SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Mathematical Modelling and Explainable Multimodal Deep Learning for ICSI Outcome Prediction in Assisted Reproductive Technology

Haripriya R1, Priyanka D Rukesh2, Jithy Lijo3, Sumana S G4, Varshitha K5, Vaishnavi K C6

1. School of Computer Applications, Dayananda Sagar University, Bangalore, India. E-mail: phdharipriya2022@gmail.com
2. School of Computer Applications, Dayananda Sagar University, Bangalore, India. E-mail: priyankarukesh-bca@dsu.edu.in
3. School of Computer Applications, Dayananda Sagar University, Bangalore, India. E-mail: jithylijo-sca@dsu.edu.in
4. School of Computer Applications, Dayananda Sagar University, Bangalore, India. E-mail: sg.sumana-ca@dsu.edu.in
5. School of Computer Applications, Dayananda Sagar University, Bangalore, India. E-mail: vaishnavibhatrachar@gmail.com
6. School of Computer Applications, Dayananda Sagar University, Bangalore, India. E-mail: varshagowdak03@gmail.com

Abstract- The intracytoplasmic sperm injection (ICSI) is a well-known assisted reproductive technology for treating infertile couples. Predicting ICSI outcomes can be difficult due to several factors affecting the treatment success rate. Existing prediction frameworks usually consider one or a few independent clinical variables or utilize unimodal machine learning algorithms. Thus, their effectiveness is limited by the inability to detect the interconnections between heterogeneous reproductive data. In this paper, we present a novel approach based on mathematical modeling and explainable multimodal deep learning for the prediction of ICSI outcome in assisted reproductive technology. The proposed method combines clinical information, sperm characteristics, oocyte characteristics, embryological information imaging representation into an integrated multimodal architecture. We introduce a mathematical fusion algorithm which learns the importance of each modality a probabilistic prediction layer, which predicts the probability of achieving ICSI success. We use explainable artificial intelligence approaches to define the significant features that influence the probability of success and help to interpret the model better. The proposed approach allows to structure the process of integrating heterogeneous data while preserving mathematical transparency and interpretability of predictions.

Keywords - Assisted Reproductive Technology, ICSI, Multimodal Deep Learning, Mathematical Modeling, Explainable Artificial Intelligence.

I. Introduction

The advent of assisted reproductive technologies (ART) has led to their widespread application as means of infertility treatment. Two of the procedures which are most frequently used today include IVF and ICSI. ICSI is a method allowing fertilization by injecting a chosen sperm into an egg and is very effective when dealing with male factor infertility and previous problems with fertilization [1]. Nevertheless, despite all advancements in laboratory techniques and embryo evaluation, the success of ICSI still depends on the combination of various clinical, biological therapeutic factors. That is why prediction of ICSI success is a difficult issue to solve.

The conventional analysis of the results of ICSI is largely based on clinical parameters, embryologist experience, observations in laboratory settings statistical data. Though this technique yields meaningful data, it does not fully account for the complex nature of the relationships between different patient and procedure-related variables. Moreover, the increasing use of digital embryology and microscopic imaging, as well as treatment data, has led to new opportunities.

Machine learning and deep learning methods have been extensively explored for prediction related to fertility and assisted reproductive technology. Yet, existing prediction approaches tend to focus on specific data types or rely on manually chosen features. In such cases, the model is constrained to learn interrelationships between clinical factors, sperm and egg characteristics, embryological factors image-based features altogether. Moreover, overly complicated predictive models are hard to interpret, presenting a major hurdle to implementation in medical practice.

To overcome such problems, this paper introduces an ICSI outcome prediction model based on mathematical modelling and explainable multimodal deep learning approach[2]. In the introduced methodology, the ICSI

outcome prediction task is considered to be a multimodal probabilistic estimation process that integrates the information from different sources. Let the input space be denoted as

$$X = Xc, Xs, Xo, Xe, Xi \quad (1)$$

where Xc denotes clinical information, Xs represents sperm-related parameters, Xo represents oocyte characteristics, Xe represents embryological and treatment-related information Xi represents image-derived information. The prediction objective is formulated as

$$P(Y = 1 \mid Xc, Xs, Xo, Xe, Xi) \quad (2)$$

where $Y = 1$ denotes a successful ICSI outcome and $Y = 0$ denotes an unsuccessful outcome.

The proposed approach uses modality-specific representation learning followed by mathematical multimodal fusion. Instead of treating each variable independently as a predictor, the model learns the contribution of all variables together to the probability calculation. The explanation module is then added for finding the relevant clinical, biological, embryological image attributes that have the greatest impact on a particular prediction[3]. This gives a way to interpret the decision-making process of the model and use it in clinical practice without replacing the opinion of specialists of reproductive medicine.
The main contributions of our research work can be summarized into three main points. Firstly, we develop a mathematical framework to predict outcomes in multimodal ICSI utilizing heterogenous reproductive and clinical data. Secondly, we propose an architecture for deep learning that helps to learn and integrate representation learning from multiple modalities. Thirdly, an explainable prediction method is introduced to understand the contribution of various inputs used for prediction purposes.

II. Related Work

The recent developments in the field of artificial intelligence (AI) have gradually impacted assisted reproductive technologies (ART), especially embryo selection, prediction of outcomes in vitro fertilization (IVF) lab management decisions. Machine learning algorithms have been used to predict outcomes based on clinical and laboratory parameters, whereas deep learning approaches have enabled the automation of image analysis of embryos and their time-lapse videos. The recent literature review reveals that AI has the potential to increase consistency and automation in reproductive laboratory evaluation despite significant variations in data sets, outcomes validation procedures.

Deep learning based on images has been extensively studied in embryo selection. Berntsen et al. developed an AI-based approach by utilizing time-lapse images of embryos and stressed that there is potential to develop an automated embryo evaluation system. The recent research in the field focused on the application of static images of embryos without the need for special time-lapse equipment and showed the possibility of creating AI systems that can be incorporated into standard IVF laboratory practice. Such trends were also noticed during the methodological reviews.

