A Case Study of the Challenges of Applied Machine Learning in Assisted Reproductive Technology

Anonymous Authors[1]

Abstract

Machine learning based pregnancy and live birth prediction in in-vitro-fertilization (IVF) field has always been a challenging task as it is hard to produce consistent performance across different studies. In this paper, we review and analyze the limitations of current work, implement a standardized machine learning pipeline as a guideline for future researchers, and propose two alternative modeling approaches (phase-by-phase and subgroup modeling). The two proposed alternatives help improve the prediction performance, provide clinically sensible explanations and timely guidance for users, and most importantly help us understand when, who and where of the entire IVF cycle is hard for ML tasks and inspire future efforts in data collection and patient engagement processes.

1. Background and Introduction

In recent years, medical practitioners have increased interest in utilizing machine learning models for pregnancy and live birth prediction. Specifically, they want to identify the significant couple factors affecting the outcome of in vitro fertilization (IVF). Previous research provides a comprehensive overview of applying machine learning techniques under the IVF scenario. However, little has been done to analyze the reproducibility and shortcomings of the previous studies. In general, many models have been tested, and some reported decent test performance. Nevertheless, few models could be adopted for real-life prognosis and diagnosis due to their lack of external validation. The models seldom perform well when researchers apply them to a real-life population. And, a domain shift can broadly compromise the models' generalizability and reproducibility. Thus, further evaluation and analysis are needed to better standardize the machine learning workflow in the future.

Previous work on IVF pregnancy and live-birth prediction using machine learning models reported inconsistent AUC results ranging from 0.6 to over 0.95. This study examined the most standard techniques from the previous work (i.e., logistic regression, decision tree, XGboost, and SVM) using Boston IVF Fertility Clinic data. Nevertheless, our evaluation also yielded inconsistent AUC results. Thus, we delved further to examine why IVF prediction is such a difficult task and propose our conjectures on two potential approaches (stage-by-stage and subgroup modeling) for improving model performance and real-life generalizability.

In vitro fertilization (IVF) treatment is a sequential and time-consuming process with many sub-cycles. Many patient and treatment factors will change during the process and affect the evaluation of potential pregnancy and live birth. Two main issues arise during the IVF treatment process when we apply machine learning models. First is the time discrepancy between feature collection and model usage. Some models require a large number of features. The dilemma is that many of these features will not be available until we run various medical tests, whereas we run the models in real time. In other words, we will not collect enough features at the time when we apply the models. The second issue is data leakage, defined as using features that result from the clinician's prediction of the treatment outcome. The data leakage issue exists vastly in machine learning and in vitro fertilization literature. Thus, it is questionable whether having a machine learning prediction of pregnancy or live birth only a few days before the real-life outcome is of practical benefit.

On the other hand, models which generate live birth/pregnancy predictions only based on preliminary results and patient demographics, may be too arbitrary and deterministic for such a dynamic treatment progress and could potentially give patients false hope or early discouragement. What clinicians really need may be a stage by stage approach to help them better outline the treatment progress and for patients it may be more practical to have machine learning models provide them with shorter term insights.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Here we divide our paper into three main parts,

1) Review of past literature and analysis of flaws

2) Attempts to reproduce a regular "ML pipeline" on our dataset, drawbacks of doing so, and insights gained for standardizing IVF ML workflow in the future

3) Discussion of potential improvement techniques for current ML for IVF models: Subgroup modeling approach and Phase by phase modeling approach.

2. Literature Review

First we conducted a comprehensive literature review into the past work from 1997 to 2021, examining the performance of N machine learning in IVF prediction papers and the difference in their approach, datasets, and interpreting their different in performance. The AUCs from the literature reviewed have a wide range of values, varying from 0.6 to 0.95. We found out that the variability in the availability of important features and changing demographics of the IVF patients, rather than the actual modeling approaches contribute to the differences in AUC values.

There are three main reasons why accurately predicting pregnancy or live birth in an IVF procedure can be challenging. To begin with, due to different data sources and volumes, it is not quite possible to replicate the findings of previous studies. The Boston IVF dataset deployed in this research has a limited size of around 6000 rows and only 2000 unique patients (can give a more accurate number). Some of the previous studies with higher AUC had a larger dataset. Previous studies all have varying data sizes and some are not evaluated on a validation set.

Secondly, after reviewing the variables in previous studies, there are important features such as BMI of males, age of the male, alcohol/smoking habits of both parents, and many other features that are not included in across datasets. Many of these variables might help with higher prediction accuracy.

