ResFTNet: Clinical Pregnancy Outcome Prediction via Multimodal Feature Fusion for Assisted Reproductive Technology

Yang Wang[1], Xiaomei Tong[2], Zhenming Yuan[1,3*]
[1]School of Information Science and Technology, Hangzhou Normal University, Hangzhou, China
[2]Assisted Reproduction Unit, Department of Obstetrics and Gynecology, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China
[3]Hangzhou Hele Tech Co.LTD
wangyang0706@stu.hznu.edu.cn, 3406028@zju.edu.cn, zmyuan@hznu.edu.cn

***Abstract*—Accurately predicting clinical pregnancy outcomes in assisted reproductive technology (ART) remains a critical and challenging problem, due to the complex interplay between embryo morphological characteristics and patient-specific clinical factors. This study proposes ResFTNet, an end-to-end multimodal fusion network for predicting clinical pregnancy outcomes (i.e., successful versus unsuccessful pregnancy after embryo transfer). In the image branch, a residual convolutional backbone enhanced with omni-dimensional dynamic convolution and squeeze-enhanced axial attention is employed to capture both fine-grained local morphology and global semantic representations from microscopic embryo images. In parallel, the tabular branch adopts an FT-Transformer to model nonlinear dependencies among structured clinical indicators. Based on these modality-specific encoders, ResFTNet jointly integrates visual and clinical representations through a Clinical Context Injection (CCI) strategy, which injects patient-specific clinical information into the visual feature space to alleviate cross-modal semantic misalignment. Experimental results on a private clinical pregnancy dataset denoted as the CP-ART dataset and a public blastocyst dataset demonstrate that ResFTNet consistently outperforms existing unimodal and multimodal approaches in terms of accuracy and robustness. Furthermore, interpretability analyses using Grad-CAM and SHAP indicate that the proposed model focuses on biologically meaningful embryonic regions and clinically relevant features, highlighting its potential as an effective and explainable decision-support tool for personalized treatment planning in reproductive medicine.**

***Index Terms*—Multimodal Fusion, Clinical Context Injection, Dynamic Convolution, Attention Mechanism, Assisted Reproductive Technology**

I. INTRODUCTION

Infertility is defined as the failure to conceive after one year of unprotected intercourse, affecting approximately one in every six couples worldwide [1]. Assisted Reproductive Technology (ART) refers to medical procedures performed in laboratory settings to manipulate human oocytes, sperm, or embryos with the aim of helping infertile patients achieve pregnancy. Among these techniques, In Vitro Fertilization (IVF) and Intracytoplasmic Sperm Injection (ICSI) are two of the most widely applied methods.

A complete ART cycle typically encompasses critical stages ranging from ovarian stimulation and oocyte retrieval to fertilization and the final embryo transfer. Within this clinical workflow, selecting the embryo with the highest developmental potential is paramount for achieving pregnancy [2]. To enhance selection accuracy, advanced technologies such as Preimplantation Genetic Testing (PGT) and time-lapse imaging have been increasingly adopted to identify chromosomal abnormalities or monitor developmental kinetics. However, despite these technological strides, traditional assessment methods often lack standardization and fail to fully account for the intricate biological variability among patients. The success rate of ART-induced clinical pregnancy remains highly variable, arising from the complex interaction between embryo morphological quality and patient-specific physiological conditions [3]. This variability makes predicting outcomes accurately a challenging task in clinical practice, highlighting the limitations of relying solely on manual grading and underscoring the need for integrated predictive models.

In recent years, data-driven approaches based on machine learning and deep learning have been increasingly applied to predict ART outcomes. Existing studies can be broadly categorized into two groups. Image-based methods focus on embryo morphology by leveraging convolutional neural networks to assess developmental potential from microscopic or time-lapse images, while tabular-based approaches utilize structured clinical data to model patient-specific risk factors using statistical or neural models. However, these single-modality methods suffer from inherent limitations. Image-only models fail to incorporate critical clinical context, whereas tabular-only models overlook the direct visual quality of the transferred embryo. Moreover, the heterogeneity between high-dimensional visual features and low-dimensional structured clinical variables poses a significant challenge for effective multimodal integration. These limitations highlight the need for a unified framework that can jointly model and fuse heterogeneous modalities to achieve complementary information learning for accurate clinical pregnancy prediction. To address these challenges, this study proposes ResFTNet, a multimodal fusion network designed to integrate embryo images and clinical tabular data for predicting clinical pregnancy outcomes in ART. The key contributions of this work are as follows:

