PriMera Scientific Engineering
Volume 4 Issue 2 February 2024
ISSN: 2834-2550

PriMera Scientific Publications

Rethinking Challenges of Machine Learning in Assisted Reproductive Technology

Type: Research Article
Received: December 30, 2023
Published: January 22, 2024

Citation:
Chenwei Wu. "Rethinking Challenges of Machine Learning in Assisted Reproductive Technology". PriMera Scientific Engineering 4.2 (2024): 15-27.

Chenwei Wu*

Electrical and Computer Engineering, University of Michigan, United States

***Corresponding Author:** Chenwei Wu, University of Michigan, 3060 Whisperwood Dr 407, Ann Arbor, Michigan, United States.

Abstract

Predicting pregnancy and live births using machine learning in the field of in-vitro fertilization (IVF) has long posed a significant challenge due to the difficulty in achieving consistent performance across various studies. In this paper, we conduct a comprehensive review and analysis of the existing limitations in current research. Additionally, we introduce a standardized machine learning pipeline, which serves as a valuable guide for future researchers. Furthermore, we propose two alternative modeling approaches: phase-by-phase modeling and subgroup FMLR modeling. These two alternatives not only enhance prediction performance but also offer clinically sensible explanations and timely guidance for users. Most notably, they shed light on the complexities of the IVF cycle, highlighting when, who, and where machine learning tasks face their greatest challenges. This insight can inspire future efforts in data collection and patient engagement processes.

Keywords: In-vitro fertilization; Machine Learning; Explainable AI; Medical AI

Abbreviations

In-vitro fertilization (IVF), Machine Learning (ML), Artificial Intelligence (AI).

Introduction

In recent years, medical practitioners have shown increased interest in utilizing machine learning models for predicting pregnancy and live births, particularly in identifying the significant factors affecting the outcome of in vitro fertilization (IVF) [10]. While previous research provides a comprehensive overview of applying machine learning techniques in the context of IVF, little attention has been given to analyzing the reproducibility and shortcomings of these studies. Although many models have undergone testing and some have reported decent performance in testing, their real-world adoption for prognosis and diagnosis remains limited due to a lack of external validation [8]. These models often fail when applied to real-life populations, and a domain shift can substantially compromise their generalizability and reproducibility. Therefore, there is a pressing need for further evaluation and analysis to standardize the machine learning workflow in future research [11].

Previous work on IVF pregnancy and live birth prediction using machine learning models has yielded inconsistent results, with AUC values ranging from 0.6 to over 0.95. In this study, we examined the most standard techniques used in previous work, including logistic regression, decision trees, XGBoost, and SVM [6], using our private dataset. However, our evaluation also revealed inconsistent AUC results. As a result, we embarked on a deeper examination of why IVF prediction is such a challenging task and propose two novel approaches, 1) stage-by-stage modeling and 2) FMLR subgroup modeling, to enhance model performance and real-life generalizability.

The IVF treatment process is a sequential and time-consuming one with numerous sub-cycles. Throughout this process, many patient and treatment factors undergo changes that can impact the evaluation of potential pregnancy and live births. Two primary issues arise when applying machine learning models to the IVF treatment process. First, there is a time discrepancy between feature collection and model usage. Some models require a large number of features, but many of these features may not be available until various medical tests are conducted, whereas the models are intended for real-time use. In other words, we may not collect enough features when we apply the models. The second issue is data leakage, which occurs when using features that result from the clinician's prediction of the treatment outcome. This data leakage issue is prevalent in both machine learning and IVF literature. Therefore, the practical benefit of having a machine learning prediction of pregnancy or live birth only a few days before the real-life outcome is questionable. Models that generate predictions of live births or pregnancies solely based on preliminary results and patient demographics may be too arbitrary and deterministic for the dynamic nature of the treatment progress, potentially giving patients false hope or early discouragement. What clinicians truly need is likely a stage-by-stage approach to better outline the treatment progress, while patients may benefit more from machine learning models that provide shorter-term insights.

IVF data also exhibits significant heterogeneity due to the intricate interplay of various biological, environmental, and procedural factors, leading to the presence of numerous latent subgroups within the data. Traditional machine learning approaches, while effective in capturing general patterns, often struggle to account for the nuanced and individualized nature of IVF outcomes. A more tailored approach is the use of mixture regression models, which inherently recognize and adapt to the heterogeneity within the data. We propose a subgroup modeling workflow, that function by identifying latent subgroups and applying distinct regression models to each, thereby capturing the unique characteristics and relationships within each subgroup. This approach not only enhances predictive accuracy but also provides deeper insights into the complex dynamics of IVF treatments.