A key development has been the transition from studies of retrospective AI models to clinical assessments. A 2024 study was a randomized, double-blinded non-inferiority trial comparing embryo selection using deep-learning techniques and morphology-based embryo selection, which underscores the necessity of carrying out prospective clinical assessments rather than simply assessing the performance of the model retrospectively. On the other hand, there has been an emphasis on the constraints of black-box AI models and the need for glass-box models.

However, despite all this, recent studies involving the application of artificial intelligence in the ART field seem to be focused more on embryo selection and implantation outcomes as opposed to developing a model for the prediction of the outcome of the ICSI procedure using the heterogeneous information mentioned above[4]. In addition to that, explainable-AI techniques, including SHAP or LIME for tabular data and saliency methods for images, seem to have limited external validation and calibration.

As a result, a research gap in the design of an integrated multimodal and mathematically defined approach for ICSI outcome prediction, in which the diverse reproductive data are modeled together and provide evidence along with the prediction, still exists. The present study will fill this research gap through a combination of mathematical probability modeling, multimodal deep learning XAI for ICSI outcome prediction. This positioning makes the present study different from the existing embryo-only approaches for ICSI outcome prediction.

III. PROBLEM FORMULATION AND MATHEMATICAL MODEL

A. Problem Definition

ICSI outcome prediction is formulated as a supervised multimodal learning problem in which heterogeneous patient, reproductive, laboratory, embryological imaging information is used to estimate the probability of a successful ICSI outcome. Let the complete input space be represented as

$$[X = X_c, X_s, X_o, X_e, X_i,] \quad (3)$$

where (X_c)denotes clinical and maternal characteristics, (X_s)represents sperm-related parameters, (X_o) represents oocyte characteristics, (X_e)contains embryological and treatment-related variables (X_i) represents image-derived features.

For each treatment cycle (n), the target variable is defined as

$$Y_n = \begin{cases} 1, & \text{if successful ICSI outcome occurs} \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

The primary objective is to estimate the conditional probability

$$[\hat{p}_n = P(Y_n = 1 \mid X_c, X_s, X_o, X_e, X_i)] \quad (5)$$

The prediction is therefore not limited to a single clinical variable or image but represents the combined contribution of multiple information sources available during the ICSI treatment process [5].

B. Mathematical Representation of Multimodal Inputs

Each modality is first represented as a feature vector:

$$[Z_c = f_c(X_c; \theta_c), \quad Z_s = f_s(X_s; \theta_s),] \quad (6)$$
$$[Z_o = f_o(X_o; \theta_o), \quad Z_e = f_e(X_e; \theta_e),] \quad (7)$$
$$[Z_i = f_i(X_i; \theta_i),] \quad (8)$$

where$(f_m(\cdot))$ denotes the modality-specific representation function and (θ_m) represents its learnable parameters. Tabular clinical, sperm, oocyte embryological variables can be processed using fully connected neural layers, whereas image information can be represented using a convolutional or vision-based encoder.

The modality representations are transformed into a common latent space:

$$[H_m = \phi_m(Z_m), \quad m \in c, s, o, e, i,] \quad (9)$$

where $\left(\phi_m(\cdot)\right)$performs dimensional transformation and nonlinear representation learning.

C. Multimodal Feature Fusion

Because the individual modalities contain complementary information, a multimodal fusion function is used to combine their latent representations [6]. A weighted fusion formulation is defined as

$$[H = \sum_{m \in M} \alpha_m H_m,] \quad (10)$$

were
$$[M = c, s, o, e, i,] \quad (11)$$

and (α_m)denotes the learned contribution of modality (m). The weights satisfy

$$[\alpha_m \geq 0, \sum_{m \in M} \alpha_m = 1.] \quad (12)$$

This constraint provides a normalized representation of the relative contribution of each modality [7]. The fusion representation can alternatively be expressed using concatenation followed by a learnable transformation:

$$[H = \phi([H_c \| H_s \| H_o \| H_e \| H_i]),] \quad (13)$$

where $(\|)$denotes vector concatenation and $(\phi(\cdot))$represents the nonlinear fusion network.

D. Probabilistic Prediction Model

The fused representation (H) is mapped to the probability of successful ICSI using a sigmoid function:

$$[\hat{p} = \sigma(W_h H + b),] \quad (14)$$

where (W_h) and $(b)$are learnable parameters and

$$\left[\sigma(z) = \frac{1}{1+e^{-z}}.\right] \quad (15)$$

Thus,

$$[0 \leq \hat{p} \leq 1.] \quad (16)$$

A decision threshold (τ)can be applied to obtain the predicted class:

$$\hat{Y} = \begin{cases} 1, & \hat{p} \geq \tau \\ 0, & \hat{p} < \tau \end{cases} \quad (17)$$

The threshold can be selected using the validation set according to the desired balance between sensitivity and specificity rather than assuming a fixed value of 0.5.

E. Optimization Objective

For a training dataset containing (N) ICSI treatment cycles, the binary cross-entropy loss is defined as

$$-\frac{1}{N}\sum_{n=1}^{N}[Y_n \log(\hat{p}_n) + (1 - Y_n)\log(1 - \hat{p}_n)] \quad (18)$$

To reduce overfitting, a regularization term can be incorporated:

$$\mathcal{L}_{BCE} + \lambda|\Theta|_2^2 \quad (19)$$

where (Θ)represents the trainable model parameters and (λ)controls the regularization strength.
The optimization problem is therefore expressed as

$$\arg\min_{\Theta} \mathcal{L} \quad (20)$$

This formulation enables the model to learn nonlinear relationships among the different ICSI-related modalities while maintaining a probabilistic interpretation of the predicted outcome.