- We propose ResFTNet, an end-to-end multimodal learning framework for predicting binary clinical pregnancy outcomes in assisted reproductive technology, which jointly models microscopic embryo images and structured clinical tabular data within a unified optimization objective.
- We design an enhanced image feature extraction module, termed OD-SEABlock, by structurally integrating Omni-Dimensional Dynamic Convolution (ODConv) and Squeeze-Enhanced Axial Attention (SEA attention) into a residual backbone. This module improves the modeling of both local morphological patterns and global semantic structures in embryo images.
- We develop a Clinical Context Injection (CCI) strategy that projects visual representations into the clinical feature space before joint learning. This alignment-based fusion enables effective cross-modal interaction and complementary information integration, leading to improved prediction performance and interpretability, as validated through extensive experiments and explainability analyses.

II. Related Work

A. Deep Learning for Embryo Image Assessment

Deep learning approaches have achieved significant progress in automating embryo quality analysis, primarily focusing on Inner Cell Mass (ICM) morphology, blastocyst grading, and implantation prediction. Chen et al. [4] introduced BlastocystMask for instance-level segmentation of ICM and TE in blastocyst images, achieving state-of-the-art performance on a public dataset. Ai et al. [5] retrospectively analyzed over 10,000 cycles, identifying ICM morphology as a critical predictor, where Grade A ICM embryos achieved a 54.62% live birth rate compared to 28.45% for Grade C. To improve generalization, Berntsen et al. [6] trained a model on a large-scale dataset of 115,832 embryos across 18 centers, achieving an AUC of 0.95 on the overall dataset, demonstrating the scalability of CNNs in diverse clinical settings. Beyond static images, time-lapse technology has emerged as a powerful tool. Theilgaard-Lassen et al. [7] developed iDAScore v2.0 using temporal data, achieving AUCs between 0.62 and 0.70 for blastocyst grading. While these visual-only models excel at morphological assessment, they often overlook patient-specific physiological factors that critically influence pregnancy outcomes.

B. Clinical Data Modeling for ART Outcome Prediction

Parallel to image analysis, machine learning on clinical tabular data has been widely adopted to assess ART risks and predict live births. Vogiatzi et al. [8] constructed an Artificial Neural Network (ANN) identifying 12 critical predictive variables, achieving 76.7% sensitivity. Similarly, Liu et al. [9] developed a predictive model using seven core clinical features, yielding a robust F1 score of 77.18%. Large-scale studies, such as Tian et al. [10], employed Bayesian networks on over 100,000 cycles to predict fertilization failure with 91% accuracy. Metello et al. [11] further confirmed the strong correlation between outcome rates and indicators like Anti-Müllerian Hormone (AMH) and Antral Follicle Count (AFC). However, these methods rely exclusively on patient history and hormone levels, ignoring the direct visual quality of the specific embryo selected for transfer.

C. Multimodal Approaches for ART Outcome Prediction

Recent research has shifted towards multimodal fusion to leverage complementary information. Liang et al. [12] combined endometrial ultrasound images with clinical data, achieving an AUC of 0.825, which validated the synergistic value of multimodal inputs. In the context of embryo assessment, Zou et al. [13] integrated time-lapse parameters with maternal age to predict implantation potential (AUC = 0.78). Barnes et al. [14] introduced STORK-A, a prominent model fusing morphological, morphokinetic, and clinical data to predict chromosomal status with 69.3% accuracy across multiple clinics. More recently, Ouyang et al. [15] proposed DeFusion, which utilizes decoupled modules to process time-series images and parental indicators.

Despite these advances, most existing fusion strategies rely on simple concatenation or specific time-lapse data. Few studies have effectively addressed the heterogeneity between static high-resolution embryo images—the most common clinical record—and complex electronic medical records data containing both continuous and categorical features. ResFTNet aims to bridge this gap through adaptive feature extraction and deep multimodal interaction.

