Deep and Hybrid Ensemble Learning Methods for Enhanced Live-Birth Prediction in Fertility Treatments

Rituraj Jain[1], Uma Shankar[2], G V Radhakrishnan[3], Saroj Date[4], Kamal Upreti[*5], S. Caroline[6], Ramesh Babu P[7], Mohit Dayal[8]
[1]Department of Information Technology, Marwadi University, Rajkot, Gujarat, India
[2]Faculty of Management and Social Sciences, Qaiwan International University, Sulaimanyah, Kurdistan, Iraq
[3]Kalinga School of Management, Kalinga Institute of Industrial Technology, Bhubaneswar, India
[4]Department of Artificial Intelligence and Data Science, CSMSS Chh. Shahu college of Engineering, Chh. Sambhajinagar, Maharashtra, India
[5]Department of Computer Science, Christ University, Delhi NCR Campus, Ghaziabad, India
[6]Department of Electronics and Communication Engineering, St. Xavier's Catholic College of Engineering, Nagercoil, Tamil Nadu, India
[7]Department of Computer Science & Engineering, Narsimha Reddy Engineering College, Secunderabad, Telangana, India
[8]Department of Information Technology, Bharati Vidyapeeth's College of Engineering, New Delhi, India
E-mail: jainrituraj@yahoo.com, uskapilavgmail.com, vrkris2002@gmail.com, saroj.date@gmail.com, kamalupreti1989@gmail.com, caroline@sxcce.edu.in, drprb2009@gmail.com, mohitdayal.md@gmail.com
[*]Corresponding author

Keywords: assisted reproductive technologies, hybrid machine learning, IVF prediction, live-birth estimation, predictive modelling

Received: March 6, 2025

The prediction of live birth outcomes using Assisted Reproductive Technologies (ART) remains a complex task owing to the high inter-patient variability and non-linear clinical interactions. This study presents a comparative evaluation of hybrid machine-learning models to improve in vitro fertilization (IVF) success prediction using a real-world anonymized dataset of 2,000 ART cases. After pre-processing (including missing value imputation, feature selection via Recursive Feature Elimination with Cross-Validation, and class balancing using SMOTE with k=5), four hybrid models were developed: stacking with XGBoost as the meta-learner, weighted ensemble, autoencoder-based feature fusion, and cascading classifiers. Models were evaluated using accuracy, AUC, precision, recall, and F1-score metrics, and compared against a baseline Random Forest classifier. The stacking model (XGBoost with Random Forest, MLP, and SVM base learners) achieved the best performance, with an accuracy and 0.999 AUC of 0.985. The weighted hybrid ensemble followed an accuracy of 0.953 and AUC of 0.994. The statistical significance of the improvements was confirmed using Wilcoxon Signed-Rank and McNemar's tests ($p < 0.05$). To enhance model transparency, SHapley Additive exPlanations (SHAP) was applied to interpret base model contributions in the stacking architecture. These results support the application of AI-driven hybrid modelling for personalized IVF treatment planning. Future work will focus on prospective validation and clinical decision support system (CDSS) integration to assess deployment feasibility.

Povzetek: Študija na 2000 primerih ART primerja hibridne modele za napoved rojstva živorojenega. Po obdelavi (RFECV, SMOTE) najboljši sistem z XGBoostom doseže najboljše rezultate. SHAP zagotovi razložljivost; Wilcoxon/McNemar potrdita izboljšave. Predvidena validacija in vključitev v klinični CDSS sistem.

1 Introduction

In ART, it is crucial to develop predictive models to forecast live birth outcomes after IVF. However, even in the era of reproductive medicine and embryology, IVF procedures have not been able to realize larger, more effective success rates, which depend on patient-specific factors, including age, hormonal levels, and the quality of embryos [1], [2]Previous attempts have employed traditional statistical models to estimate IVF success probabilities, which are, unfortunately, never overly accurate since biological and clinical factors are complex and their effects are still heterogeneous [3], [4]. Machine Learning (ML) and Deep Learning (DL) models have recently emerged as powerful tools for enhancement of predictive abilities with reproductive medicine [5], [6]. Nevertheless, standalone models often suffer from overfitting, data bias, and low generalizability [7]. To address such challenges, hybrid machine learning solutions can offer a worthy possibility through the combination of several algorithms to exploit the positive aspects of multiple models and increase prediction accuracy and robustness [8], [9]. Successful IVF treatment

relies on a range of variables including patient characteristics, clinical indicators, and embryo developmental aspects [2], [10], [11], [12], [13]. Traditionally implemented predictive models have limited generalizability because IVF-related data contain high dimensions and non-linear patterns [14], [15], [16]. The main reason for adopting hybrid models is the six core difficulties.