Moreover, the demographics of patients who receive IVF treatments have changed over the past decades. For instance, when the study was done in the 90s, Caucasians made up the majority of the patients, constituting 91.5% of the patient population from 1994 to 1998. African Americans, Asians, and Hispanics accounted for 4%, 3%, and 1.5% of the population, respectively. IVF treatments have become more readily available to a more diverse demographic group. Although race and ethnicity information is not included in our dataset, it is reasonable to assume that patients receiving treatments today vary greatly from the data collected in the 90s.

We were also able to identify and experiment on some common flaws of the existing ML+IVF literature and the whole literature review table is as follow. These flaws could be summarized into two categories: exclusion of sub-populations and data leakage.

Exclusion of certain sub-populations is one of the potential drawbacks we found throughout the current work. For example, in —Individualized decision-making in IVF: calculating the chances of pregnancy—, 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. Also, in —Predicting live birth, preterm delivery, and low birth weight in infants born from in vitro fertilisation: a prospective study of 144,018 treatment cycles—, the research excluded cases where the embryo is related to donor/surrogate mother and the cycles using frozen embryos, which is in fact a large component of common IVF data points. These exclusions of sub-populations artificially increases the variance in patient populations and deteriorates the model's ability to generalize.

Another of the mistakes that occurs most frequently when modeling based on IVF data is using information that would not be available to the model at the time it needs to make its prediction. Machine learning models operating in a clinical setting typically have to run in real time, and at the time they are run, not every feature available in the dataset might be available for the model to use. In fact, there are several tests that are inside the IVF dataset along with other demographical features but would not deliver results until pregnancy or at least a short time before pregnancy, such as B-HCG. Another form of this mistake, commonly referred to as data leakage is to use features whose value is obtained as a result of the clinician's prediction of the outcome of the patient. Using these features leaks information about the true label that the model is trying to predict, even though in practice the model would be used to help the clinician come up with their prediction in the first place, and therefore it wouldn't have access to these data-leaking features. For example, in one of the papers that have the most "outstanding" performance, —Computational prediction of implantation outcome after embryo transfer—, authors actually used B-HCG as a predictor. We plugged in B-HCG as a predictor in our model and also achieved AUROC as high as over 90 per cent, however, HCG is a hormone produced in the body during pregnancy and such prediction is not useful at all.

These two existing flaws inspired us to come up with two alternatives to the traditional ML workflow, phase by phase modeling and subgroup modeling, which will be discussed in later sections.

3. Analysis of Standard Modeling Workflow

Next we replicated a 'regular' machine learning workflow for pregnancy and live birth on our dataset using as comprehensive methods in the existing literature as possible, which could be broken down into exploratory data analysis, data pre-processing,and model fitting sections, so as to test the limitations of common ML+IVF pipelines.

Exploratory data analysis

To better understand the dataset, we performed exploratory data analysis on distribution of variables, total number of cycles and dropout rate.

To begin with, we studied the distribution of 15 predictor variables, including age, BMI and other test results, between the pregnant and or pregnant groups. From the histogram, it is difficult to uncover significant difference in the distribution of these variables.

We further studied the total number of cycles that each patient has in the dataset. Most patients went through only one cycle and the maximum number of cycles in this study is 8. There are in total 5196 cycles and 4050 of those are cycle 1.

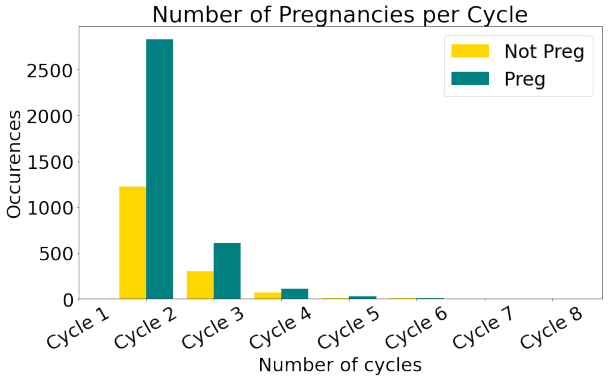

Figure 1. Number of pregnancies vs non-pregnancies per cycle

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 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's no successful pregnancy for cycles after 6.

After understanding the trend between number of cycles and pregnancy, we want to further analyze patients behavior and gain insight into why certain patients stayed for extra cycles after a failure and some chose to stop the treatment.

We analyze the dropout rate for each cycle. Dropout rate for each cycle is defined as the percentage of people who didn't continue another cycle after not getting live birth. The

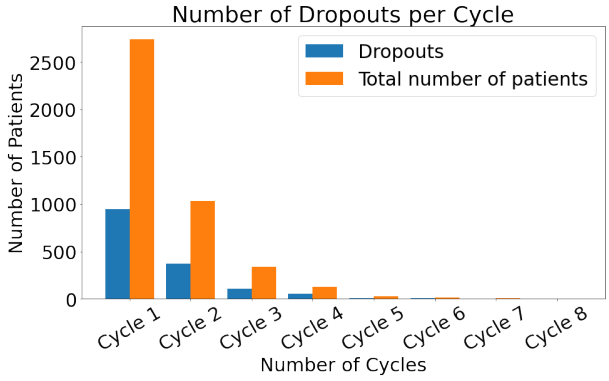

Figure 2. Number of dropouts per cycle

dropout rate remains relatively stable for the first 3 cycles, 0.34, 0.357 and 0.317 respectively.