We structure our paper into three main parts:

1. A comprehensive review of past ML+IVF literature and analysis of the flaws.
2. Experiments of a standard ML pipeline on our dataset and why it fails.
3. Two novel alternatives: the Subgroup modeling approach and the Phase-by-phase modeling approach.

Materials and Methods

Materials I. Literature Analysis: Pitfalls of Applied Machine Learning in IVF

In our preliminary investigation, we conducted an extensive literature review spanning from 1997 to 2021. This review aimed to assess the performance of various machine learning models in IVF prediction papers and understand the differences in their approaches, datasets, and their resulting performance variations, as indicated by AUC (Area under the ROC Curve) values. Notably, the AUCs reported in the reviewed literature exhibited a broad spectrum, ranging from 0.6 to 0.95. We identified that these discrepancies in AUC values are primarily attributed to the variability in the availability of crucial features and the changing demographics of IVF patients, rather than divergent modeling techniques.

Several factors contribute to the inherent challenge of accurately predicting pregnancy or live birth in IVF procedures. Firstly, the inability to replicate findings from previous studies arises due to differences in data sources, volumes, and study sizes. These studies did not consistently employ a left-out test set for evaluation.

Secondly, upon reviewing variables used in previous studies, we identified that important features such as the BMI of males, age of the male partner, alcohol and smoking habits of both parents, are not always present across datasets. Inclusion of these variables may enhance prediction accuracy.

Thirdly, the demographics of patients undergoing IVF treatments have evolved over the decades. For instance, during the 1990s, Caucasians accounted for the majority of patients, constituting 91.5% of the patient population from 1994 to 1998. African Americans, Asians, and Hispanics represented 4%, 3%, and 1.5% of the population, respectively. IVF treatments have since become accessible to a more diverse demographic group. Although our dataset lacks race and ethnicity information, it is reasonable to assume that today's patients receiving treatments differ significantly from those included in data collected in the 1990s.

Our examination also revealed common flaws in existing ML+IVF literature. These shortcomings can be categorized into two main groups: the exclusion of subpopulations and data leakage.

The exclusion of specific subpopulations emerged as a recurring issue in current research. For example, in [5], IVF/ICSI cycles were excluded in cases involving oocyte or embryo donation, surgically retrieved spermatozoa, patients positive for human immunodeficiency virus, modified natural IVF, and cycles canceled due to poor ovarian stimulation, ovarian hyperstimulation syndrome, or other unforeseen medical or non-medical reasons. Similarly [4], excluded cases involving embryos related to donor/surrogate mothers and cycles using frozen embryos, which constitute a significant portion of common IVF data points. These exclusions artificially increase the variance in patient populations and impair the model's generalization capabilities.

Another common mistake in modeling IVF data is the use of information that would not be available to the model at the time it needs to make predictions. Machine learning models used in clinical settings often operate in real time, and they lack access to every feature available in the dataset at the moment of prediction. Tests and demographic features exist in the IVF dataset but may not yield results until pregnancy or shortly before pregnancy, such as B-HCG. Another form of this error, known as data leakage, involves using features whose values are obtained as a result of the clinician's prediction of the patient's outcome. This use of data-leaking features divulges information about the true label the model aims to predict, even though, in practice, the model should assist the clinician in making their prediction initially, without having access to these data-leaking features. For instance, in one paper with exceptionally high performance [10], authors used B-HCG as a predictor. When we incorporated B-HCG as a predictor in our model, we achieved an AUROC as high as over 95 percent. However, HCG is a hormone produced in the body during pregnancy, rendering such predictions meaningless. These two identified flaws inspired us to devise two alternatives to the traditional ML workflow: phase-by-phase modeling and subgroup modeling, which will be explored in subsequent sections.

A summary of the literature review could be found below:

#	*Data Year and Size*	*Model & Performance*	*Important Features*	*Used Features*	*Outcome Variable*	*External Validation*	*Limitations*
[1]	1991-1994, 36961	Logistic Regression	Age	Age, Treatment	Livebirth Success	N	Cycles that involved gamete or embryo donation, frozen embryo transfer, or micromanipulation and unstimulated cycles were excluded.