III. Method

Given an embryo image $\mathbf{x}_I \in \mathbb{R}^{H\times W\times C}$ and the corresponding clinical tabular features $\mathbf{x}_T \in \mathbb{R}^d$, the objective of this study is to predict the probability of achieving clinical pregnancy after embryo transfer. The proposed multimodal model learns a mapping

$$\hat{y} = f(\mathbf{x}_I, \mathbf{x}_T), \tag{1}$$

where $\hat{y} \in [0,1]$ denotes the predicted probability of clinical pregnancy.

ResFTNet is an end-to-end multimodal framework designed to integrate embryo microscopic images and structured clinical data. As illustrated in Fig. 1, the architecture comprises three main components: (1) an image feature extraction branch based on a ResNet backbone enhanced with ODConv and SEA attention to capture morphological details and global semantic features; (2) a clinical feature extraction branch employing the FT-Transformer to model nonlinear dependencies among

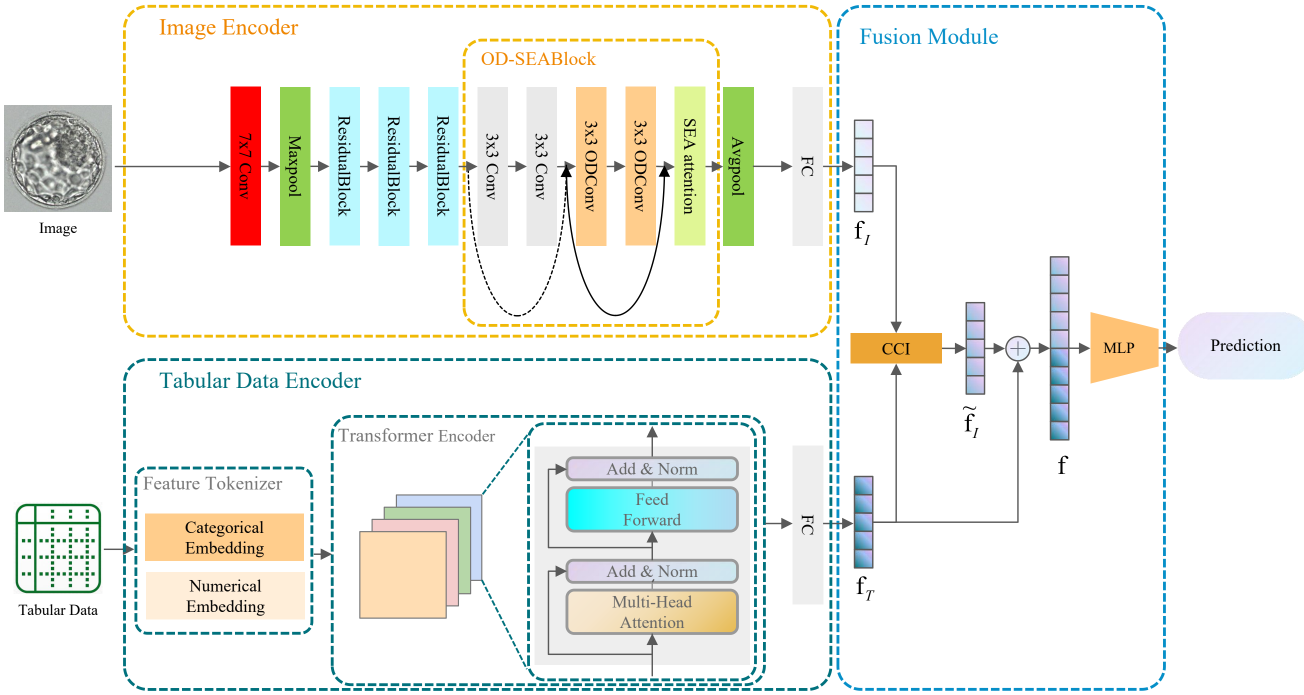

Fig. 1. The overall architecture of the proposed ResFTNet framework.

tabular variables; and (3) a fusion module that integrates these heterogeneous representations to predict clinical pregnancy probability via an MLP. By performing joint optimization within a unified objective, ResFTNet leverages complementary cross-modal information, thereby enhancing both predictive accuracy and interpretability compared to conventional single-modality approaches.