There is large number of variables in the dataset as a result of the lab tests that doctors order for each patient. It is important to understand whether some of these variables are correlated, so we could potentially group them in later modeling approaches. From the Pearson correlation graph in figure 3, we have identified that some data collected from the partners of IVF participants are somewhat correlated. Findings of correlated variables will help inform our feature selection process and phase by phase modeling approach.

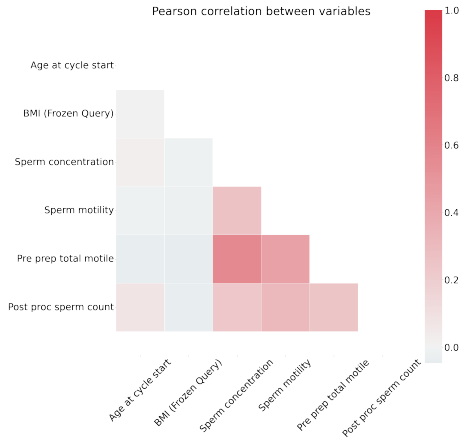

Figure 3. Pearson correlation between variables

Data preprocessing

Our prediction problem is challenging because the dataset we got is of limited size as it includes 6000 rows and only 2000 unique patients. Also, there are important features such as BMI of male, age of male, AFC, duration of infertility, alcohol/smoke for both parents, number of follicles measuring between 10 and 14 mm in diameter on the day of hCG injection and many other features that are not included in our dataset. Some data are hard to collect such as the

living habits of the male partner, and the absence of these features could be a potential reason why regular machine learning models couldn't excel on this dataset.

Feature selection is another big challenge for the Boston IVF Clinic dataset as many important features have over 0.5 missingness, which are tricky to impute. Hence we tried out three different approaches:

In the data processing step, since many features have over 50% missingness, it doesn't make much sense to impute those variables. The missingness could be a result of the clinician recognizing some symptoms and asking specific patients to get the tests. This results in missingness for the patient population who were not ordered the test. How we handled it it to drop features with over 0.5 missingness. In total, 51 features, most of which are outcome related (e.g. live birth-related information and dates), are dropped. We then imputed the remaining of the missing data using KNN iterative imputer. In terms of feature selection, there are 364 total features after one hot encoding. We experimented with three feature selection techniques including , L1-based feature selection, tree-based feature selection and clustering features to check we are not missing out on important features. There are X features that went into the training of the model.

Using Imputation: We experimented with different imputing mechanisms, including iterative imputing, which refers to a process where each feature is modeled as a function of the other features, and knn imputation, where each sample's missing values are imputed using the mean value from n_neighbors nearest neighbors found in the training set. Selecting which of these two imputation techniques do not have significant impacts on the performance. The best results on pregnancy using Logistic, Random Forest and XGboost models are around 0.65 and on live birth are around 0.63, quite hard to break through the 0.7 clinically viable threshold.

No Imputation: We looked at the missing values across all subjects and found that some "boundary predictors" that are quite important are around the 0.5 missingness drop threshold, such as Max_E2 (Frozen Query), Max_LH (Frozen Query), and pregnant outcome for previous fresh cycles. Hence it is highly possible that the poor prediction performance could be caused by the absence of these predictors. Performance may increase with these predictors added back even at the trade-off of smaller sample sizes. We instead experimented with directly dropping rows that have the original set of predictors and our decided boundary predictors and test if there are significant changes in results. The best results on pregnancy using Logistic, Random Forest and XGboost models are around 0.67 and on live birth are around 0.69, quite hard to break through the 0.7 clinically viable threshold.

No Imputation but add an Indicator: In medical settings, sometimes missingness could be a hidden indicator of the patient's overall situation and may not need imputation. Take an arbitrary example, the missing of a male fertility test and corresponding sperm quality value may imply that a patient is using sperms with high quality from donors and the clinician thinks there is no need for a sperm quality test at all. Here imputing with either the mean of all sperm quality or KNN inferred value of the most similar patients' sperm quality may mislead the model. Thus we are adding an indicator value of whether a test exists to all the data in order to help the model capture this information. The best results on pregnancy using Logistic, Random Forest and XGboost models are around 0.68 and on live birth are around 0.69.