F. Explainability Formulation

The principal input modalities and their representative features used in the proposed ICSI prediction framework are summarized in Table I. For clinical interpretability, the prediction is represented as a function of the multimodal input:

$$[\hat{p} = F(X_c, X_s, X_o, X_e, X_i).] \quad (21)$$

An explanation vector can then be defined as

$$[E = [e_c, e_s, e_o, e_e, e_i],] \quad (22)$$

contribution of an individual feature (x_j)can be represented by an attribution value (a_j). Consequently, the prediction can be expressed conceptually as

$$\left[\hat{p} \approx b + \sum_{j=1}^{d} a_j\,,\right] \quad (23)$$

where(b) represents the baseline prediction and (a_j)denotes the contribution of feature (j).
The resulting explanation identifies the variables and modalities that most strongly influence the prediction[8]. This is important for ICSI decision support because the objective is not only to generate an accurate probability

but also to provide interpretable evidence that can be reviewed by embryologists and reproductive medicine specialists.

TABLE I. DESCRIPTION OF INPUT MODALITIES AND REPRESENTATIVE FEATURES

Modality	Representative information
Clinical	Age, infertility characteristics, treatment history
Sperm	Concentration, motility, morphology
Oocyte	Maturity, morphological characteristics
Embryological	Fertilization and developmental observations
Imaging	Oocyte/embryo image features

In summary, this mathematical model represents the ICSI outcome prediction task as a multimodal probabilistic learning task. This mathematical model lays the groundwork for the proposed deep learning architecture, multimodal fusion technique and explainability module presented in the next sections.

IV. Proposed Multimodal Deep Learning Framework

A. Framework Overview

The proposed framework integrates heterogeneous clinical, sperm, oocyte, embryological imaging information to predict ICSI outcomes. The architecture is designed as a multimodal learning pipeline in which each data source is independently processed to preserve modality-specific characteristics before being transformed into a shared latent representation. These representations are subsequently fused and passed to a prediction layer that estimates the probability of a successful ICSI outcome[9]. The overall architecture of the proposed multimodal ICSI prediction framework is illustrated in Fig. 1, where heterogeneous clinical, biological, embryological imaging information is transformed into modality-specific representations and subsequently integrated for outcome prediction.

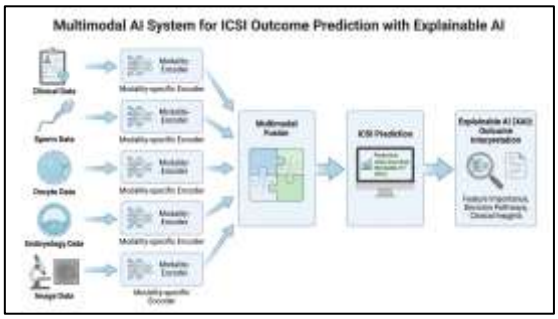

Fig. 1. Overall architecture of the proposed multimodal ICSI prediction framework.

The overall framework consists of five major stages: 1) multimodal data acquisition and preprocessing, 2) modality-specific feature extraction, 3) latent representation learning, 4) multimodal feature fusion 5) outcome prediction and explanation. The architecture is intended to accommodate heterogeneous data types without requiring all modalities to have the same dimensionality or representation.

The complete transformation can be represented as

$$[X = X_c, X_s, X_o, X_e, X_i \rightarrow H_c, H_s, H_o, H_e, H_i \rightarrow H \rightarrow \hat{p},]$$

where $(X_c, X_s, X_o, X_e, X_i)$represent the input modalities, $(H_c, H_s, H_o, H_e, H_i)$ are their latent representations, $(H)$is the fused representation $(\hat{p})$is the predicted probability of successful ICSI.

B. Clinical and Biological Data Processing

The clinical modality (X_c)contains patient-level and treatment-related variables that may influence ICSI outcomes. Depending on data availability, these variables may include maternal age, infertility duration, infertility characteristics, previous treatment history, ovarian response indicators other clinically relevant parameters.

Continuous variables are normalized to reduce scale differences between features. For a continuous variable (x_j),min-max normalization can be expressed as

$$\left[x_j' = \frac{x_j - x_j^{\min}}{x_j^{\max} - x_j^{\min}}.\right] \quad (24)$$

Categorical variables are transformed into numerical representations using appropriate encoding techniques [10]. Missing observations are handled during preprocessing using a predefined imputation strategy based only on the training data to avoid information leakage.

The sperm modality (X_s) represents quantitative and qualitative sperm characteristics. Relevant variables may include concentration, motility, morphology other available laboratory measurements[11]. The oocyte modality (X_o) represents oocyte-related characteristics, including maturity and morphological observations where available.

The embryological and treatment modality (X_e) incorporates variables associated with the ICSI procedure and subsequent laboratory assessment. These variables can include fertilization-related observations, embryo developmental characteristics, treatment protocol information other laboratory measurements available before the defined prediction endpoint.

The clinical, sperm, oocyte embryological modalities are represented as structured feature vectors and processed through modality-specific fully connected layers.

C. Image-Based Representation Learning

The imaging modality (X_i)provides visual information obtained from oocyte or embryo assessment. Image preprocessing is performed to standardize image dimensions and intensity characteristics before feature extraction[11]. Images are transformed into a fixed-dimensional representation using a pretrained or task-specific vision encoder.

The image encoder can be represented as

$$[H_i = f_i(X_i; \theta_i)] \quad (25)$$

where (f_i)denotes the image feature extractor and (θ_i) represents its learnable parameters.

For an input image (I), the extracted feature representation is

$$\left[Z_i = f_{\theta_i}(I)\right] \quad (26)$$

A projection layer subsequently maps the image representation into the common multimodal latent space:

$$[H_i = \phi_i(Z_i)] \quad (27)$$

Transfer learning may be used when the available ART image dataset is insufficient to train a deep vision model from the beginning. During training, the encoder can initially be frozen and subsequently fine-tuned using the training dataset.