1. The complex connections between IVF data elements, including patient information, hormone patterns, and embryonic results, remain inaccessible to the standard statistical techniques for interpretation.
2. Many ART predictive features display non-linear behavior with specific effects on different patient populations, thus creating obstacles for developing a unified generalized model.
3. Class Imbalance exists because successful live births occur less frequently than unsuccessful cases, thus causing an unbalanced dataset that affects the model performance.
4. Medical staff need interpretable predictive models to demonstrate their analytical reasoning, rather than forcing them to use machine learning methods as unexplainable systems.
5. Multiple sensory data sources influence IVF success rates through clinical records combined with patient imaging results and time-based medical histories, which require sophisticated data-fusion systems to extract involvement information.
6. There is a lack of interpretable, personalized insights for supporting clinical decision-making in IVF treatments.

This study presents and evaluates five novel cross-hybrid models for predicting IVF live birth outcomes, further refining the specificity of prediction through multiple machine learning paradigms. The hybrid approaches used in this study included stacking (Layered Learning), weighted hybrid ensembles, cascading models, feature-level fusion with autoencoder networks, and SHAP-enhanced gradient boosting. These approaches were devised to handle data heterogeneity, temporal dependencies, feature importance selection, and automated hyperparameter optimization, which are highly relevant to IVF outcome prediction. This study aimed to present a robust and scalable predictive framework for these hybrid approaches to support clinical decision-making and personalized treatment planning in Assisted Reproductive Technology (ART).

A hybrid modelling approach provides a synergistic blending of several machine-learning models to overcome the weaknesses of individual models, resulting in increased predictive accuracy and robustness [17], [18], [19]. This study also introduces SHAP-based interpretability to align predictive modelling with clinical transparency, making it easier for fertility specialists to interpret model outputs.

This study aimed to bridge the gap between computational intelligence and clinical decision-making in reproductive medicine, improve the prediction of IVF success, and aid fertility specialists in designing optimal treatment plans tailored to individual patients.

1.1 Novelty and contribution of proposed system

This paper proposes a novel hybrid framework that incorporates stacking, weighted ensembles, cascading classifiers, autoencoder-based feature fusion, and SHAP-enhanced XGBoost explainability. Compared to existing methods, the proposed system employs multi-stage learning to address class imbalance and enhances decision reliability through trust refinement techniques such as RFECV-based feature selection and ensemble diversity. Table 1 summarizes the existing research gaps and explains how the proposed models were designed to overcome them through synergistic learning strategies and explainable outputs.

1.2 Structure of paper

The remainder of this paper is structured as follows: Section 2 presents a comprehensive literature survey of prior works on prediction models of IVF using machine learning applications in ART. Section 3 explains the methodology, dataset used, pre-processing, and a detailed description of the four hybrid modelling strategies used in this study. Section 4 presents the results and a comparison of the models. Section 5 summarizes the main findings and concludes the study.

Table 1: Research questions, identified research gaps and research contribution

Research Questions	Identified Research Gaps	Research Contributions
RQ1: How can predictive models generalize better for complex IVF data?	Existing models struggle with generalizing across diverse patient data and clinical variability.	Hybrid models integrate multiple learning approaches to improve generalization and robustness.
RQ2: How can feature selection be optimized for IVF prediction?	Conventional methods may overlook critical IVF-specific features or suffer from overfitting.	Hybrid models use RFECV, Bayesian Optimization, and autoencoders to refine feature selection.
RQ3: How can we handle class imbalances in IVF datasets?	Standard classifiers are biased towards majority classes, reducing predictive accuracy for minority cases.	SMOTE, weighting mechanisms, and cascading models dynamically handle class imbalance.
RQ4: How can we quantify the uncertainty in IVF predictions?	Traditional models lack interpretability and fail to quantify prediction uncertainty.	Hybrid ensemble methods provide confidence estimates, whereas cascading models refine the uncertain predictions.
RQ5: How can deep feature representations be utilized for improved prediction?	Feature extraction techniques do not effectively capture deep, non-linear IVF patterns.	Autoencoder fusion and hybrid deep learning methods enhance the feature representations.
RQ6: How can hyperparameter tuning be	Manual or grid search-based tuning is inefficient and	Hybrid models integrate multiple learning approaches to

Research Questions	Identified Research Gaps	Research Contributions
improved to achieve better performance?	computationally expensive.	improve generalization and robustness.

2 Literature review

Models of increasing machine learning have been increasingly used to predict IVF outcomes, with the possibility of providing the benefits of predicting IVF outcomes for infertile couples and in healthcare systems. These models often use a given set of data to predict the IVF outcomes. Existing machine learning-based approaches for IVF outcome prediction generally involve the use of medical and reproductive history, biochemical indicators, and data regarding reproductive tract examination, along with information from previous IVF cycles [13]. These models have proven useful for assessing subfertile couples and providing clues for treatment. Nevertheless, they are constrained by their dependency on conventional clinical parameters and do not reveal the factors that affect IVF success.