Model fitting and performance

We tested Logistic Regression, Random Forests and XGBoost models and tuned the hyperparameters over a large set of combinations using grid search. For Logistic Regression, we searched over the solver, norm of the penalty λ and magnitude of λ; for Random Forests, we searched over the max depth of trees, gamma,regularization lambda, scale positive class weight, subsample rate and number of trees over 2000 combinations; for XGBoosts, we searched over the same set of hyperparameters as Random Forests.

Here is a performance chart of different combinations of methods we used:

Preprocess	Best Model	AUC$_{Preg}$	AUC$_{Birth}$
Impute	Random Forest	0.65	0.63
Drop	XGBoost	0.67	0.69
Indicator	XGBoost	0.68	0.69

Table 1. Standardized ML Modeling Performance

4. Alternative Modeling Approaches

As discussed in previous sections, IVF birth and pregnancy modelling has been a difficult task to tackle and there have been flaws in previous literature that has not been addressed. We would like to evaluate why this is such a hard problem for ML models to digest, more specifically which phases of the frozen cycle IVF are the hardest to model, and which patients subgroups display the most unpredictable characteristics. These two modeling approaches also help provide early insights into overall success rate and designs of clinical procedures for the medical practitioners.

Stage by stage modeling

Apart from these findings that could possibly help us standardize the machine learning workflow for IVF prediction, we also consider whether a direct estimate of live

birth/pregnancy is plausible, given the literature review above, and helpful for clinical decisions or patient self-assessment. Our proposal is that a phase-by-phase approach may be a better solution.

Due to IVF's sequential nature, a frozen cycle could be divided in to consultation, thaw, uterus prep and transfer phases and finally the pregnancy/live birth outcome. A directly estimation of live birth/ pregnancy usually requires collection of all results for all of these stages and run these features through a machine learning model. Doing so may not be plausible and provide timely insights for patients or clinicians.

So instead, we divided a full frozen cycle into In each phase of the IVF treatment, into consultation, thaw, uterus prep and transfer phases, excluding the final post-transfer phase which would be equivalent to predicting a pregnancy/live birth outcome. For each phase, we identify a number of important measurable outcomes and build models to predict them using only the available predictors at that stage. The architecture of stage-by-stage model could be described by the figure below:

The outcomes at each phase, plus other general measurements and fresh cycle data (if the patient has been through fresh cycle before), become the input data for model building in the next phase. Our dataset is frozen cycle patient data with some patients having previous fresh cycles, hence these general measurements will include the patient's previous data at fresh cycles, pathology and demographics of themselves and their partners.

The initial consultation, the very first step in the IVF process, is an opportunity for the Boston IVF clinical team to learn more about patients' medical history and begin to design a customized IVF treatment plan that addresses your goals and maximizes your success. Once a comprehensive work-up of the Day 3 hormone levels and other preliminary tests have been completed, the care team will prepare the patients to begin their IVF cycle.

The thaw phase for Frozen embryo transfer, or FET, is an assisted reproductive technology procedure in which a previously frozen embryo is thawed and transferred it into an appropriately prepared uterus in order to have a baby.

At consultation →thaw phase, we expect to have patients general demographics data such as BMI, age, smoking or not, previous abortions, prior IVFs, gravida, para and other fresh cycle features. The outcome variables we are trying to classify at consultation to thaw stage include whether some tests should be conducted like PGS TE bx thaw, PGS D3 bx thaw, and some clinical outcomes like Concatenated Embryo Quality, number of embryos survived, Thaw from cryo all, number of AH'd, and number of vials thawed. These predictions achieve much higher performances than directly

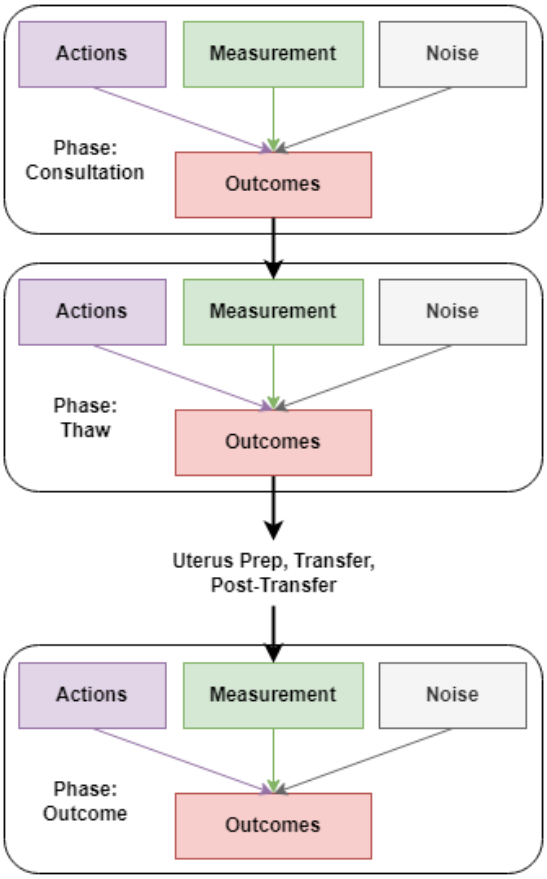

Figure 4. Phase by Phase Model Workflow

estimating the live birth or pregnancy outcome. Here is a table of the performance metrics for our consultation →thaw prediction model.