[2]	Year range unknown, 554	Logistic Regression (no interaction terms)	Maternal age (negative), Number and quality of embryos (positive)	Maternal age, Cause for intervention, Donor insemination, Rank of attempt, Serum LH and E2 levels on day of hCG administration, Embryo transfer catheter (flexible vs rigid), Number of embryos transferred of each morphologic type and developmental stage, Sperm parameters (concentration, percentage motility and rate of progression) before and after Percoll processing, Sperm concentration at insemination, Number and quality of retrieved oocytes, Human factor	pregnancies, live births, and multiple birth deliveries vs. IVF failure	N	Small Sample Size
[3]	1993-1998, 642 women undergoing their first IVF treatment cycle in which no more than two embryos were transferred.	Multivariate logistic regression (AUC 0.68)	Development stage, Morphology score of the 2 best embryos, Age	Woman's age (per y), Duration of infertility (per y), Secondary type of infertility. Indication for IVF: Tubal, Male factor, Idiopathic infertility, Others, Total no. of sperm cells (per 107/mL), Progressive motile sperm cells (per %), Estrogen level (per 103 pmol/L), No. of preovulatory follicles (per follicle), No. of retrieved oocytes (per oocyte), Proportion of oocytes fertilized (per 10%). Day of ET: Day 3, Day 4, Day 5, No. of embryos suitable for transfer (per embryo). Stage development of the best embryo: Retarded, Appropriate, Advanced. Stage development of the second best embryo: Retarded, Appropriate, Advanced, Morphology score of the best embryo (range 1–4), Morphology score of the second best embryo (range 1–4)	singleton and twin pregnancy	N	No external validation
[4]	2003-2007, 144018	Logistic Regression (AUC 0.6335)	previous IVF live birth		preterm birth, low birth weight, and macrosomia	N	Excluded donor/surrogate mother, frozen embryos

[5]	2001- 2009, 2621	Logistic Regression (AUC 0.68)		Female age, duration of subfertility, previous ongoing pregnancy, male subfertility, diminished ovarian reserve, endometriosis, basal FSH, number of failed IVF cycles	ongoing pregnancy	N	IVF/ICSI cycles were excluded in the case of oocyte or embryo donation, surgically retrieved spermatozoa, patients positive for human immunodeficiency virus, modified natural IVF and cycles cancelled owing to poor ovarian stimulation, ovarian hyperstimulation syndrome or other unexpected medical or non-medical reasons.
[6]	2014- 2018, 7188	Logistic regression, Random forest, XGBoost, SVM (AUC 0.71, 0.73, 0.73, 0.71)		Age, AMH, BMI, duration of infertility, previous live birth, previous miscarriage, previous abortion and type of infertility	the live birth chance prior to the first IVF treatment	N	Limited generalization of the model to other populations. Model could only be used for couples who have never accepted IVF treatment, limited application. Failed to account for family genetic history and lifestyle factors
[7]	1999- 2008, 184269	Logistic regression, backwards selection (AUC 0.73)	Age	Pretreatment model: number of complete cycles, patient characteristics. Post-treatment model: number of complete cycles, patient characteristics, treatment information at first complete cycle	Cumulative chances of a first live birth	N	

[8]	2007- 2015, 526	multivariable logistic regression (AUC 0.62)	LH, male testosterone level, sperm motility	Type of infertility (primary/secondary); Duration of infertility (months); Female age (years); Parity (n); Average menstrual cycle length (days); Uterine abnormalities (yes/no); Antral follicle count before stimulation (number of follicles ¡11 mm); Alcohol use (self-reported; yes/no) for male and female; Smoking status (self reported; yes/no) for male and female; BMI at baseline (kg/m2) for male and female; Male age (years); Male testosterone (nmol/l); Male inhibin B (ng/l); Male FSH (IU/l); Male LH (IU/l); Total testicular volume (cc); Suspected primarily diagnosis of azoospermia (OA/NOA) before sperm retrieval. Number of TESE-ICSI cycles; Spermatozoa (fresh or frozen–thawed); Motility of spermatozoa (oocytes injected with motile spermatozoa/immotile spermatozoa or a combination of both for each individual cycle); Number of oocytes retrieved.	live birth in couples undergoing ICSI after successful testicular sperm extraction (TESE-ICSI)	Y	Paternal BMI as a predictor may help improve the model if the values are not missing a lot.
[9]	2012- 2016, 739	Binary regression with out interaction terms (AUC 0.688)	women's age, AFC, AMH; ovarian reserve measures	AMH, AFC, women's and men's age, body mass index (BMI) both for men and women, smoking status, previous diagnosis, type of treatment (IVF/ICSI), having had previous deliveries, ethnicity	Live birth in fresh cycle	N	Subgroups were created after a post hoc analysis of the data and this might be a source of bias.
[10]	April 2016 to February 2018, 500	KNN, SVM, Neural Networks, Naive Bayes, Random Forest, Decision Tree (AUC 0.87 - 0.97)	FSH/HMG dosage, contraception duration and the number of germinal vesicle (GV) quality oocytes	Clinical data, Female pathology data, Male pathology data, Embryological data, Semen analysis data	HCG	Y	Authors used BHCG as a predictor. However, HCG is a hormone produced in the body during pregnancy and thus should not appear as a predictor for pregnancy.

