

A Survey on Deep Learning for IVF Live Birth Prediction using Embryological, Imaging Based, Data Driven Approaches with Performance Metrics

Ms. V. Kalaiselvi¹, Dr. S. Poongodi²

¹*Research Scholar, Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore, Tamilnadu, India.*

²*Associate Professor, Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore, Tamilnadu, India.*

Abstract— Choosing the embryo most likely to result in a live birth is the focal point of the context of in vitro fertilization. Recently, deep learning strategies have shifted the emphasis from conventional grading toward objective models capable of considering a broad range of variables associated with the prediction of clinical outcomes. The aim of this paper is to review the literature on the use of deep neural networks in the prediction of live birth through the combined analysis of embryological, imaging, and clinical variables. The findings are grouped according to the type of data used for training the model and the architecture of the algorithms. For each work, a brief description of the methodology is given, and the evaluation results are underlined. In particular, models that include both imaging data and clinical information showed the best performance in most of the studies with AUC up to 0.97 and accuracy from 74% to 82% on retrospective data. The main limitations of the current evidence, as well as the directions for further research are discussed.

Index Terms— Convolutional neural networks (CNN), Recurrent neural networks (RNN), Long short-term memory (LSTM), Vision transformers (ViT), Deep neural networks, Machine learning, Classification algorithms, Predictive modeling, Blastocyst images, Morphokinetic analysis, Image-based prediction

I. INTRODUCTION

IVF treatments are a popular choice among couples who want to have children but are infertile. However, the success rate per cycle is relatively low, and a significant portion of IVF failures can be attributed to embryo selection. The conventional embryo assessment is based on morphological criteria

established by embryologists. More recently, time-lapse embryo incubators have been used to track embryo development and allow for more information to be obtained about embryo viability. Despite that, embryo assessment remains a subjective process that does not guarantee that the transferred embryo will lead to a live birth.

Meanwhile, deep learning can be applied to find associations between various pieces of data related to the embryo and the outcome of the treatment. Unlike traditional machine learning, neural networks can detect patterns in high-dimensional data such as embryo images or videos directly. Most of the research into applying deep learning to embryo assessment is focused on finding predictive power in different types of clinical data, such as the morphological features of the embryo or its developmental timing. The ultimate goal for embryo assessment is to predict live birth; however, most studies into embryo assessment only deal with intermediate outcomes such as implantation or clinical pregnancy.

The purpose of this paper is to review and analyze the application of deep learning to embryo assessment in IVF treatments. The works presented in this study are sorted according to the type of data they utilize for the prediction of the embryo's viability. Moreover, this paper will discuss the neural network architectures utilized in these studies, as well as the metrics used to evaluate the performance of these architectures. Finally, the article will summarize the key points of each paper and attempt to highlight the areas that could be subject to further research.

II. CLINICAL CONTEXT AND OUTCOME DEFINITIONS

2.1 Where AI fits into the IVF workflow

In an IVF cycle, eggs are extracted, fertilized and cultured in a laboratory for a few days. The embryos created are then graded to choose the ones that will be implanted or frozen for future use. The new artificial intelligence tools can help in this decision-making process by predicting the likelihood that an embryo will result in a live birth if it is transferred to the uterus.

2.2 What “success” means

Studies may differ in what they define as a successful outcome:

Implantation, which is the attachment of the embryo to the uterine lining.

Clinical pregnancy which is typically defined by the presence of a fetal heartbeat that can be picked up on ultrasound.

Live birth means a birth of at least one living infant after viable gestational age (which may vary by region and study).

Live birth is a more clinically relevant endpoint than implantation or other surrogate outcomes because it is subject to additional variables such as the receptivity of the uterus and overall patient health not related to the embryo itself. As such, predictive models that are trained or validated on live birth outcome data are preferable to those that do not.