A. Image Encoder

The image encoder is built upon a ResNet backbone to extract high-level visual representations from embryo images. To enhance feature extraction capability, we integrate a customized module termed OD-SEABlock into the residual architecture. OD-SEABlock combines ODConv [16] and SEA attention [17] to jointly model fine-grained local morphology and long-range spatial dependencies.

Specifically, ODConv improves convolutional adaptability by dynamically adjusting kernel responses across spatial and channel dimensions, enabling robust modeling of morphological variations. SEA attention captures global contextual information along axial directions with low computational overhead. By embedding OD-SEABlock into the image encoder, the proposed model obtains more discriminative and semantically enriched visual representations for subsequent multimodal fusion.

The image encoder maps the input image to a compact visual feature vector:

$$\mathbf{f}_I = \phi_I(\mathbf{x}_I), \tag{2}$$

where $\phi_I(\cdot)$ denotes the image encoding function.

B. Tabular Data Encoder

The tabular data encoder adopts an FT-Transformer to model structured clinical data. By embedding numerical and categorical variables into a shared latent space and applying multi-head self-attention, the FT-Transformer effectively captures nonlinear dependencies among clinical indicators [18].

The FT-Transformer models dependencies among clinical variables using self-attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \tag{3}$$

where Q, K, and V denote the query, key, and value matrices, respectively.

The resulting tabular representation is given by:

$$\mathbf{f}_T = \phi_T(\mathbf{x}_T), \tag{4}$$

where $\phi_T(\cdot)$ denotes the tabular data encoding function.

C. Multimodal Feature Fusion and Classification Module

To effectively integrate heterogeneous visual and clinical information, we employ a CCI strategy. Since visual features and clinical indicators originate from disjoint semantic spaces with different dimensions, direct interaction is challenging. This module aligns the clinical information to the visual domain, using it as a semantic bias to recalibrate the embryo representation.

Specifically, the image features $\mathbf{f}_I$ are projected to the target dimension. These are integrated through an additive mechanism followed by non-linear activation:

$$\tilde{\mathbf{f}}_I = \mathrm{ReLU}\left(\mathbf{W}_I \mathbf{f}_I + \mathbf{W}_{align} \mathbf{f}_T\right), \tag{5}$$

where $\mathbf{W}_I$ and $\mathbf{W}_{align}$ are learnable projection matrices for image and tabular branches, respectively. By adding the clinical representation as a dynamic bias term, the visual features are adaptively shifted based on the patient's physiological background.

The aligned image feature $\tilde{\mathbf{f}}_I$ is then concatenated with the original tabular feature to form the unified multimodal representation:

$$\mathbf{f} = [\tilde{\mathbf{f}}_I; \mathbf{f}_T]. \tag{6}$$

Finally, the predicted probability is produced by a multilayer perceptron:

$$\hat{y} = \mathrm{MLP}(\mathbf{f}). \tag{7}$$

Due to the class imbalance between positive and negative samples, the model is trained using a weighted binary cross-entropy loss. The loss function is defined as

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \Big[-w^+ y_i \log \hat{y}_i - w^- (1 - y_i) \log(1 - \hat{y}_i) \Big], \tag{8}$$

where $\hat{y}_i \in [0, 1]$ denotes the predicted probability of clinical pregnancy for the i-th sample, and w^+ and w^- are class weights determined by the ratio of positive and negative samples in the training set.

IV. Experiments

A. Datasets

This study employs two datasets to evaluate the proposed model, both labeled in a binary manner indicating whether clinical pregnancy was achieved after embryo transfer.

The first dataset is a private clinical pregnancy dataset, referred to as the CP-ART dataset, collected from a reproductive medicine center. It includes 245 valid embryo transfer cases across diverse age groups. Each sample includes two modalities: high-resolution embryo images acquired before transfer and structured clinical data comprising ten variables: maternal age, paternal age, body mass index (BMI), AMH, basal follicle stimulating hormone (FSH), basal estradiol (E2), oocytes retrieved, available embryos, protocols for controlled ovulation stimulation (COS), and fertilization methods. All data were anonymized, and this retrospective study was conducted in accordance with institutional regulations.