Metric\Outcome name	AUROC	Accuracy
PGS TE bx thaw	0.92	0.86
PGS D3 bx thaw	0.84	0.9
Concatenated Embryo Quality	0.92	0.86
Thaw from cryo all	0.8	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

Table 2. Phase-by-Phase Modeling Performance : Consultation →Thaw Phase

By interpreting the results, we saw that Prior_FET (Frozen Query), Trigger_Day_P4 (Fresh Humm) drug, Para, Gravida (Frozen Query), Protocol (Fresh Humm)_OCP-Long Lupron, Abortions are dominant factors on the test variables and BMI, Age, Gravida and fresh cycle protocols are contributing most to the prediction of concatenated embryo qualities. Features contributing to the prediction of surviving embryos include Luteal Estrogen, FSH, number of embryos transferred at previous fresh cycle and number of abortions.

A FET-IVF cycle with hormonal support starts at the end of the previous menstrual cycle, much like a conventional IVF cycle. Injections of a drug meant to control and shut down the reproductive cycle are given. Usually, the GnRH agonist Lupron is used, but other pituitary-suppressing medications may be chosen instead. Once you get your period, abaseline ultrasound and blood work are ordered. If all looks good, estrogen supplementation is started.

At Thaw →Uterus Prep phase, we include all the features in the previous phase plus the outcome variables predicted in the previous phase, and predict these variables: Max_P4 (Frozen Query), and Last_Endo_Thickness_Before_Transfer. Here are their prediction performance:

Metric\Outcome name	AUROC	Accuracy
Max_P4 (Frozen Query)	0.72	0.64
Last_Endo_Thickness	0.6	0.6

Table 3. Phase-by-Phase Modeling Performance : Thaw →Uterus Prep Phase

To those who fit the criteria for treatment, elective single embryo transfer (eSET) will be conducted, which is the process of transferring one single healthy embryo, rather than a few. This is the first choice in selected couples because we know that transferring multiple embryos is associated with multiple pregnancies, and the greatest chance for a healthy pregnancy comes from a single-child pregnancy. Three to five days after egg retrieval and fertilization, embryos are transferred into the woman's uterus. This procedure occurs in-clinic using a catheter inserted through the cervix. Most women are able to resume regular activities the next day. If an embryo sticks to the uterine lining and grows, pregnancy results. Any unused embryos may be frozen to allow the option of future implantation.

At Uterus prep →Transfer phase, we include all the features in the previous phase plus the outcome variables predicted in the previous phase, and predict the number of embryos transferred.

Metric\Outcome name	AUROC	Accuracy
#Embryos Transferred	0.99	0.97

Table 4. Phase-by-Phase Modeling Performance : Uterus Prep →Transfer Phase

After transfer, patients will undergo post transfer period : To help thicken the uterine lining in order to make it easier for the embryo to implant, they will continue progesterone therapy for two weeks after embryo transfer. A blood test for pregnancy is performed several days after the embryo transfer. in the affirmative case pregnancy is confirmed by ultrasound. Patient is monitored until after delivery. The Transfer →Outcome phase is basically same as our baseline models implemented in section 3 and results would not be repetitively included here.

These results may suggest that phase by phase modeling would be a good approach for clinicians to dissect the evaluation of potential pregnancy step by step and the interpretation of these higher performance predictions indicates data from fresh cycles are great predictors of the following results in frozen cycles.

More importantly, this helps us learn that thaw →Uterus Prep phase and Transfer →Outcome phase indicates highest levels of difficulty for the models to predict, which could be the potential reason to explain why IVF in general is still a fathomable field to machine learning specialists. However, such results still proved that ML is very helpful for providing insights for the clinicians at consultation →thaw phase and uterus prep →transfer phase.

Subgroup modeling

Since modeling using the full dataset didn't yield result with promising AUC, we wanted to explore alternative methods to predict pregnancy or live birth. In the EDA and literature review process, we proposed a hypothesis that there are certain variables that are more highly correlated with the outcome of IVF. Subgrouping the data by such variables can provide insightful information in terms of how well the

machine learning model performs for different groups. Age and ICM grade are the two variables that we use to subgroup our dataset.