Table 1: A comprehensive review of machine learning for IVF literature.

Materials II. "Standardized" Machine Learning Pipeline: Why does it always fail to replicate?

Our machine learning analysis relies on our private IVF dataset, comprising approximately 6,000 data instances and around 2,000 unique patients. We embarked on replicating a "standard" machine learning workflow for predicting pregnancy and live births using our dataset. This replication encompassed the comprehensive methods found in the existing literature. Our approach could be delineated into three distinct phases: exploratory data analysis, data pre-processing, and model fitting. The primary goal was to assess the limitations of common machine learning workflows in the context of IVF prediction.

To gain a better understanding of the dataset, we conducted exploratory data analysis focusing on the distribution of variables, the total number of cycles, and the dropout rate. Initially, we examined the distribution of 15 predictor variables, including age, BMI, and other test results, across the 'pregnant' and 'non-pregnant' groups. The majority of patients underwent only one cycle, with the maximum number of cycles in this study being eight. There are in total 5196 cycles, and 4050 of those are cycle 1. The pregnancy rate for the first 3 cycles gradually decreases from 0.699 to 0.669 to 0.617, which is reasonable as the patients would continue the IVF treatment if they didn't get pregnant or achieve a live birth in the previous cycles. The pregnancy rate for the later cycles fluctuates from 0.813 in cycle 4, to 0.462 in cycle 5 to 0.5 in cycle 6. There were no successful pregnancies recorded for cycles beyond the sixth.

Having understood the trend between the number of cycles and pregnancy rates, we aimed to further analyze patient behavior to understand why certain patients opted for additional cycles after a failure, while others chose to discontinue treatment. Dropout rate for each cycle is defined as the percentage of people who didn't continue another cycle after not achieving a live birth. The dropout rate remains relatively stable for the first 3 cycles, 0.34, 0.357, and 0.317 respectively.

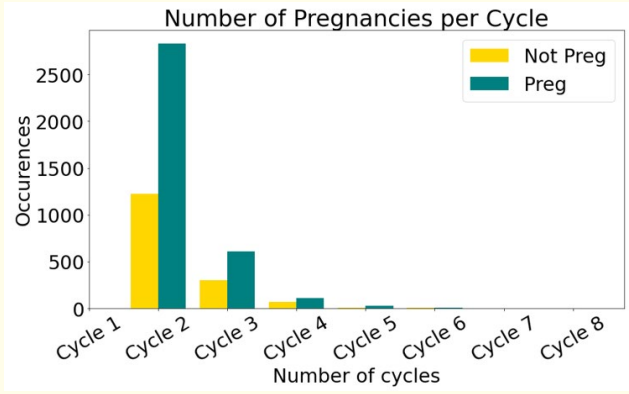

Figure 1: Number of pregnancies vs non-pregnancies per cycle.

During the data processing phase, as many features exhibit over 50% missingness, it may not be prudent to impute these variables. In medical settings, missingness can sometimes be a latent indicator of the patient's overall condition, rendering imputation unnecessary. Consider, for instance, the absence of a male fertility test and corresponding sperm quality values; this might suggest the use of high-quality donor sperm, with the clinician deeming a sperm quality test unnecessary. In this context, imputing data using either the mean sperm quality or a K-Nearest Neighbors (KNN) inferred value from similar patients' data could mislead the model. Therefore, we have introduced an indicator for the presence of a test in all data entries to assist the model in capturing this information.

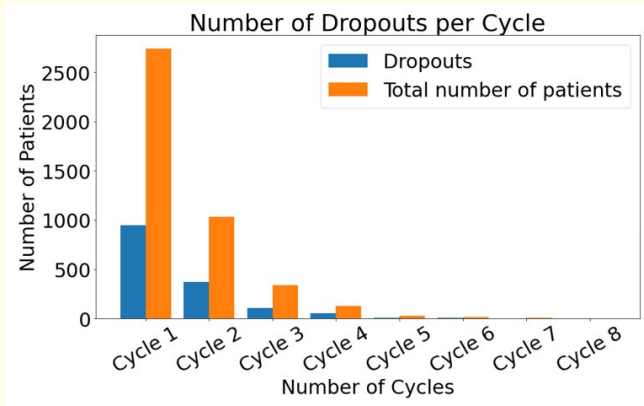

Figure 2: Number of dropouts per cycle.