The second dataset is the publicly available human blastocyst dataset by Kromp et al. [19], comprising 2,344 images from 837 patients along with associated clinical parameters for 752 fresh transfer cases. After preprocessing and matching, 732 paired image–clinical samples were obtained. This dataset provides complementary information for evaluating multimodal prediction performance.

B. Experimental Setup

The proposed model is implemented using the PyTorch framework. In the experiments, a differentiated multimodal training strategy is adopted, in which the image and tabular branches are optimized separately using the Adam optimizer to accommodate the distinct feature learning characteristics of each modality. Specifically, the ResNet18 backbone is employed as the image feature extractor with a learning rate of 1×10^{-5} and a weight decay of 1×10^{-6}. The FT-Transformer is used for tabular feature extraction with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}. For the fusion MLP, the learning rate is set to 1×10^{-4}. To stabilize the optimization process, gradient clipping with a maximum norm of 10 is applied. For datasets from two different sources, a five-fold cross-validation strategy is employed for both model training and evaluation. To enable fair comparison with existing studies, four widely used evaluation metrics are adopted: accuracy, precision, recall, and F1 score.

C. Experimental Results

To comprehensively evaluate the effectiveness of ResFTNet, comparative experiments were conducted against several representative methods on two datasets. The compared models encompass three methodological categories. The first group includes image-based models, namely EfficientNet [20] and ResNet, which perform binary classification using only embryo images. The second group consists of tabular-based models, including SVM and MLP, which rely exclusively on clinical tabular data. The third group comprises multimodal models, such as Perceiver, the approach proposed by Liu et al., STORK-A, and DeFusion. To ensure fair comparison on our static dataset, the STPE module of DeFusion was excluded, allowing all models to jointly utilize visual and clinical modalities for prediction.

As shown in Table I, ResFTNet delivers the best overall performance on the CP-ART dataset, achieving an accuracy of 65.71%, a precision of 71.08%, a recall of 76.77%, and an F1-score of 73.08%. The model surpasses unimodal baselines by more than ten percentage points in accuracy and twelve percentage points in F1-score, clearly demonstrating the benefit of combining embryo morphological information with patient-specific clinical indicators. Even when compared with advanced multimodal methods, ResFTNet maintains a consistent advantage, outperforming DeFusion and STORK-A by approximately two to three percentage points in both precision and F1-score. This improvement stems from the inclusion of the dynamic feature enhancement modules ODConv and SEA attention, which collectively strengthen adaptive local feature extraction and global semantic representation within a unified end-to-end fusion framework.

A similar result is observed on the Kromp et al. Blastocyst Dataset, as summarized in Table II. ResFTNet once again achieves the highest accuracy of 67.76% and an F1-score of 64.35%, outperforming all competing approaches, including

TABLE I
PERFORMANCE COMPARISON ON THE CP-ART DATASET.

Model	Modality		Performance Metrics (%)			
	Image	Table	Accuracy	Precision	Recall	F1 Score
EfficientNet	✓		50.61	59.74	60.93	60.33
ResNet	✓		56.73	64.15	67.55	65.81
SVM		✓	55.51	63.82	64.24	64.03
MLP		✓	58.37	66.23	66.23	66.23
Perceiver	✓	✓	61.22	67.50	71.52	69.45
Liu et al.	✓	✓	59.18	65.09	72.85	68.75
STORK-A	✓	✓	62.45	69.03	70.86	69.93
Defusion	✓	✓	63.67	69.14	74.17	71.57
Ours	✓	✓	**65.71**	**71.08**	**76.77**	**73.08**

TABLE II
PERFORMANCE COMPARISON ON THE KROMP ET AL. BLASTOCYST DATASET.