ICM stands for inner cell mass and is graded into three categories: A, many tightly packed cells; B, many loosely grouped cells; and C, very few cells.(Reference) ICM is studied to be an indicator of pregancy outcome. According to Reproductive Medicine Center's study in 2021, the miscarriage rate of blastocysts with ICM grade A was lower, compared with ICM grade C. Since ICM grade is something that can be evaluated before birth, our hypothesis is that patients with ICM grade A is more likely to deliver live birth than patients with ICM grade C. In our experiment, we want to explore whether our model can predict patients with the same category of ICM grade with higher accuracy.

We manually divided the patients into ICM groups A, B and C and excluded the ICM variable itself from the modelling pipeline. Our results for different groups are shown below.

	AUC	Accuracy
ICM_A	0.65	0.6
ICM_B	0.63	0.65
ICM_C	0.75	0.75

Table 5. ICM-Grade Subgroup Modeling Performance

Age is also an important subgroup to take into consideration. 2 out of 3 women who start IVF before age 35 will take home a baby within three IVF cycles. Women under 30 have a 44% chance of a live birth in their first IVF cycle; Women under 30 have a live-birth rate of between 69% and 92% after seven cycles. Women aged 40-44 have an 11% chance of a live birth in their first IVF cycle; Women aged 40-44 have a live-birth rate of between 21-37% after eight cycles. The average change of taking home a baby with each IVF cycle is 30%; 33% of moms undergoing IVF get pregnant during their first IVF cycle; 54-77% of women undergoing IVF get pregnant by the eighth cycle; The average chance of taking home a baby with each IVF cycle is 30%. Hence our hypothesis is that different age groups shall also display different levels of prediction difficulties.

Same as the ICM subgroups modelling pipeline, we manually divided the patients into age groups Under 35, 35-40, 40-45 and above 45 and excluded the age variable itself from the modelling pipeline. Our results for different group are shown below.

Which parts of IVF and what patients are the hardest to model?

As for parts of IVF across phases, consultation →thaw and uterus prep →transfer phases are much easier to predict than thaw →uterus prep and transfer →outcome phases, and could potentially act as clinically viable machine learning models to suggest the designs and expectations of IVF processes.

	AUC	Accuracy
Under 35	0.62	0.59
35-40	0.64	0.63
40-45	0.68	0.65
Above 45	0.6	0.68

Table 6. Age Subgroup Modeling Performance

As for different composition of patient populations, age groups have not shown any difference while ICM grade group C is significantly easier to predict than other two classes.

Potential Clinical Insights from the models

Phase by phase models could help us interpret what clinical factors are important to measurements and actions taken at each phase of the IVF process. Here we illustrate some clinically sensible examples of the modeling results. In the Thaw →Uterus Prep phase, we have successfully predicted the "Embryos Trans" variable, which stands for the number of embryos successfully transferred at the transfer phase, and here are some top clinical factors contributing to our prediction:

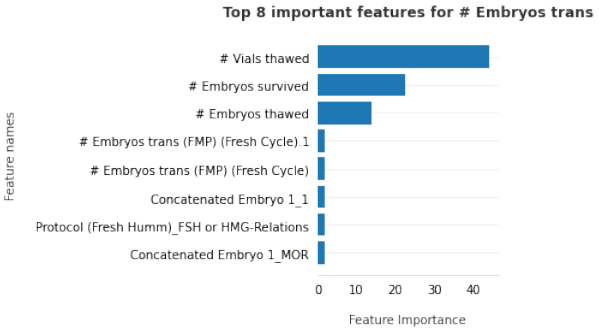

Figure 5. Embryos Transferred Feature Importance

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 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.

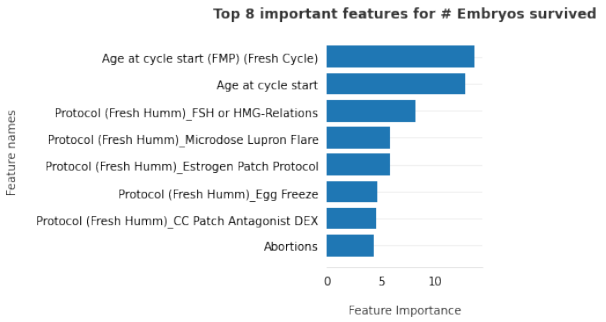

Figure 6. Embryos Survived Feature Importance

5. Conclusion

In this paper, we performed a throrough literature review to analyze the limitations of current work of machine learning in IVF. We also re-implemented and compared against approaches mentioned in previous literature and illustrated the methods and improved results of two alternatives we proposed: phase by phase model as well as the subgroup modeling approach.