The optimal results for predicting pregnancy, using Logistic Regression, Random Forest, SVM, Neural Network and XGBoost models, are approximately 0.68, and for predicting live birth, they are around 0.69. We tested and tuned hyperparameters employing a grid search over an extensive set of combinations. For Logistic Regression, our search encompassed the solver, norm of the penalty λ, and the magnitude of λ. For Random Forests and XGBoost, we investigated parameters such as the maximum depth of trees, gamma, regularization lambda, scale of positive class weight, subsample rate, and the number of trees, covering over 2000 combinations. Below is a performance chart depicting various combinations of methods employed:

Model	***AUC Pregnancy***	***AUC LiveBirth***
Logistic	0.64	0.65
Random Forest	0.65	0.67
SVM	0.65	0.63
Neural Network	0.67	0.69
XGBoost	0.68	0.69

Table 2: Standardized ML Pipeline Results.

Despite its performance agreeing with most of the previous work [6, 8, 9], a standardized ML pipeline is not able to replicate the stellar performance of up to 90 per cent AUC in [10]. This prediction problem presents significant challenges due to the limited size of the dataset, which comprises only 6000 rows and 2000 unique patients. Furthermore, crucial features are missing, such as the BMI and age of the male partner, Antral Follicle Count (AFC), duration of infertility, and alcohol/smoking habits of both parents. Additionally, data on specific metrics, like the number of follicles measuring between 10 and 14 mm in diameter on the day of hCG injection, and many other pertinent features, are absent. Some data, such as the living habits of the male partner, are challenging to collect, and the lack of these features could be a key reason why conventional machine learning models struggle with this dataset. Therefore, given these limitations, it would be unrealistic to expect a single model trained on this dataset to perform exceptionally across multiple datasets.

Methods I. Stage-by-Stage Modeling

Due to the sequential nature of IVF, a frozen cycle can be divided into consultation, thaw, uterus prep, and transfer phases, culminating in the pregnancy/live birth outcome. Directly estimating live birth or pregnancy outcomes usually requires collecting results from all these stages and processing them through a machine learning model. However, this approach may not be feasible nor provide

timely insights for patients or clinicians.

Therefore, we segmented a full frozen cycle into distinct phases: consultation, thaw, uterus prep, and transfer, excluding the final post-transfer phase, which would involve predicting pregnancy/live birth outcomes. At each phase, we identified key measurable outcomes and built models to predict them using only the available predictors at that stage. The architecture of this stage-by-stage model is illustrated in the figure 3 below.

The outcomes at each phase, along with general measurements and data from any previous fresh cycles, become the input for model building in the subsequent phase. Our dataset includes frozen cycle patient data, with some patients having undergone fresh cycles, hence these general measurements incorporate the patient's previous data from fresh cycles, pathology, and demographics. The initial consultation, the first step in the IVF process, is an opportunity for the clinical team to learn more about the patient's medical history and begin designing a customized IVF treatment plan. After a comprehensive work-up of Day 3 hormone levels and other preliminary tests, the care team prepares patients to begin their IVF cycle.

In the thaw phase of Frozen Embryo Transfer (FET), a previously frozen embryo is thawed and transferred into a prepared uterus for the possibility of having a baby. From consultation to the thaw phase, we expect to have general demographic data such as BMI, age, smoking status, previous abortions, prior IVFs, gravida, para, and other fresh cycle features. The outcome variables we aim to classify at this stage include tests like PGS TE bx thaw, PGS D3 bx thaw, and clinical outcomes like Concatenated Embryo Quality, number of embryos survived, and number of vials thawed. These predictions achieve much higher performance than directly estimating live birth or pregnancy outcomes. The FET-IVF cycle with hormonal support starts at the end of the previous menstrual cycle, similar to a conventional IVF cycle. Medications like GnRH agonist Lupron are administered to control the reproductive cycle. Following menstruation, a baseline ultrasound and blood work are ordered, and if favorable, estrogen supplementation begins.