Model	Modality		Performance Metrics (%)			
	Image	Table	Accuracy	Precision	Recall	F1 Score
EfficientNet	✓		58.06	51.12	58.10	54.38
ResNet	✓		61.20	54.47	60.00	57.10
SVM		✓	61.75	55.26	58.41	56.79
MLP		✓	63.25	56.89	60.32	58.55
Perceiver	✓	✓	64.89	59.12	59.68	59.40
Liu et al.	✓	✓	61.48	54.98	57.78	56.35
STORK-A	✓	✓	67.35	**61.66**	63.81	62.71
Defusion	✓	✓	66.80	60.47	66.03	63.13
Ours	✓	✓	**67.76**	61.38	**67.62**	**64.35**

the strong multimodal baselines DeFusion and STORK-A, by one to two percentage points across all evaluation metrics. The stable performance across both private and public datasets underscores the model's robust generalization capability. This robustness can be attributed to its dynamic convolution- and attention-based fusion design, which effectively adapts to domain variability and captures complementary modality-specific information. Overall, the results confirm that ResFTNet not only enhances predictive accuracy but also improves the consistency and interpretability of predicting clinical pregnancy outcomes in ART.

The visualization results in Fig. 2 illustrate the attention distributions of five multimodal models across representative test samples. Competing approaches, including Perceiver, Liu et al., STORK-A, and DeFusion, exhibit scattered or coarse activation areas that often fail to consistently align with morphologically critical regions, frequently drifting towards the zona pellucida or background noise. In contrast, ResFTNet produces clearer and more concentrated activation responses over biologically relevant zones. Specifically, the model successfully localizes the ICM and delineates the trophectoderm boundary, indicating a robust capability to identify key patterns associated with developmental potential. This improvement arises from the synergy between the ODConv and SEA attention modules, which together enhance both local structural perception and global semantic reasoning.

The interpretability of the clinical feature branch is further examined through Fig. 3, which presents the SHAP-based

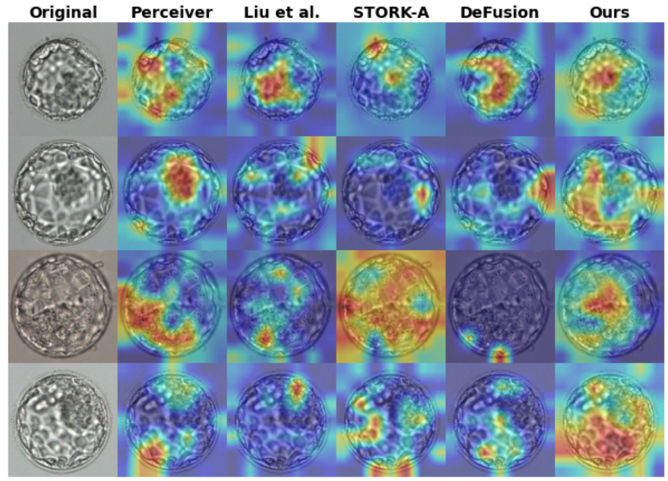

Fig. 2. Visual comparison of Grad-CAM heatmaps. The first two samples show that ResFTNet precisely localizes the ICM, whereas competing methods like STORK-A and DeFusion exhibit scattered focus. The last two samples demonstrate our model's ability to capture the trophectoderm boundary while suppressing background noise. Unlike other approaches that are often distracted by the zona pellucida, ResFTNet consistently highlights biologically significant features.

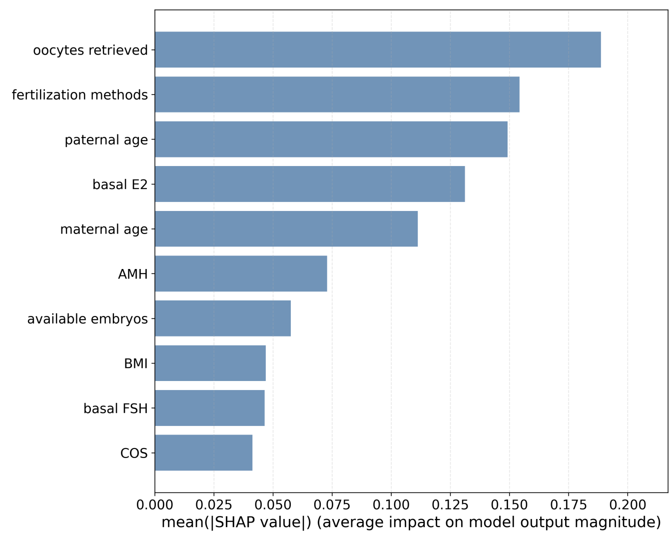

Fig. 3. SHAP summary plot of clinical tabular features on the private dataset.

feature importance ranking derived from the tabular modality. The top contributing variables include the total number of retrieved oocytes, fertilization methods, and paternal age. These indicators align with clinical knowledge in reproductive medicine, confirming that the model captures physiologically meaningful relationships between patient characteristics and pregnancy outcomes.