It is possible that these model architecture might not be the solution to this problem, given many important data points could be missing from the data collection process. Many male-related features as well as ethnicity and race-related features are poorly represented in the dataset which could be one of the reasons why it is difficult to achieve a high AUC. Some potential next steps could be 1) work with clinicians to design more informative patient intake surveys, 2) include Bayesian modeling and linear mixture effect model to study latent variables to better capture clusters within the patient groups.

6. Citations and References

Acknowledgements

Do not include acknowledgements in the initial version of the paper submitted for blind review.

If a paper is accepted, the final camera-ready version can (and probably should) include acknowledgements. In this case, please place such acknowledgements in an unnumbered section at the end of the paper. Typically, this will include thanks to reviewers who gave useful comments, to colleagues who contributed to the ideas, and to funding agencies and corporate sponsors that provided financial support.

References

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Ai, J., Jin, L., Zheng, Y., Yang, P., Huang, B., Dong, X. (2020). The Morphology of Inner Cell Mass Is the Strongest Predictor of Live Birth After a Frozen-Thawed Single Embryo Transfer. Frontiers in Endocrinology. https://doi.org/10.3389/fendo.2021.621221

A. You *can* have an appendix here.

You can have as much text here as you want. The main body must be at most 8 pages long. For the final version, one more page can be added. If you want, you can use an appendix like this one, even using the one-column format.

Published Year	Paper	Train Data Year Range, Size	Val Data Year Range, Size	ML Models	Outstanding Features	AUROC	Outcome Y	Used X	External Validation?	Potential Limitations
1996	Factors that affect outcome of in-vitro fertilisation treatment	1991-1994,36961	Not mentioned	Logistic		Not mentioned				
1997	Multivariate Analysis of Factors Predictive of Successful Live Births in In Vitro Fertilization (IVF) Suggests Strategies to Improve IVF Outcome	range unknown, 554	Not mentioned	Logistic (no interaction terms)	maternal age (negative), number and quality of embryos (positive)	Unknown (model fit was not evaluated)	pregnancies, live births, and multiple birth deliveries vs. IVF failure	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”	NO	
2002	A prediction model for selecting patients undergoing in vitro fertilization for elective single embryo transfer	1993-1998, 642 women undergoing their first IVF treatment cycle in which no more than two embryos were transferred.	Not mentioned	multiple logistic regression	the development stage, the morphology score of the two best embryos available for transfer, the age of the patient.	AUC: 0.68 for pregnancy, 0.71 for twin pregnancy.	the chances of singleton and twin pregnancy when one or two embryos are transferred.	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)”	No	

Published Year	Paper	Train Data Year Range, Size	Val Data Year Range, Size	ML Models	Outstanding Features	AUROC	Outcome Y	Used X	External Validation?	Potential Limitations
2011	Predicting live birth, preterm delivery, and low birth weight in infants born from in vitro fertilisation: a prospective study of 144,018 treatment cycles	2003-2007, 144018	Not mentioned	Logistic		0.6335	Live birth			Excluded donor/surrogate mother, frozen embryos
2013	1Individualized decision-making in IVF: calculating the chances of pregnancy	2001-2009, 2621	2009-2011, 500	Logistic	No	0.68	ongoing pregnancy	Female age, duration of subfertility, previous ongoing pregnancy, male subfertility, diminished ovarian reserve, endometriosis, basal FSH, number of failed IVF cycles.		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.
2016	Personalized prediction of live birth prior to the first in vitro fertilization treatment: a machine learning method	2014-2018, 7188 women, 70% for train	30% for validation	logistic regression, random forest, extreme gradient boosting (XGBoost), support vector machine (SVM)		AUC: logistic regression (0.71), random forest (0.73), XGBoost (0.73), SVM (0.71)	the live birth chance prior to the first IVF treatment	Age, AMH, BMI, duration of infertility, previous live birth, previous miscarriage, previous abortion and type of infertility	No	1. limited generalization of the model to other populations. 2.can only be used for couples who have never accepted IVF treatment, limited application. 3.failed to account for family genetic history and lifestyle factors.