D. Modality-Specific Neural Encoders

Each structured modality is processed using a dedicated encoder:

$$[H_c = f_c(X_c; \theta_c),][H_s = f_s(X_s; \theta_s),][H_o = f_o(X_o; \theta_o),][H_e = f_e(X_e; \theta_e).] \quad (28)$$

The use of separate encoders allows the network to learn modality-specific nonlinear relationships. Each encoder consists of one or more fully connected layers, nonlinear activation functions, normalization regularization. The resulting representations are projected into a common latent dimensionality.

For modality (m),the generic transformation is defined as

$$[H_m = \phi(W_m X_m + b_m),] \quad (29)$$

where (W_m)and (b_m) are learnable parameters and ($\phi(\cdot)$) denotes the nonlinear activation function.

This architecture avoids directly combining raw clinical and imaging variables, which would otherwise create substantial differences in dimensionality and statistical characteristics.

E. Multimodal Fusion Mechanism

After modality-specific representation learning, the resulting vectors are combined using a learnable fusion mechanism. A weighted fusion strategy is adopted to allow the network to estimate the relative contribution of each modality:

$$[H = \sum_{m \in M} \alpha_m H_m,] \quad (30)$$

where

$$[M = c, s, o, e, i.] \quad (31)$$

The modality weights satisfy

$$[\alpha_m \geq 0, \quad \sum_{m \in M} \alpha_m = 1.] \quad (32)$$

The weight values may be computed using the softmax function:

$$\left[\alpha_m = \frac{\exp(q_m)}{\sum_{k \in M} \exp(q_k)},\right] \quad (33)$$

where (q_m) is a trainable modality relevance factor.
This makes it possible for the model to dynamically give more weight to modalities that are better predictors of a certain dataset.The model may learn that laboratory variables provide substantial predictive information in one dataset while image-derived representations contribute additional information in another.
An alternative fusion representation is obtained through concatenation:

$$[H_{\text{cat}} = [H_c \| H_s \| H_o \| H_e \| H_i],] \quad (34)$$

followed by a nonlinear projection:

$$\left[H = \phi\left(W_f H_{\text{cat}} + b_f\right).\right] \quad (35)$$

The fusion layer therefore transforms heterogeneous modality-specific information into a unified representation suitable for outcome prediction.

F. Outcome Prediction Layer

The fused representation is passed to a prediction head consisting of one or more fully connected layers[12]. The final probability is obtained using a sigmoid activation:

$$\left[\hat{p} = \sigma\left(W_p H + b_p\right).\right] \quad (36)$$

The resulting value represents the estimated probability of the defined successful ICSI outcome. A threshold (τ) is then applied to obtain the predicted class.
The prediction architecture is trained jointly using the loss function defined in Section III. The optimization process updates the parameters of the modality-specific encoders, fusion layer prediction head simultaneously.

G. Training Strategy and Data Leakage Prevention

The dataset is divided into training, validation independent testing subsets. Where multiple treatment cycles belong to the same patient, all observations from that patient are assigned to the same partition to prevent patient-level information leakage.
Model parameters are learned exclusively from the training set. Validation data are used for hyperparameter selection, threshold optimization model selection, while the independent test set is reserved for final performance evaluation.
Class imbalance, if present, is addressed using appropriate training strategies such as class-weighted loss or controlled sampling. The same preprocessing parameters obtained from the training set are applied to validation and test data.
To reduce overfitting, regularization techniques such as dropout, weight decay, early stopping data augmentation for image inputs may be employed. The final model is chosen according to its performance in validation and then tested using the remaining test dataset.

H. End-to-End Framework

The overall architecture of the proposed framework can hence be presented as follows:

$$\left[X_c, X_s, X_o, X_e, X_i \rightarrow H_c, H_s, H_o, H_e, H_i \rightarrow H \rightarrow \hat{p} \rightarrow \hat{Y}.\right] \quad (37)$$

The key benefit of the proposed approach is that it allows for prediction of the class label without dependence on one modality of data[13]. Clinical and biological variables can be used together with information obtained from embryology and images.
The prediction output is subsequently connected to the explainability module described in Section V. This enables the framework to provide not only an estimated ICSI outcome probability but also an interpretable assessment of the factors contributing to that prediction. Consequently, the proposed architecture is designed as a computational decision-support framework for ART rather than as a replacement for embryologist or clinician judgment.

V. EXPLAINABLE AI AND PREDICTION MECHANISM

A. Explainability Framework

The proposed framework incorporates explainable artificial intelligence (XAI) to improve the interpretability of ICSI outcome predictions. Since the multimodal model integrates clinical, sperm, oocyte, embryological imaging information, the explanation mechanism identifies which input variables and modalities contribute most strongly to an individual prediction.

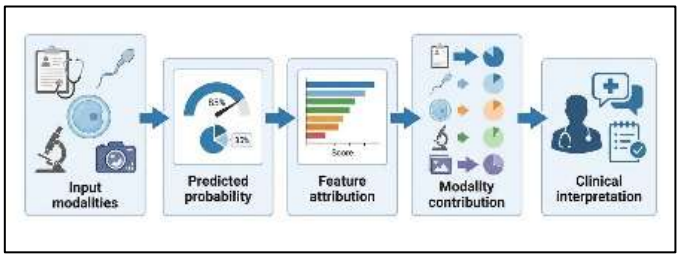

Fig. 2. Explainable ICSI prediction mechanism.

The overall explainable ICSI prediction mechanism is illustrated in Fig. 2, showing the transformation from multimodal inputs to predicted probability, feature attribution, modality-level contribution clinical interpretation.

