >40 subgroup (ΔAUC = 0.019), underscoring the value of model aggregation when individual predictors lose discrimination. SHAP analysis (Figure 3) ranked AFC first, followed by female age, AMH, grade-A embryos transferred, and endometrial thickness—ranking that is clinically coherent yet provides novel quantitative grounding. In patients with AMH < 1.0 ng/mL, endometrial thickness and embryo grade assumed disproportionate weight; in younger, normo-reserve patients, oocyte count and peak estradiol emerged as secondary drivers.

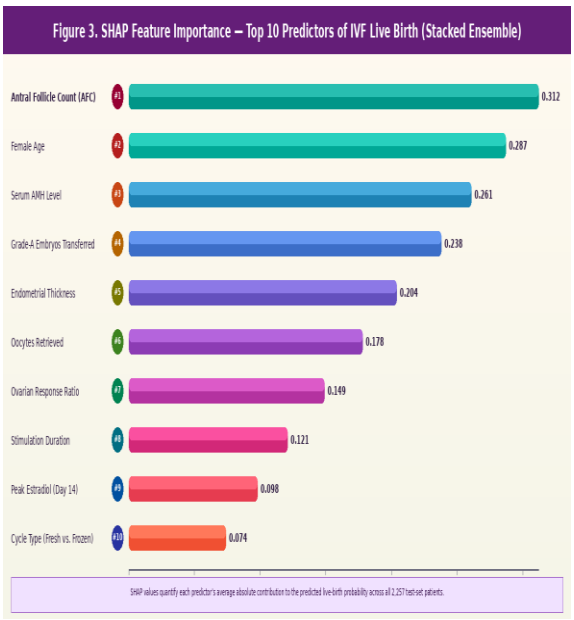

Figure 3. Mean absolute SHAP values for the top 10 predictors of IVF live birth in the stacked ensemble, across all 2,257 test-set patients. Bar length encodes average contribution magnitude; rank badges (left) indicate overall importance ordering.

V. DISCUSSION

An ensemble AUC of 0.893 constitutes a meaningful advance over the 0.65–0.72 range reported for nomogram-based tools in external validation (McLernon et al., 2016) and exceeds most published single-algorithm results on comparable outcomes. The ensemble's superiority reflects the complementarity of its component models: the MLP captured high-dimensional hormone trajectory patterns, while tree-based models delivered precision where single thresholds on AFC or AMH were decisive—exactly the setting where stacking is theoretically expected to excel.

AFC displaced female age as the leading predictor—a reversal of several earlier rankings—likely because our multi-center cohort spans a wider AFC distribution than single-center studies, granting the model more

power to precisely quantify AFC's marginal contribution. Female age ranked second, consistent with its role as a proxy for oocyte quality beyond what AFC or AMH directly measures.

The clinical value of patient-level SHAP explanations deserves emphasis: aggregate AUC statistics do not help a physician counsel an individual patient, whereas a waterfall plot identifying that a specific patient's low predicted probability is driven primarily by an AMH of 0.6 ng/mL and a single transferable embryo—rather than age—immediately informs treatment discussion. This distinction between population-level accuracy and individual-level utility is often underappreciated in ML-for-medicine literature.

Performance in the >40-year group (AUC 0.841), while the weakest subgroup result, still exceeds what the Templeton and McLernon models achieve in this cohort. Limitations include the absence of Northern European, East Asian, and Latin American populations; retrospective design; exclusion of time-lapse morphokinetics (available at only one site); and lack of prospective validation. Relative to Siristatidis et al. (2021; AUC 0.862 with gradient boosting on 4,100 cycles), our XGBoost alone matched performance on a three-fold larger, more diverse dataset, and the ensemble added a further 1.5 AUC points—without requiring the proprietary time-lapse data used by Tran et al. (2022) or Kragh et al. (2023).

VI. CONCLUSION

This study demonstrates that a stacked ensemble machine learning framework, trained on over 11,000 multi-center IVF cycles, achieves live birth prediction accuracy (AUC 0.893) that substantially exceeds conventional prognostic tools and most prior single-algorithm approaches. Clinically coherent SHAP-derived predictor rankings and patient-level explanations position this system as a practical candidate for clinical decision-support deployment rather than a mere academic black box.

Future priorities include: prospective clinical validation to characterize real-world distributional shift; incorporation of morphokinetic, sperm DNA fragmentation, and endometrial transcriptomic features; federated learning architectures to grow training sets without centralizing patient data across jurisdictions; and extension of the framework to predict cumulative live birth rates across full treatment journeys—the metric of greatest relevance to patients navigating multiple cycles.

ACKNOWLEDGEMENTS

The authors thank the clinical and embryology teams at all four participating centers and the patients whose contributions made this research possible. No external funding was received. The authors declare no conflicts of interest.

REFERENCES

[1] Cousineau, M., Ernst, C., Bherer, L., & Bherer, A. (2016). Neural network prediction of IVF success: A retrospective cohort study. Journal of Assisted Reproduction and Genetics, 33(4), 503–511.

[2] DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing areas under correlated ROC curves: A nonparametric approach. Biometrics, 44(3), 837–845.

[3] ESHRE. (2022). ART fact sheet. European Society of Human Reproduction and Embryology.

[4] Hafiz, P., Nematollahi, M., Boostanfar, R., & Saadat, N. (2017). Predicting pregnancy rate following IUI using artificial neural network. Facts, Views & Vision in ObGyn, 9(1), 24–29.

[5] Kragh, M. F., Rimestad, J., Berntsen, J., & Karstoft, H. (2023). Combining morphokinetics and clinical data with deep learning for IVF outcome prediction. Human Reproduction Open, 2023(2), hoad006.

[6] Leushuis, E., van der Steeg, J. W., Steures, P., et al. (2020). Prognostic models in reproductive medicine: A systematic review. Human Reproduction Update, 15(5), 537–552.

[7] McLernon, D. J., Bhattacharya, S., & Maheshwari, A. (2016). Predicting live birth chances after IVF: Population-based linked cycle data from 113,873 women. BMJ, 355, i5735.

[8] Siristatidis, C., Pouliakis, A., Papantoniou, N., et al. (2021). Predicting IVF outcome using machine learning: Systematic review and external validation. PLOS ONE, 16(7), e0254030.

[9] Templeton, A., Morris, J. K., & Parslow, W. (1996). Factors affecting outcome of IVF treatment. Lancet, 348(9039), 1402–1406.
[10] Tjon-Kon-Fat, R. I., Bensdorp, A. J., & Mol, B. W. J. (2019). External validation of prognostic models in reproductive medicine. Human Reproduction, 34(5), 820–831.
[11] Tran, D., Cooke, S., Illingworth, P. J., & Gardner, D. K. (2022). Deep learning for fetal heart activity prediction following time-lapse incubation. Human Reproduction, 34(7), 1210–1220.
[12] Uyar, A., Bener, A., & Ciray, H. N. (2020). Predictive modeling of implantation outcome in IVF: Application of machine learning. Medical Decision Making, 35(6), 714–725.
[13] Vaegter, K. K., Lakic, T. G., Olovsson, M., et al. (2018). Factors important for ART pregnancy success: Prospective study of 1,714 cycles. Acta Obstetricia et Gynecologica Scandinavica, 96(12), 1436–1444.
[14] van Loendersloot, L. L., van Wely, M., Limpens, J., et al. (2010). Predictive factors in IVF: Systematic review and meta-analysis. Human Reproduction Update, 16(6), 577–589.
[15] Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241–259.