Published Year	Paper	Train Data Year Range, Size	Val Data Year Range, Size	ML Models	Outstanding Features	AUROC	Outcome Y	Used X	External Validation?	Potential Limitations
2016	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	1999-2008, 184269	Not mentioned	Logistic regression, backwards selection	increasing age of the woman, increasing number of eggs collected and the cryopreservation of embryos, treatment year	pretreatment model 0.73 (0.72 to 0.74); post-treatment model 0.72 (0.71 to 0.73)	Cumulative chances of a first live birth for a couple having up to six complete cycles of IVF. One complete cycle included all fresh and frozen embryo transfers resulting from one episode of ovarian stimulation.	Pretreatment model: number of complete cycles, patient characteristics (woman's age, duration, treatment type, year first complete cycle started, tubal infertility, male factor infertility, unexplained infertility, anovulatory infertility, previous pregnancy in couple). Post-treatment model: number of complete cycles, patient characteristics (woman's age, duration, year first complete cycle started, tubal infertility, previous pregnancy in couples), treatment information at first complete cycle (cryopreservation of embryos, number of eggs collected, stage of embros transferred).	NO	
2016	Prediction model for live birth in ICSI using testicular extracted sperm	2007-2015, 526 couples undergoing 1006 TESE-ICSI cycles	2007-2015, 289 couples undergoing 553 TESE-ICSI cycles.	multivariable logistic regression	a lower male LH, a higher male testosterone level, sperm motility	AUC: 0.62	live birth in couples undergoing ICSI after successful testicular sperm extraction (TESE-ICSI).	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.	Maybe	Paternal BMI as a predictor may help improve the model if the values are not missing a lot.

Published Year	Paper	Train Data Year Range, Size	Val Data Year Range, Size	ML Models	Outstanding Features	AUROC	Outcome Y	Used X	External Validation?	Potential Limitations
2019	Can we predict the IVF/ICSI live birth rate?	2012-2016, 739 IVF/ICSI cycles	Not mentioned	Binary regression without interaction terms	categorized or exponentialized women's age, categorized or logarithmized AFC, categorized or logarithmized AMH; ovarian reserve measures, couples undergoing treatment for ovulation disorder or pure male factor	0.688 (0.649-0.728)	Live birth in fresh cycle	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	NO	Subgroups were created after a post hoc analysis of the data and this might be a source of bias.

Published Year	Paper	Train Data Year Range, Size	Val Data Year Range, Size	ML Models	Outstanding Features	AUROC	Outcome Y	Used X	External Validation?	Potential Limitations
2020	Computational prediction of implantation outcome after embryo transfer	April 2016 to February 2018, 500 patients (one cycle each)	train-test split by 80/20 (but also said using 10-fold cross-validation)	KNN, SVM, Neural Networks, Naive Bayes, Random Forest, Decision Tree	FSH/HMG dosage, contraception duration and the number of germinal vesicle (GV) quality oocytes	87 to .97	-HCG (human chorionic gonadotropin)/Pregnancy	Clinical data (patient-related data): Age of female, Age of male, BMI (body mass index), Family relation of couples Smoking, Type of infertility, Infertility duration, Contraception duration, Infertility in family, G (gravida/gravidity), P (para/parity), Ab (abortion), EP (ectopic pregnancy), L (living children), Comorbidity diseases, Anemia, Thyroid, disease, Prolactin hormone disorders, Drug usage. Female pathology data: Amenorrhea (absence of menstruation), Dysmenorrhea (painful periods), Period status, Hirsutism (excessive body hair in women), Galactorrhea (abnormal milky breast discharge), Gynecological surgery, Oocyte donation, AFC (antral follicle count), Endometrium (tissue lining of the uterus) thickness Three-line (regular/normal), endometrium, Uterus depth, Size of follicles, Tubal factor, Pelvic factor, Cervical factor, Ovulatory factor, PCOS (polycystic ovary syndrome), Uterine factor, Endometriosis (abnormal growth of endometrium in outside of the uterus cavity), Endometrial factor, Vaginitis, RIF (repeated implantation failure), RPL (recurrent pregnancy loss). Male pathology data: Male factor, Male genital surgery, Varicocele (abnormal enlargement of the testicular veins), TESE (testicular sperm extraction), PESE (percutaneous epididymal sperm extraction) Fresh/freeze sperm. Semen analysis data: Sperm count, Normal morph, Immotile. Lab tests: FSH (follicle-stimulating hormone), LH (luteinizing hormone) Estradiol, vitD3 Levels, Oocyte stimulation and morphology: FSH/HMG (human menopausal gonadotropin) dosage, GnRH (gonadotropin-releasing hormone) antagonists Dosage, GnRH agonists dosage, Duration of stimulation (days), Estradiol dosage, No. estradiol days, Number of retrieved oocytes, Number of MII (metaphase II) quality oocytes, Number of MI (metaphase I) quality oocytes, Number of GV (germinal vesicle) quality oocytes, Number of degenerated quality oocytes, Quality of injected MII oocytes. Embryological data: Number of 2PN (pronuclear), Number of developed embryos, Quality of developed embryos, Quality of vitelline space, ET (embryo transfer) strategies, ET day, Number of transferred embryos, Number of blastomeres, Quality and stages of transferred embryos Experience of ET. Others: PRP (platelet-rich plasma), ID (identification)	Maybe	Authors used B-HCG as a predictor. However, HCG is a hormone produced in the body during pregnancy and thus should not appear as a predictor for pregnancy.