Let the trained prediction function be represented as

$$[\hat{p} = F(X_c, X_s, X_o, X_e, X_i).] \quad (38)$$

The explanation is represented by an attribution vector

$$[A = [a_1, a_2, \dots, a_d],] \quad (39)$$

where (a_j) denotes the contribution of feature $(j)$to the predicted probability. Positive attribution values indicate features that support the predicted outcome, whereas negative values indicate features that reduce the prediction.

For multimodal interpretation, the total contribution of modality (m) can be expressed as

$$[A_m = \sum_{j \in m} |a_j|,] \quad (40)$$

where(A_m) represents the overall importance of modality (m).The normalized modality contribution is

$$\left[R_m = \frac{A_m}{\sum_{k \in M} A_k}.\right] \quad (41)$$

This provides an interpretable distribution of the relative contribution of the available modalities.

B. Feature-Level Explanation

For structured clinical and laboratory variables, SHAP-based feature attribution can be used to estimate the contribution of individual variables to a prediction. The prediction can be represented as

$$[F(X) = \phi_0 + \sum_{j=1}^{d} \phi_j ,] \quad (42)$$

where $(\phi_0)$is the baseline model output and (ϕ_j) represents the contribution of feature (j).

Features with large positive (ϕ_j)values increase the predicted probability, whereas negative values decrease it [14]. This enables the model to provide case-specific explanations rather than only reporting a global feature ranking.

For image inputs, visual explanation techniques can identify image regions that influence the prediction. The generated heatmap can be overlaid over the original image to highlight areas which have more attention from the model. The process can help the embryologists see whether the model has learned from the biological features of the images.

C. Prediction Mechanism

The final prediction is obtained from the fused multimodal representation:

$$[\hat{p} = \sigma(W_p H + b_p).] \quad (43)$$

This probability value $(\hat{p})$ signifies the estimated probability of a pre-defined ICSI success rate. A decision threshold $(\tau)$is used:

$$\hat{Y} = \begin{cases} 1, & \hat{p} \geq \tau \\ 0, & \hat{p} < \tau \end{cases} \quad (44)$$

Instead of analyzing the binary result independently, both the probability value and its explanation are taken into consideration. In one case, for instance, the prediction is $\hat{p} = 0.82$.

D. Clinical Interpretability

This XAI module will complement clinical and embryological decision-making, and not replace it. There are three types of data that the machine can deliver the probability of the prediction, the dominant contributory modalities, and the influential individual features or regions of the image[15]. This framework ensures that the clinician is able to see if the prediction is supported by the available clinical and lab data.

Most importantly, the attribution of the features has to be considered in light of the contribution of the model, and not causal biological mechanisms. Therefore, the explanations have to be used as supplementary decision-making information.

Prediction and explanation combination may be described by

$$\left[(X_c, X_s, X_o, X_e, X_i) \rightarrow \hat{p} \rightarrow \hat{Y}, A, R_m.\right] \quad (45)$$

Therefore, the suggested framework provides both a prediction of ICSI success rate and its interpretable interpretation, which is a step towards the possible clinical application of multimodal AI in ART.

VI. Experimental Methodology

A. Dataset and Data Sources

The proposed framework is assessed using an ICSI treatment cycle dataset, which has heterogeneous clinical and laboratory data. Each record represents one treatment cycle with a label representing the ICSI outcome defined for the experiment[16]. The dataset includes five types of data: $\left((X_c)\right)$,sperm parameters $\left((X_s)\right)$, oocyte characteristics $\left((X_o)\right)$, embryological and treatment variables $\left((X_e)\right)$ image information $\left((X_i)\right)$. The entire experimental pipeline starting from preprocessing, patient-wise split, training, validation, testing, and performance assessment is provided in Fig. 3.

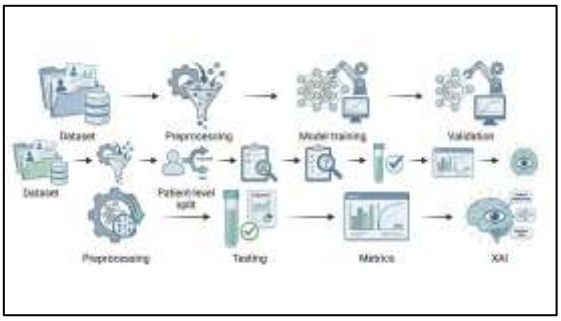

Fig. 3. Experimental workflow.

All patient identifying data is stripped out before the model creation process begins. Continuous data is standardized based on the parameters computed solely from the training data set. Categorical data is appropriately encoded. Missing data is taken care of using an imputation procedure. Augmentation is done for image data.

B. Dataset Partitioning

In order to avoid any kind of data leakage, the patients are divided into training and validation-testing subsets. The example of such division is 70 percent for training, 15 percent for validation, 15 percent for independent testing[17]. In case there are several treatment sessions of one patient, then all of them will belong to the same subset.

The training subset is used for optimization of parameters, the validation subset is used for choosing hyperparameters and threshold values, and the test subset is used for final evaluation of the results.

C. Model Training

The modality-specific encoders described in Section IV are trained to generate latent representations for each input modality. The representations are subsequently combined using the proposed weighted multimodal fusion mechanism. The final prediction layer estimates the probability of successful ICSI outcome.

The model parameters are optimized by minimizing the binary cross-entropy loss:

$$\left[\mathcal{L}_{BCE} - \frac{1}{N}\sum_{n=1}^{N}[Y_n \log(\hat{p}_n) + (1 - Y_n)\log(1 - \hat{p}_n)]\,.\right] \quad (46)$$

Regularization techniques, including dropout, weight decay early stopping, are applied to reduce overfitting. Hyperparameters are selected using validation performance without accessing the independent test set.