III. TYPES OF DATA USED FOR PREDICTION

3.1 Embryological and morphokinetic features

Embryologists have been using manually annotated events (e.g., t2, t3, t4, etc.) and blastocyst formation as indicators of embryo viability. Morphology scores (e.g., Gardner scoring for blastocysts) represent additional information useful in embryo assessment.

These manually annotated parameters can be valuable in deep learning pipelines either as inputs to more traditional machine learning or shallow networks.

As targets to be predicted by models that estimate such parameters from time-lapse movies.

Auxiliary tasks to regularize the learning process of deeper architectures.

Recent works highlight the importance of standardized benchmarks for these manually annotated parameters

to judge the predictive power of the learned models before attempting to use them as an input to downstream tasks such as outcome prediction.

3.2 Imaging data: single frames versus time-lapse videos

Static images represent a moment of the embryos' development (for example, day 3 or day 5) and are easily captured, stored, processed, and analyzed; however, they only capture information of a specific moment. The continuous or frequent capturing of embryos allows one to observe the morphological changes associated with embryo development and extract more information, such as temporal patterns. The method is more data-intensive (and requires more processing power and storage space) and is inconsistent between different forms of capture or equipment. The reviewed articles on the use of AI in embryo assessment primarily focused on accuracy and AUC-ROC values, which imply a significant emphasis on discriminatory power.

3.3 Patient-level clinical variables

The quality of an embryo is not the sole indicator of pregnancy success. Maternal characteristics such as age, body mass index, ovarian reserve markers, infertility diagnosis, previous IVF attempts, and stimulation protocol are important predictors of live birth potential that, when combined with morphological parameters, improve prognostic modeling capability.

There is growing evidence that predictive models using both image-based and clinical data outperform those using either type of information.

IV. NEURAL NETWORK ARCHITECTURES IN USE

4.1 Convolutional networks for still images

The convolutional neural networks (CNN) such as ResNet, EfficientNet, or DenseNet architectures are most commonly applied to process embryo images in the task of predicting live birth. The typical pipeline consists of the following stages:

Image preprocessing (cropping the embryo area, normalization, augmentations).

The CNN backbone pretrained on natural or domain-specific images (e.g., breast imaging) with the classification head.

The optional stage of combining image features with the corresponding clinical data.

Note that CNN classifiers typically show moderate performance (AUC around 0.68–0.73) for the problem of prediction of live birth or related outcomes.

4.2 Sequence models for time-lapse data

Time-lapse videos are inherently sequential. There are several common architectures of such models:

3D CNNs: apply convolution operations in both spatial and temporal dimensions.

CNN + recurrent layer: CNN feature extractor is followed by a recurrent layer (LSTM/GRU) which learns video features.

Transformers/Video Transformers: use self-attention mechanism to learn relations between different developmental stages.

Hybrid models: combine CNNs and Transformers, typically using CNN for feature extraction and Transformer for temporal up sampling.

The results for time-lapse based approaches are highly variable (AUC ranges from roughly 0.64 to 0.97) probably due to different dataset and varying model challenges.

4.3 Multimodal fusion strategies

The models that predict live birth outcomes most effectively are those that use a mix of data sources:

An image encoder (usually a CNN or Transformer) that processes embryo imagery.

A distinct clinical encoder (frequently a smaller multilayer perceptron) to capture patient and cycle-specific variables.

A fusion module that merges these two representations either at a late (prediction-level) or early (feature-level) stage (or through some intermediate fusion with attention-based weighting).

A review of literature, so far indicates that these multimodal architectures typically attain AUC scores around 0.97 and accuracy in the low-to-mid 70s (or even low 80s) on retrospective datasets.

4.4 Foundation models and transfer learning

Large and labeled IVF datasets tend to be scarce, and many groups have started from the foundation model pretrained on generic images. More recently, some groups proposed domain specific foundation model pretrained on millions of embryo images. These models can learn general representation of various

aspects of embryo structure and development that can be used for downstream task such as prediction of ploidy, quality score, and ultimately live birth rate. The results show that they enable data efficient and generalizable model compared to training from scratch.