At the Thaw → Uterus Prep phase, we include all previous phase features plus the predicted outcomes, focusing on variables like Max P4 (Frozen Query), and Last Endo Thickness Before Transfer. Elective Single Embryo Transfer (eSET) is conducted for suitable cases, transferring a single healthy embryo to maximize the chance of a healthy pregnancy. Three to five days after egg retrieval and fertilization, embryos are transferred into the uterus. If an embryo implants and grows, pregnancy results. Any unused embryos may be frozen for future use.

During the Uterus Prep → Transfer phase, we incorporate all previous phase features and outcomes, predicting the number of embryos transferred. After transfer, patients undergo a post-transfer period where they continue progesterone therapy for two weeks. Pregnancy is confirmed by blood test and ultrasound, and the patient is monitored until delivery.

Methods II: Subgroup Modeling: Finite Mixtures of Binomial Logit Regressions

In the realm of in-vitro fertilization (IVF) combined with machine learning (ML), a pervasive challenge emerges in the form of heterogeneity. The inherent diversity of patient populations, each with its unique set of characteristics, treatment protocols, and demographics, contributes to this complexity. Collecting comprehensive data variables that encompass the entirety of this heterogeneity is an arduous task, often hindered by logistical and ethical constraints. Notably, variables critical for predicting IVF outcomes, such as patient-specific health metrics, lifestyle factors, and individualized treatment plans, vary significantly across patient groups. Consequently, attempting to create a unified predictive model that accounts for this heterogeneity becomes a formidable challenge. It is in this context that the utilization of a Finite Mixture of Logistic Regression (FMLR) proves to be both justified and highly valuable [12]. FMLR provides an elegant framework to accommodate the diverse subpopulations within IVF data, allowing for the development of more tailored and accurate predictive models [13].

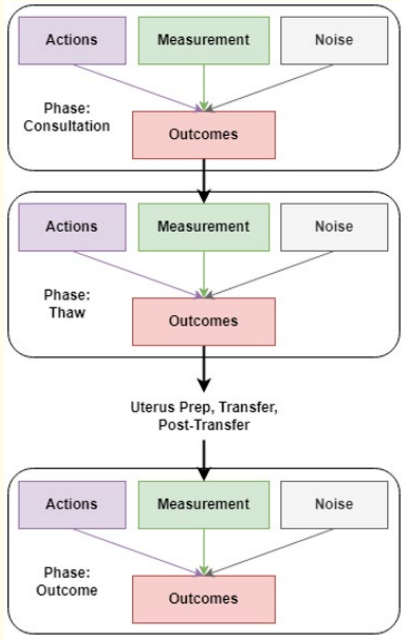

Figure 3: Phase-by-Phase Modeling.

The FMLR model can be represented as follows: Let Yi represent the binary outcome for the ith observation (0 for failure, 1 for success), and let Xi denote the vector of predictor variables for that observation. The FMLR model is expressed as:

$P(Y_i = 1) = \sum_{j=1}^{J} \pi_j \cdot P_j(Y_i = 1 \mid X_i, \beta_j)$, where j represents the number of latent classes or subgroups in the mixture model. Πj is the probability of belonging to the jth latent class. Pj is the logistic regression probability of success for the jth class, where βj denotes the vector of regression coefficients specific to that class.

Results and Discussion

Results Interpretation I. Stage-by-Stage Modeling

Table 3 illustrates a breakdown of our phase-by-phase prediction of different clinical variables. These findings may suggest that a phase-by-phase modeling approach holds high promise for clinicians, enabling them to systematically analyze the potential for pregnancy at each step of the IVF process. Furthermore, the interpretation of these improved predictions indicates that data from fresh cycles can serve as strong predictors for subsequent results in frozen cycles. The transition from thawing to the Uterus Preparation phase and from Transfer to the Outcome phase presents the greatest challenges for the models in terms of prediction accuracy. This insight potentially explains why IVF, as a field, still poses considerable complexities for machine learning specialists. Our results underscore the valuable role of ML in offering insights to clinicians, particularly during the consultation-to-thaw and uterus prep-to-transfer phases.

Variable Name	*AUROC*	*Accuracy*
Consultation →Thaw Phase		
PGS TE bx thaw	0.92	0.86
PGS D3 bx thaw	0.84	0.90
Concatenated Embryo Quality	0.92	0.96
Thaw from cryo all	0.80	0.73
# Embryos thawed	0.73	0.68
# Embryos survived	0.78	0.71
# AH'd	0.84	0.82
# Embryos available at thaw	0.72	0.67
# Vials thawed	0.76	0.70
Thaw →Uterus Prep Phase		
Max P4 (Frozen Query)	0.72	0.74
Last Endometrium Thickness	0.76	0.73
Uterus Prep →Transfer Phase		
#Embryos Transferred	0.95	0.94

Table 3: Phase-by-Phase Modeling Performance.