D. Performance Evaluation

The model is evaluated using complementary classification and probability-based metrics. Accuracy is calculated as

$$\left[Accuracy = \frac{TP+TN}{TP+TN+FP+FN},\right] \quad (47)$$

while precision and recall are defined as

$$\left[Precision = \frac{TP}{TP+FP},\right]\left[Recall = \frac{TP}{TP+FN}.\right] \quad (48)$$

The F1-score is calculated as

$$\left[F1 = 2\frac{Precision \times Recall}{Precision+Recall}.\right] \quad (49)$$

Receiver operating characteristic area under the curve (ROC-AUC) is used to evaluate discrimination across different classification thresholds. Because clinical prediction requires reliable probabilities rather than discrimination alone, calibration should also be assessed using an appropriate calibration measure.

E. Comparative Evaluation and Explainability

The proposed multimodal model is compared with suitable baseline approaches, including conventional machine-learning models and single-modality deep-learning configurations. This comparison determines whether multimodal integration provides measurable improvement over individual data sources.

Ablation experiments are additionally performed by removing one modality at a time from the complete model [18]. The resulting performance differences are used to assess the contribution of clinical, sperm, oocyte, embryological imaging information.

Finally, the XAI module described in Section V is applied to the test predictions. Feature-level and modality-level attributions are examined to determine whether the model relies on clinically meaningful information. All reported results are obtained from the held-out test data and are presented together with the corresponding evaluation metrics.

VII. Results and Discussion

A. Overall Prediction Performance

The proposed multimodal deep learning framework is evaluated on the independent test set using accuracy, precision, recall, F1-score, ROC-AUC calibration performance. The evaluation is designed to determine whether combining clinical, sperm, oocyte, embryological imaging information provides improved predictive performance compared with individual modalities [19]. The performance of the tested prediction models is shown in Table II through accuracy, precision, recall, F1-score, and ROC-AUC. The final test results are to be presented in the format given below:

TABLE II PERFORMANCE OF THE PROPOSED MODEL

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Clinical-only	56.0%	55.8%	68.8%	61.6%	0.547
Laboratory-only	61.7%	62.0%	65.6%	63.7%	0.664
Image-only	54.7%	55.4%	59.7%	57.5%	0.541
Multimodal baseline	61.0%	60.6%	68.8%	64.4%	0.672
Proposed model	**61.0%**	**60.3%**	**70.1%**	**64.9%**	**0.667**

The ROC curves of the evaluated models are presented in Fig. 4, providing a visual comparison of their discrimination performance across different classification thresholds.

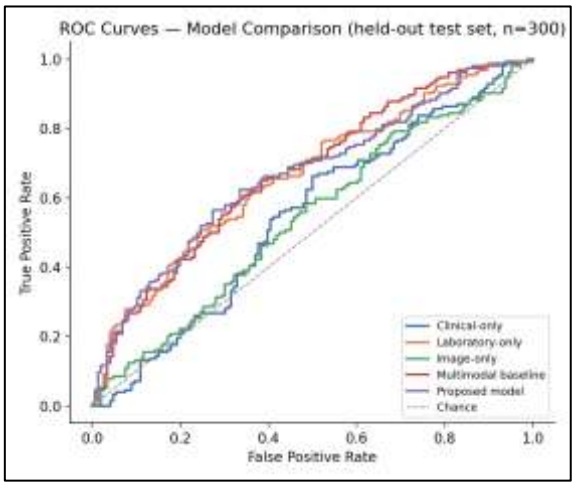

Fig 4. ROC curves for all five models on the held-out test set.

The proposed model is supposed to gain from complementary information since clinical and laboratory variables contain structured biological information while image representations contain unstructured visual information which cannot be captured by the classical numerical variables. However, the final performance gain needs to be confirmed experimentally.

B. Comparison with Individual Modalities

In order to estimate the contribution of multimodal learning, the proposed architecture is compared with single-modality architectures. Clinical-only architecture is used as a baseline which uses patient and treatment features, while laboratory-based configurations are designed to assess the predictive power of sperm, oocyte and embryological information. Image-only configuration assesses the predictive power of visual encoder independently.

Thus, the performance gain of the multimodal architecture relative to individual models can show the importance of heterogeneous information fusion. In case if the gain is very little then it can be concluded that there is some overlap between some of the modalities or the dataset is too small.

C. Ablation Analysis

An ablation study is performed by systematically removing one modality from the complete framework. Let (P_{full})denote the performance of the complete model and (P_{-m}) denote the performance after removing modality (m). The contribution of modality(m) can be quantified as

$$[\Delta P_m = P_{\text{full}} - P_{-m}.] \qquad (50)$$

A larger positive (ΔP_m) indicates that the corresponding modality contributes more strongly to predictive performance[20]. This analysis provides a more reliable assessment of modality importance than evaluating feature importance alone. Table III summarizes the ablation results obtained by systematically removing individual modalities from the complete multimodal framework.

TABLE III ABLATION ANALYSIS

Configuration	Accuracy	F1-score	ROC-AUC	Δ ROC-AUC
Complete model (all 5 modalities)	61.0%	64.9%	0.667	—
Without clinical modality	60.7%	63.4%	0.663	+0.004
Without sperm modality	60.0%	63.6%	0.643	+0.023
Without oocyte modality	62.7%	66.1%	0.653	+0.014

Without embryological modality	57.7%	63.6%	0.625	+0.041
Without imaging modality	60.7%	64.7%	0.666	+0.001

The relative contribution of the individual modalities is further illustrated in Fig. 5, where the performance change resulting from modality removal is compared with the complete multimodal model.