D. Ablation Study

To investigate the contribution of each component within the proposed framework, we conducted a series of ablation experiments on the CP-ART dataset, as summarized in Table III. We first evaluated the unimodal baselines. The image

TABLE III
ABLATION RESULTS OF RESFTNET COMPONENTS.

Component			Performance Metrics(%)			
Image	**Table**	**Module**	**Accuracy**	**Precision**	**Recall**	**F1 Score**
✓			56.73	64.15	67.55	65.81
	✓		60.82	66.67	72.85	69.62
✓	✓		63.27	68.48	74.83	71.52
✓	✓	✓	**65.71**	**71.08**	**76.77**	**73.08**

branch utilizing the ResNet backbone achieved an F1-score of 65.81%, while the tabular branch based on the FT-Transformer reached 69.62%. This indicates that patient-specific clinical information provides more direct predictive cues for pregnancy outcomes than embryo morphology alone. When these two modalities were combined through simple feature concatenation, the model performance further improved to an F1-score of 71.52%, demonstrating the inherent benefit of integrating complementary multimodal information. Finally, by incorporating the proposed ODConv and SEA attention modules into the image branch, the complete ResFTNet architecture achieved the best performance with an accuracy of 65.71% and an F1-score of 73.08%. This consistent improvement confirms that our dynamic feature enhancement design effectively strengthens the representation of embryo morphology, enabling more synergistic learning between visual and clinical features.

V. CONCLUSION

In this work, we proposed ResFTNet, an end-to-end multimodal framework for predicting clinical pregnancy outcomes. The architecture employs an FT-Transformer to capture complex clinical dependencies and an enhanced image encoder integrating ODConv and SEA attention to extract robust morphological features. To bridge the semantic gap between modalities, we introduced a clinical context injection (CCI) strategy that projects patient-specific clinical indicators into the visual feature space as a dynamic semantic bias, ensuring that embryo assessment aligns with the patient's physiological background. Experimental results on both private and public datasets demonstrate that ResFTNet consistently outperforms state-of-the-art methods in accuracy and generalization. Interpretability analyses further confirm that the model focuses on biologically meaningful cues, highlighting its potential as a trustworthy clinical decision-support tool. Future work will explore extending this framework to temporal embryo data and multi-center cohorts.

ACKNOWLEDGMENT

This study was funded by Key Scientific & Technological Project of XPCC (2025AB090).

REFERENCES

[1] P. K. Shah and J. M. Gher, "Human rights approaches to reducing infertility," *International Journal of Gynecology & Obstetrics*, vol. 162, no. 1, pp. 368–374, 2023.

[2] M. E. Graham, A. Jelin, A. H. Hoon Jr, A. M. Wilms Floet, E. Levey, and E. M. Graham, "Assisted reproductive technology: Short-and long-term outcomes," *Developmental Medicine & Child Neurology*, vol. 65, no. 1, pp. 38–49, 2023.

[3] L. Jin, K. Si, Z. Li, H. He, L. Wu, B. Ma, X. Ren, and B. Huang, "Multiple collapses of blastocysts after full blastocyst formation is an independent risk factor for aneuploidy—a study based on AI and manual validation," *Reproductive Biology and Endocrinology*, vol. 22, no. 1, p. 81, 2024.

[4] L. Chen, X. Tong, J. Zhang, X. Sun, Y. Wang, and Z. Yuan, "BlastocystMask: An instance segmentation of internal structure in human blastocyst images," in *International Conference on Intelligent Computing*. Springer, 2025, pp. 260–272.

[5] J. Ai, L. Jin, Y. Zheng, P. Yang, B. Huang, and X. Dong, "The morphology of inner cell mass is the strongest predictor of live birth after a frozen-thawed single embryo transfer," *Frontiers in Endocrinology*, vol. 12, p. 621221, 2021.