Phase-by-phase models offer a valuable avenue for interpreting the significance of clinical factors at each stage of the IVF process. In this context, we provide illustrative examples of modeling results that align with clinical sensibility. Within the Thaw → Uterus Prep phase, we achieved successful predictions for the "Embryos Trans" variable, representing the number of embryos successfully transferred during the subsequent transfer phase. Notably, several pivotal clinical factors emerged as key contributors to our predictive model, shedding light on their importance in shaping the outcome:

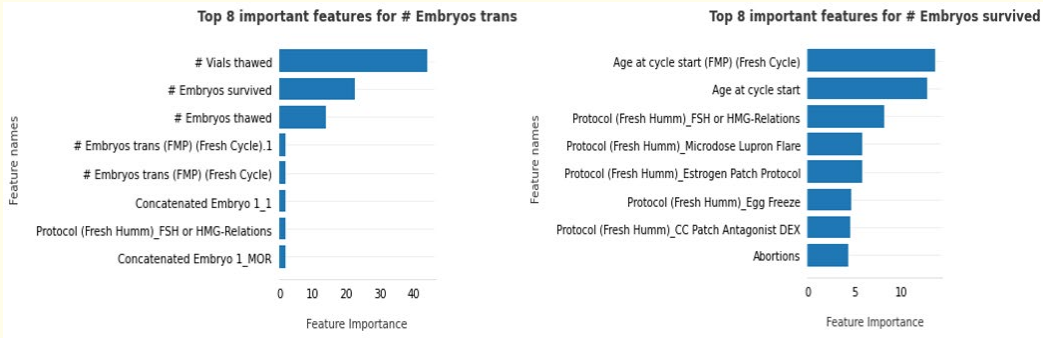

Figure 4: Interpretable results from Phase-by-phase modelling.

We could see that number of vials thawed, number of embryos survived, and number of embryos thawed, are all outcome variables in our previous phase models (thaw), and this means our design of a phase-by-phase split is clinically sensible and successful. Another example here would be the "Embryos survived" variable at the Thaw Phase. Here we could see that the model uses clinically sensible factors such as age, FSH hormone, and previous abortions.

Results Interpretation II. Subgroup Modeling

We picked age, number of embryos transferred and ICM grade as predictors, constructed age as a polynomial variable with a degree k from 1 to 5 and apply the FMLR model having an increasing number of components C from 1 to 5 to fit. We used a combination of Bayesian Information Criterion (BIC) and Cohen's Kappa as the indicator to select the most appropriate number of components. Our best performing model was with C=4 and k=3, with a test Cohen's Kappa of 0.889 and accuracy of 0.95 on 1389 test data points.

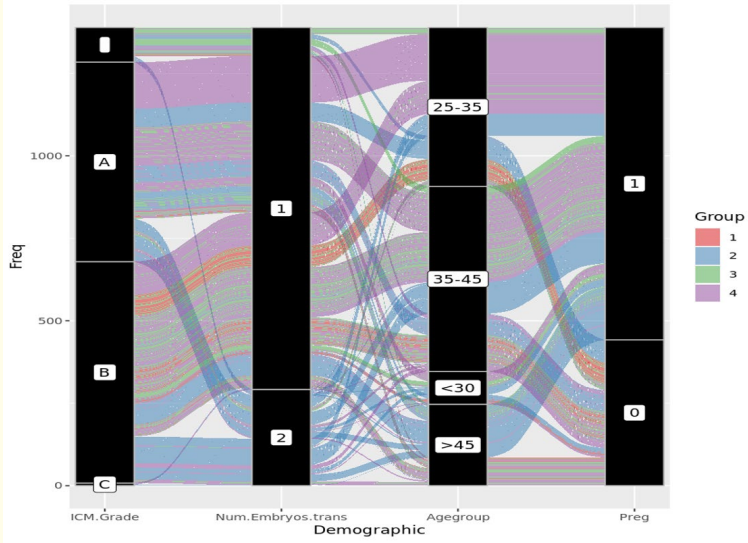

Figure 5: Alluvial Diagram of Identified 4 Subgroups.