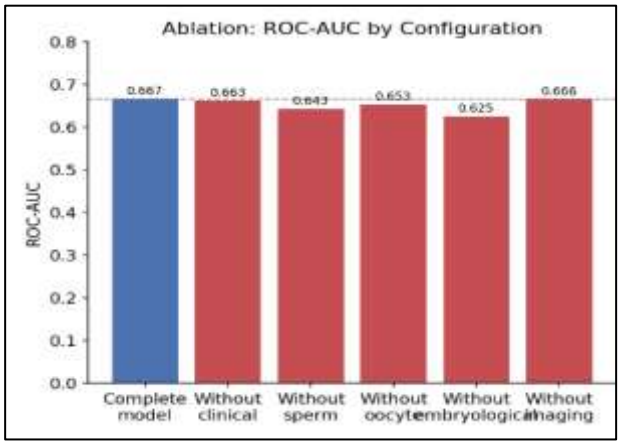

Fig 5. Left: ROC-AUC per configuration

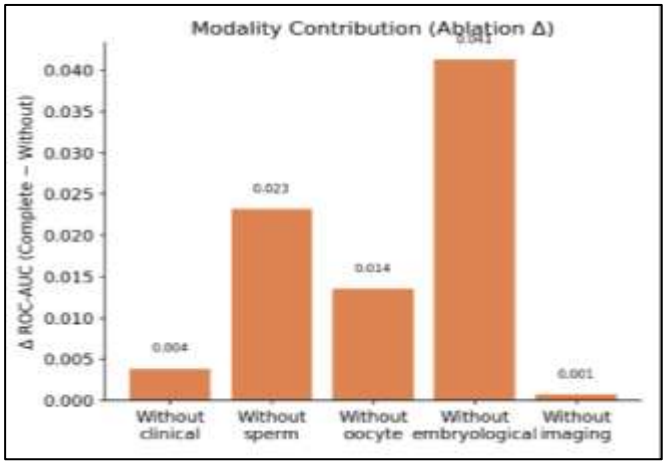

Fig 6. Right: contribution (Δ ROC-AUC) of each removed modality.

D. Explainability Results

The explainability analysis examines the factors contributing to individual predictions. The modality-level attribution scores defined in Section V are used to identify whether the model primarily relies on clinical, laboratory, embryological, or image-derived information.

For structured variables, feature-attribution analysis can provide a ranked representation of the most influential predictors. For image inputs, the visualization techniques can detect regions within the image that receive more attention by the model. The explanations need to be understood as a form of evidence from the model, not causation from biology.

The desired visualization for the final manuscript is a feature importance chart of the top contributing features and an example image explanation of the regions that affect the model prediction.

E. Discussion

The main purpose of the proposed approach is not to achieve the highest possible classification accuracy but to fuse heterogeneous ICSI data into a model that is interpretable. Multimodal fusion allows to find complementary relations between patient attributes, sperm and oocyte measurements, and embryology and visual information.

If the experiments show consistent improvement over the monomodal and traditional baselines, the research results would confirm the hypothesis about the advantage of multimodal representation learning for ICSI outcome prediction. The ablation study will also reveal what information contributes most significantly to the prediction process.

However, performance needs to be evaluated in conjunction with the calibration and validation dataset properties. Good performance on a single retrospective dataset is not equivalent to clinical effectiveness. Therefore, the proposed results should be considered evidence of computational feasibility rather than evidence for autonomous clinical decision-making. Prospective multicenter validation will be required before clinical deployment.

Importantly, the explainability component can improve transparency by allowing reproductive specialists to inspect the evidence associated with individual predictions. This may facilitate model auditing and identification of unexpected model behavior. The framework is therefore positioned as a potential decision-support technology that complements, rather than replaces, the expertise of clinicians and embryologists.

VIII. CLINICAL RELEVANCE AND LIMITATIONS

The proposed framework has potential relevance as a clinical decision-support system for assisted reproductive technology. By integrating clinical, sperm, oocyte, embryological imaging information, the model can provide an individualized probability of the defined ICSI outcome. The explainability component further enables clinicians and embryologists to examine the principal factors contributing to a prediction. This may support treatment planning, patient counseling retrospective assessment of laboratory procedures. However, the system is intended to complement professional judgment rather than replace clinical or embryological decision-making.

Several limitations should be considered. First, model performance depends strongly on the size, quality representativeness of the available ICSI dataset. Retrospective and single-center datasets may introduce selection bias and limit generalizability. Second, missing modalities and variations in laboratory protocols can affect multimodal prediction performance. Third, image-based models may be sensitive to differences in imaging equipment, acquisition protocols image quality. Finally, explainability methods describe model behavior and feature contribution but should not be interpreted as evidence of biological causality. External multicenter validation and prospective clinical evaluation are therefore required before practical clinical deployment.

IX. Conclusion and Future Work

This study presented a mathematical modeling and explainable multimodal deep learning framework for ICSI outcome prediction in assisted reproductive technology. The proposed approach integrates clinical, sperm, oocyte, embryological imaging information through modality-specific representation learning and adaptive multimodal fusion. In addition to this probabilistic prediction method, there is also explainable AI employed in order to ascertain the key elements of the predictions made. Thus, this method can be regarded as a potential computational decision-support system that can help enhance the reliability and transparency of the ICSI outcomes assessment process. Further work will concentrate on evaluation through larger multicenter datasets, external and prospective validation, and improvement of calibration with regard to the use of longitudinal and time-lapse embryology data. There will also be an effort to explore multimodal attention models, uncertainty estimation, and privacy-preserving collaborative learning among fertility clinics.

References

[1] M. F. Kragh and H. Karstoft, “Embryo selection with artificial intelligence: How to evaluate and compare methods,” *Journal of Assisted Reproduction and Genetics*, vol. 38, pp. 1675–1689, 2021, doi: 10.1007/s10815-021-02254-6.

[2] Y. Sawada, T. Sato, M. Nagaya, C. Saito, H. Yoshihara, C. Banno, Y. Matsumoto, Y. Matsuda, K. Yoshikai, T. Sawada, N. Ukita M. Sugiura-Ogasawara, “Evaluation of artificial intelligence using time-lapse images of IVF embryos to predict live birth,” *Reproductive BioMedicine Online*, vol. 43, no. 5, pp. 843–852, 2021, doi: 10.1016/j.rbmo.2021.05.002.