[6] J. Berntsen, J. Rimestad, J. T. Lassen, D. Tran, and M. F. Kragh, "Robust and generalizable embryo selection based on artificial intelligence and time-lapse image sequences," *PloS One*, vol. 17, no. 2, p. e0262661, 2022.

[7] J. Theilgaard Lassen, M. Fly Kragh, J. Rimestad, M. Nygård Johansen, and J. Berntsen, "Development and validation of deep learning based embryo selection across multiple days of transfer," *Scientific Reports*, vol. 13, no. 1, p. 4235, 2023.

[8] P. Vogiatzi, A. Pouliakis, and C. Siristatidis, "An artificial neural network for the prediction of assisted reproduction outcome," *Journal of Assisted Reproduction and Genetics*, vol. 36, no. 7, pp. 1441–1448, 2019.

[9] L. Liu, H. Liang, J. Yang, F. Shen, J. Chen, and L. Ao, "Clinical data-based modeling of IVF live birth outcome and its application," *Reproductive Biology and Endocrinology*, vol. 22, no. 1, p. 76, 2024.

[10] T. Tian, F. Kong, R. Yang, X. Long, L. Chen, M. Li, Q. Li, Y. Hao, Y. He, Y. Zhang *et al.*, "A bayesian network model for prediction of low or failed fertilization in assisted reproductive technology based on a large clinical real-world data," *Reproductive Biology and Endocrinology*, vol. 21, no. 1, p. 8, 2023.

[11] J. L. Metello, C. Tomás, and P. Ferreira, "Can we predict the IVF/ICSI live birth rate?" *JBRA Assisted Reproduction*, vol. 23, no. 4, pp. 402–407, 2019.

[12] X. Liang, J. He, L. He, Y. Lin, Y. Li, K. Cai, J. Wei, Y. Lu, and Z. Chen, "An ultrasound-based deep learning radiomic model combined with clinical data to predict clinical pregnancy after frozen embryo transfer: a pilot cohort study," *Reproductive BioMedicine Online*, vol. 47, no. 2, p. 103204, 2023.

[13] Y. Zou, Y. Pan, N. Ge, Y. Xu, R. Gu, Z. Li, J. Fu, J. Gao, X. Sun, and Y. Sun, "Can the combination of time-lapse parameters and clinical features predict embryonic ploidy status or implantation?" *Reproductive BioMedicine Online*, vol. 45, no. 4, pp. 643–651, 2022.

[14] J. Barnes, M. Brendel, V. R. Gao, S. Rajendran, J. Kim, Q. Li, J. E. Malmsten, J. T. Sierra, P. Zisimopoulos, A. Sigaras *et al.*, "A non-invasive artificial intelligence approach for the prediction of human blastocyst ploidy: a retrospective model development and validation study," *The Lancet Digital Health*, vol. 5, no. 1, pp. e28–e40, 2023.

[15] X. Ouyang, J. Wei, W. Huo, X. Wang, R. Li, and J. Zhou, "Defusion: An effective decoupling fusion network for multi-modal pregnancy prediction," *arXiv preprint arXiv:2501.04353*, 2025.

[16] C. Li, A. Zhou, and A. Yao, "Omni-dimensional dynamic convolution," *arXiv preprint arXiv:2209.07947*, 2022.

[17] Q. Wan, Z. Huang, J. Lu, G. Yu, and L. Zhang, "SeaFormer++: Squeeze-enhanced axial transformer for mobile visual recognition," *International Journal of Computer Vision*, vol. 133, no. 6, pp. 3645–3666, 2025.

[18] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 18 932–18 943, 2021.

[19] F. Kromp, R. Wagner, B. Balaban, V. Cottin, I. Cuevas-Saiz, C. Schachner, P. Fancsovits, M. Fawzy, L. Fischer, N. Findikli *et al.*, "An annotated human blastocyst dataset to benchmark deep learning architectures for in vitro fertilization," *Scientific Data*, vol. 10, no. 1, p. 271, 2023.

[20] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning (ICML)*. PMLR, 2019, pp. 6105–6114.