V. DATASETS, LABELING, AND PRACTICAL CONSTRAINTS

5.1 Characteristics of existing datasets

Key properties affecting the model behavior are:

Scale: Number of embryos, cycles, and patients.

Center diversity: single-center versus multi-center studies.

Imaging hardware: Different time-lapse systems yield different image quality.

Lab protocols: Variations in culture media, incubators, and procedures.

Outcome definitions: Live-birth definitions, multiple embryo transfers, and methods for dealing with censored cases.

Most published models are based on single-center retrospectively collected datasets resulting in concerns about the validity of their findings.

5.2 Annotation challenges

The main weaknesses associated with the present study are as follows:

Morphokinetic labels could have been annotated manually (with potential inter observer discrepancies) or automatically generated.

Outcome labels were inconsistent between studies that utilized live birth and those that used implantation or clinical pregnancy as the outcome of interest.

Data screening and selection, including the removal of images or recordings with poor quality or insufficient data, was not consistently reported.

The absence of publicly available standardized datasets is also a major limitation of the current research, challenging the credibility and objectivity of the conducted comparison.

VI. HOW PERFORMANCE IS MEASURED

6.1 Metrics commonly reported

The embryo assessment literature most frequently cites the following performance metrics:

Accuracy (ACC): Proportion of correctly predicted outcomes, easy to interpret but can be misleading in case of class imbalance.

AUC-ROC: Probability that a randomly chosen positive instance will be ranked higher than a negative instance, insensitive to threshold selection and less affected by class imbalance.

Precision, recall, and F1-score: Become critical when the costs of false positives and false negatives are different and when outcomes such as live birth rates are low.

Calibration measures (e.g., Brier score, calibration plots): Are increasingly viewed as essential for clinical decision-support but are underreported.

A few studies report average precision, Jaccard index, or task-specific evaluation metrics.

6.2 Interpreting numbers in the IVF setting

A few studies report average precision, Jaccard index, or task-specific evaluation metrics. AUC: Scores above 0.80 are typically regarded as good, while 0.70-0.80 is considered acceptable, and values under 0.7 are poor. A pooled analysis of AI used for embryo selection reports an average AUC of roughly 0.7-0.81, depending on the outcome and the type of model.

Accuracy: This can be misleading when live birth rates are low, and it should be read in light of prevalence and other metrics.

Precision/recall: It may be better to have high recall so that viable embryos are not discarded or high precision so that unnecessary embryo transfers with low odds of success are avoided. This depends on the clinician's context and priorities.

Calibration: This is essential for interpreting model estimates as probabilities (for example, an embryo has a 30% chance of leading to a live birth”).

6.3 Common evaluation pitfalls

Frequent methodological issues include:

Inadequate sample size or analysis design that leads to imprecise estimation of key metrics.

Data leakage during train/test split (e.g., at embryo level instead of patient/cycle level).

Reporting best possible performance with no consideration of decision criteria or thresholds.

Lack of external validation, and multiple comparisons.

Adequate methods would include patient-based split of training/test set, reporting of confidence intervals for relevant metrics such as sensitivity and specificity, and using clinically relevant frameworks to assess performance (e.g. decision-curve analysis) as well as external validation and avoid multiple comparisons by using pre-specified models and analysis metrics.

6.4 What the Numbers Look Like in Practice

Single-modality models:

The static, time-lapse (TL) and clinical variables (CV) only models perform relatively poorly compared to the combined models.

Static image-only models typically showed the worst performance, frequently achieving accuracy in the low-to-mid 60s percent and often had a low area under the curve (AUC).

TL only models had a broad range of performance, from AUC of approximately 0.64-0.97 and accuracy of approximately 64-78%.

CV- only models could have an AUC of up to approximately 0.80 and accuracy of approximately 76-78%.