Figure 5 shows an alluvial diagram of the 4 hidden subgroups, characterized by different age, embryos transferred, ICM grade distributions that lead to different pregnancy outcomes. Group 1 has mid ICM grade, low number of embryos transferred, 25-45 years old; group 2 has mid-to-high ICM grade and usually older than 30 years old; group 3 has mid tier ICM grades and usually older; group 4 has highest ICM grade, small number of embryos transferred and youngest age. This analysis underscores the value of subgroup modeling in helping clinicians discern hidden patterns within heterogeneous data. By revealing these distinct subgroups, clinicians gain a deeper understanding of the patient population, enabling them to tailor treatments and interventions more effectively to specific subgroups, ultimately leading to improved patient outcomes. Distinctively, group 1s has below 40 chance of getting pregnant while Group 4 has over 88% of chance of getting pregnant, which could help clinicians separate out these patient subgroups early and design different treatment correspondingly.

Conclusion

In conclusion, the endeavor to predict pregnancy and live births in the realm of in-vitro fertilization (IVF) has presented formidable challenges over the years, marked by inconsistent performance across various studies. This paper has been dedicated to a thorough review and critical analysis of the limitations inherent in current research practices. To chart a course towards more robust predictive models in the future, we have introduced a standardized machine learning pipeline, offering invaluable guidance to forthcoming researchers in this field.

Moreover, the core contribution of this work lies in the proposition of two alternative modeling approaches: phase-by-phase modeling and subgroup Finite Mixture of Logistic Regression (FMLR) modeling. These innovative approaches exhibit a harmonious blend of interpretability and high-performance prediction. Notably, they not only enhance predictive accuracy but also provide clinicians with clinically sensible explanations, aiding them in making informed decisions at crucial junctures of the IVF process.

One of the key takeaways from our research is the illumination of the intricate nature of the IVF cycle, offering insights into the challenges encountered by machine learning tasks in discerning "when," "who," and "where" these difficulties manifest most acutely. Such revelations have the potential to invigorate future endeavors in data collection and patient engagement processes, further refining the efficacy of IVF prediction models.

Conflict of interest

None.

References

1. Templeton Allan, Joan K Morris and William Parslow. "Factors that affect outcome of in-vitro fertilisation treatment". The Lancet 348.9039 (1996): 1402-1406.
2. Minaretzis Demetrios., et al. "Multivariate analysis of factors predictive of successful live births in in vitro fertilization (IVF) suggests strategies to improve IVF outcome". Journal of assisted reproduction and genetics 15 (1998): 365-371.
3. Hunault Claudine C., et al. "A prediction model for selecting patients undergoing in vitro fertilization for elective single embryo transfer". Fertility and sterility 77.4 (2002): 725-732.
4. Nelson Scott M and Debbie A Lawlor. "Predicting live birth, preterm delivery, and low birth weight in infants born from in vitro fertilisation: a prospective study of 144,018 treatment cycles". PLoS medicine 8.1 (2011): e1000386.
5. Van Loendersloot LL., et al. "Individualized decision-making in IVF: calculating the chances of pregnancy". Human Reproduction 28.11 (2013): 2972-2980.
6. Qiu Jiahui., et al. "Personalized prediction of live birth prior to the first in vitro fertilization treatment: a machine learning method". Journal of translational medicine 17.1 (2019): 317.
7. McLernon David J., et al. "Predicting the chances of a live birth after one or more complete cycles of in vitro fertilisation: population based study of linked cycle data from 113 873 women". BMJ 355 (2016).
8. Meijerink AM., et al. "Prediction model for live birth in ICSI using testicular extracted sperm". Human Reproduction 31.9 (2016): 1942-1951.
9. Metello José Luis, Claudia Tomás and Pedro Ferreira. "Can we predict the IVF/ICSI live birth rate?". JBRA Assisted Reproduction 23.4 (2019): 402-407.
10. Raef Behnaz, Masoud Maleki and Reza Ferdousi. "Computational prediction of implantation outcome after embryo transfer". Health Informatics Journal 26.3 (2020): 1810-1826.
11. Goyal Ashish, Maheshwar Kuchana and Kameswari Prasada Rao Ayyagari. "Machine learning predicts live-birth occurrence before in-vitro fertilization treatment". Scientific reports 10.1 (2020): 20925.
12. Kaptein Maurits and Paul Ketelaar. "Maximum likelihood estimation of a finite mixture of logistic regression models in a continuous data stream". arXiv preprint arXiv:1802.10529 (2018).
13. Li Gen. "Application of finite mixture of logistic regression for heterogeneous merging behavior analysis". Journal of Advanced Transportation (2018): 1-9.