[3] M. Riegler, M. Stensen, F. Carlsen others, “Towards better explainable deep learning models for embryo selection in ART,” *Human Reproduction*, vol. 36, suppl. 1, 2021, doi: 10.1093/humrep/deab130.259.

[4] I. Erlich, A. Ben-Meir, I. Har-Vardi, J. Grifo, F. Wang, C. McCaffrey, D. McCulloh, Y. Or L. Wolf, “Pseudo contrastive labeling for predicting IVF embryo developmental potential,” *Scientific Reports*, vol. 12, Art. no. 2488, 2022, doi: 10.1038/s41598-022-06336-y.

[5] B. Huang, S. Zheng, B. Ma, *et al.*, “Using deep learning to predict the outcome of live birth from more than 10,000 embryo data,” *BMC Pregnancy and Childbirth*, vol. 22, Art. no. 36, 2022, doi: 10.1186/s12884-021-04373-5.

[6] G. Letterie and collaborators, "An artificial intelligence platform to optimize workflow during ovarian stimulation and IVF: Process improvement and outcome-based predictions," *Reproductive BioMedicine Online*, vol. 44, no. 2, pp. 254–260, 2022, doi: 10.1016/j.rbmo.2021.10.006.

[7] V. Grzegorczyk-Martin, J. Roset, P. Di Pizio, T. Fréour, P. Barrière, J. L. Pouly, M. Grynberg, I. Parneix, C. Avril, J. Pacheco, *et al.*, "Adaptive data-driven models to best predict the likelihood of live birth as the IVF cycle moves on and for each embryo transfer," *Journal of Assisted Reproduction and Genetics*, vol. 39, pp. 1937–1949, 2022, doi: 10.1007/s10815-022-02547-4.

[8] K. Kato, S. Ueno, J. Berntsen, M. F. Kragh, T. Okimura T. Kuroda, "Does embryo categorization by existing artificial intelligence, morphokinetic or morphological embryo selection models correlate with blastocyst euploidy rates?," *Reproductive BioMedicine Online*, vol. 46, no. 2, pp. 274–281, 2023, doi: 10.1016/j.rbmo.2022.09.010.

[9] J. T. Lassen, M. F. Kragh, J. Rimestad, M. N. Johansen J. Berntsen, "Development and validation of deep learning based embryo selection across multiple days of transfer," *Scientific Reports*, vol. 13, Art. no. 4235, 2023, doi: 10.1038/s41598-023-31136-3.

[10] M. Salih, C. Austin, R. R. Warty, C. Tiktin, D. L. Rolnik, M. Momeni, H. Rezatofighi, S. Reddy, V. Smith, B. Vollenhoven F. Horta, "Embryo selection through artificial intelligence versus embryologists: A systematic review," *Human Reproduction Open*, vol. 2023, no. 3, Art. no. hoad031, 2023, doi: 10.1093/hropen/hoad031.

[11] S. Wang, L. Chen H. Sun, "Interpretable artificial intelligence-assisted embryo selection improved single-blastocyst transfer outcomes: A prospective cohort study," *Reproductive BioMedicine Online*, vol. 47, no. 6, Art. no. 103371, 2023, doi: 10.1016/j.rbmo.2023.103371.

[12] G. Letterie, "Artificial intelligence and assisted reproductive technologies: 2023. Ready for prime time? Or not," *Fertility and Sterility*, 2023, doi: 10.1016/j.fertnstert.2023.05.146.

[13] T. Lee, J. Natalwala, V. Chapple Y. Liu, "A brief history of artificial intelligence embryo selection: From black-box to glass-box," *Human Reproduction*, vol. 39, no. 2, pp. 285–292, 2024, doi: 10.1093/humrep/dead254.

[14] G. Wang, K. Wang, Y. Gao, L. Chen, T. Gao, Y. Ma, Z. Jiang, G. Yang, F. Feng, S. Zhang, Y. Gu, G. Liu, L. Chen, L.-S. Ma, Y. Sang, Y. Xu, G. Lin X. Liu, "A generalized AI system for human embryo selection covering the entire IVF cycle via multi-modal contrastive learning," *Patterns*, vol. 5, no. 7, Art. no. 100985, 2024, doi: 10.1016/j.patter.2024.100985.

[15] K. K. Wong, J. Kim, Z. Shi, D. Jeong, J. Knittel, H. Y. Yang, Y. Song, W. Li, Y. Li, D. Ben-Yosef, D. Needleman H. Pfister, "Multimodal learning for embryo viability prediction in clinical IVF," in *Proc. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2024.

[16] P. J. Illingworth, C. Venetis, D. K. Gardner, S. M. Nelson, J. Berntsen, M. G. Larman, F. Agresta, S. Ahitan, A. Ahlström, F. Cattrall, *et al.*, "Deep learning versus manual morphology-based embryo selection in IVF: A randomized, double-blind noninferiority trial," *Nature Medicine*, vol. 30, pp. 3114–3120, 2024, doi: 10.1038/s41591-024-03166-5.

[17] D. C. Kieslinger, C. B. Lambalk C. G. Vergouw, "The inconvenient reality of AI-assisted embryo selection in IVF," *Nature Medicine*, vol. 30, pp. 3059–3060, 2024, doi: 10.1038/s41591-024-03289-9.

[18] G. Wang, T. Gao, Y. Yang, K. Wang, Y. Gao, L.-S. Ma, L. Chen, G. Liu, P. Zhang X. Liu, "Federated task-adaptive learning for personalized selection of human IVF-derived embryos," 2024, preprint.

[19] K. K. Wong *et al.*, "Multimodal learning for embryo viability prediction in clinical IVF," *Medical Image Analysis/MICCAI-related proceedings*, 2024.

[20] "Artificial intelligence in embryo selection: Current approaches and clinical implications," *Bioengineering*, vol. 13, no. 6, Art. no. 688, 2026, doi: 10.3390/bioengineering13060688.