6.5 Multimodal systems

Models that merged embryo images with clinical information proved to be reliably better:

Accuracy: Often around 74-82%, while certain specific methods reached even higher values.

AUC: The best multimodal architectures are able to reach approximately 0. with many high-quality models achieving values between 0.85 and 0.91.97, alongside.

Precision/recall/F1: When these are reported, they generally fall between 0.75 and 0.91 for fusion frameworks that work well.

As an illustration, a single fusion architecture combining embryo images and clinical data attained roughly 82% accuracy, 91% average precision, and 0.91 AUC in the prediction of clinical pregnancy, clearly surpassing unimodal approaches. A further study that applied attention-based deep learning and refined the feature selection process reported about 80.5% accuracy and 85.96% AUC across outcome categories that included live birth.

VII. SURROGATE ENDPOINTS VERSUS LIVEBIRTH

Many papers used implantation or clinical pregnancy as surrogates for live birth. When analyzing live birth as the endpoint, performance was slightly worse (possibly due to extra noise introduced by post transfer events), but the same hierarchy (best: multimodal, next time-lapse, then only clinical and finally static imaging) was observed.

VIII. OPEN CHALLENGES AND RESEARCH DIRECTIONS

8.1 Generalization across clinics and devices

Model trained in one setting often fail to generalize to another because of differences in:

Imaging equipment and settings.

Culture media, incubators, and laboratory workflows.

Patient demographics and treatment protocols.

Methods such as domain adaptation, invariant representation learning and multi-center training have been proposed, but are not yet widely used.

8.2 Uncertainty and interpretability

When it comes to clinical applicability, achieving accurate predictions alone is not enough; one must also be capable of describing the level of certainty and explaining the reasoning behind a particular prediction;

Probabilistic models offer a way to estimate uncertainty in predictions: for instance, Bayesian frameworks, model ensembles, or conformal prediction.

Interpretability may be attained via saliency maps, attention visualizations, and concept-based reasoning, yet the clinical value of these approaches remains to be rigorously assessed.

8.3 Fairness and bias

Potential sources of inequity may include:

Under representation of certain age groups, ethnic backgrounds, or infertility diagnoses.

Clinic practices associated with different socioeconomic statuses.

Stratified performance reporting and fairness audits are needed to ensure that benefits are distributed equitably.

8.4 From retrospective studies to prospective trials

The majority of the evidence collected thus far comes from retrospective studies:

In order to demonstrate real-world effectiveness, the field requires prospective studies that compare the outcomes of AI-assisted and standard embryo selection, using live birth as the primary endpoint. Such trials ought to report metrics concerning time to live birth, the number of embryo transfers needed, multiple pregnancy rates, and cost-effectiveness. Preliminary evidence indicates that algorithms can enhance the speed of embryo selection in the setting of multiple embryo transfers by identifying the embryos most likely to lead to a live birth, though further research is needed to establish whether this translates into faster success for patients.

IX. REPRODUCIBILITY AND COMMUNITY INFRASTRUCTURE

The progress of this field is currently limited by a number of factors:

The private nature of many datasets: Raw images and annotations are not publicly available.

Inconsistent reporting: Different studies report varying metrics, thresholds, or data splits.

A lack of standardized benchmarks: Few studies provide publicly available tasks for morphokinetic prediction or live birth prediction.

Community efforts would benefit from building upon existing resources such as:

De-identified multi-center repositories with standardized outcome labels.

Standardized benchmark tasks with agreed-upon train/test splits and evaluation protocols.

Open-source baseline implementations to reduce barriers to entry and enable meaningful performance comparisons.

X. CONCLUSION

Deep learning has enabled new opportunities for selecting IVF embryos with the highest chances of resulting in a live birth. By utilizing embryological measurements, imaging data, and patient variables, recent studies have shown that multimodal models are capable of achieving close to 0.97 AUC and around 75-85% accuracy rates in retrospective settings. Nevertheless, there is a need to address generalizability, reproducibility, uncertainty, and

prospective clinical utility challenges to truly bring such models into the clinical reality and improve the much-needed live birth rates.

REFERENCES

- [1] Sengupta, G. Majumdar, A. Choudhury, P. Gupta, D. Modi, and P. Narad, "Artificial intelligence-enhanced in-vitro fertilization outcome prediction for the Indian subpopulation: Integrating pre-treatment parameters and Bayesian-optimized ensemble techniques," *Int. J. Community Med. Public Health*, vol. 12, no. 5, pp. 2272–2279, May 2025, doi: 10.18203/2394-6040.ijcmph20251387.
- [2] K. R. Reddy, M. Asif, G. Deotale, and V. G. S. Priya, "Artificial intelligence in embryo selection: Enhancing precision and overcoming traditional limitations in in vitro fertilization," *Int. J. Reprod., Contracept., Obstet. Gynecol.*, vol. 15, no. 2, pp. 789–797, Jan. 2026, doi: 10.18203/2320-1770.ijrcog20260214.
- [3] P. Malathi and D. M. Gomathi, "Deep learning classification of embryo quality based on blastocyst," *Int. J. Innov. Res. Technol.*, vol. 11, no. 8, pp. 1284–1289, 2025.
- [4] P. Malathi and M. Gomathi, "Enhancing in-vitro fertilization outcomes through machine learning: A comparative analysis of predictive models and dataset evaluation," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 12, no. 4, pp. 156–163, 2024, doi: 10.22214/ijraset.2024.65652.
- [5] R. Kodali, V. R. Dhulipalla, V. S. K. Tatavarty, M. Nadakuditi, B. Thiruveedhula, S. Gunnam, D. P. Bavirisetti, and G. P. Reddy, "Interpretation of deep learning model in embryo selection for in vitro fertilization (IVF) treatment using CNN-LSTM architecture," *arXiv preprint arXiv:2506.06680*, 2025, doi: 10.48550/arXiv.2506.06680.
- [6] M. Maheshwari and I. V. Reddy, "Artificial intelligence in embryo selection: Revolutionising IVF laboratory practices," *Int. J. Innov. Sci. Res. Technol.*, vol. 10, no. 5, pp. 2492–2494, May 2025, doi: 10.38124/ijisrt/25may1396.
- [7] S. Divya and S. Geetharani, "A comprehensive survey on deep learning techniques for embryo quality assessment in IVF," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 12, no. 3, pp. 892–901, 2024.
- [8] M. Ishaq, S. Raza, H. Rehar, S. E. Z. U. Abadeen, D. Hussain, R. A. Naqvi, and S.-W. Lee, "Assisting the human embryo viability assessment by deep learning for in vitro fertilization," *Mathematics*, vol. 11, no. 9, Art. no. 2023, 2023, doi: 10.3390/math11092023.
- [9] S.-A. Sadegh-Zadeh, S. Khanjani, S. Javanmardi, B. Bayat, Z. Naderi, and A. M. Hajiyavand, "Catalyzing IVF outcome prediction: Exploring advanced machine learning paradigms for enhanced success rate prognostication," *Front. Artif. Intell.*, vol. 7, Art. no. 1392611, 2024, doi: 10.3389/frai.2024.1392611.
- [10] L. Boucret, F. Chabrun, M. Boguenet, *et al.*, "Deep-learning model for embryo selection using time-lapse imaging of matched high-quality embryos," *Sci. Rep.*, vol. 15, Art. no. 28068, 2025, doi: 10.1038/s41598-025-10531-y.
- [11] S. K. Tadepalli and P. V. Lakshmi, "Deep learning in IVF to predict embryo quality from blastocyst images," in *Advances in Intelligent Computing and Communication*, 2020, pp. 1145–1153, doi: 10.1007/978-981-15-7961-5_136.