Consequently, scientists are developing increasingly sophisticated applications that integrate omics data with AI. The recent introduction of metabolomics, transcriptomics, and biomarkers in conjunction with deep machine learning assessment of oocytes, sperm, and embryos has been proposed as a novel tool [13]. The proposed algorithm provides a way to develop artificial neural network models that can better objectively and accurately predict this outcome than the traditional methods used in couples with unexplained infertility or repeated implantation failures. Several earlier techniques have been used for multiple-attribute selection methods to predict outcomes more accurately and efficiently through IVF prediction.

In one example, researchers integrated omics and artificial intelligence evidence to suggest the best treatment options and increase IVF success rates; thus, they developed a novel tool [13]. First, the lifestyle and demographic parameters of the subfertile couples, metabolomics, transcriptomics, and biomarkers were obtained, and the oocytes, sperm, and embryos were evaluated using deep machine learning. This study also emphasizes the value of omics data in facilitating optimal embryo selection and improving personalized IVF treatment.

Similarly, a separate study used the XGBoost machine learning system [20] to minimize multiple embryo gestation rates in IVF by creating a hierarchical model. It concomitantly learns embryo implantation potential and double embryo transfer. The variables identified by the researchers for single-embryo transfer pregnancies were age, IVF attempts, estradiol level on hCG day, and endometrial thickness. For double embryo transfer, the other variables, including P1 and P2, were significant. For SET pregnancy, DET pregnancy, and DET twin risks, the model exhibited AUC of 0.7945, 0.8385, and 0.7229, respectively.

Issues related to data quality and feature selection are important in predicting IVF outcomes. It is also in the context of challenges in learning deep features and extracting high-level patterns [21]. Feature selection is important because it can select redundant and irrelevant features to remove dimensionality and enhance model generalization [22]. Specifically, regarding IVF, the integration of omics data (metabolomics and transcriptomics) with classic clinical parameters has both advantages and disadvantages when applied to feature selection [13].

However, the interpretability of the model remains challenging. Machine learning methods are capable of modelling flexibility and robustness; however, it is difficult to interpret individual features using sophisticated algorithms [23]. One problem with this lack of interpretability is that it hinders the identification of important biomarkers needed to develop novel hypotheses for the prevention, diagnosis, and treatment of complex conditions such as infertility.

Interestingly, the choice of modelling approach and feature selection method strongly depends on the purpose of the analysis. In [24], it was clearly recommended that the goal of model selection be specified as data exploration, inference, or prediction, as it serves the purpose of selecting the appropriate model to ensure that there is no confusion when selecting a statistical model.

However, demographic and clinical factors determine the IVF outcomes. However, important predictors for both single and double embryo transfer pregnancies include age, number of previous IVF attempts, estradiol level on hCG day, and endometrial thickness [20]. The live birth and implantation rates for women aged 35 years or younger with a caesarean section defect were significantly lower than those for women with a history of vaginal delivery [25].

Among the causes of low intrauterine insemination success, semen parameters (sperm concentration and motility) and female body mass index (BMI) were identified as the most important predictors [26]. Nevertheless, a meta-analysis of semen quality (concentration, motility, and morphology) and outcomes of assisted reproduction technologies [27] could not determine a significant correlation between these two variables. This contradiction indicates complications in predicting IVF success.

Environmental factors have a bearing on IVF outcomes. Fresh embryo transfer (FET) cycles result in lower chances of biochemical pregnancy, clinical pregnancy, and live birth during exposure to air pollutants, particularly ozone (O3), nitrogen dioxide (NO2), and carbon monoxide (CO), at various stages of IVF treatment [28].

The Area Under the Curve (AUC) of a Receiver Operating Characteristic (ROC) curve is commonly used. The AUC for live birth prediction was 0.905, and that for clinical pregnancy with fetal heartbeat was 0.722 [1]. The AUCs reported for single embryo transfer (SET), DET pregnancy, and DET twin risks were 0.7945, 0.8385, and 0.7229, respectively [20]. AUCs from 0.70 to 0.78 were obtained for ploidy prediction [29].

Other metrics included the accuracy, precision, recall, and F1 scores. Zou et al. achieved an accuracy of 0.77, precision of 0.79, recall of 0.86, and F1 score of 0.83 [29]. Values reported by [1] for live births and clinical pregnancies with fetal heartbeat of 1.12 and 0.77, respectively, were also used in the Observed: Expected (O: E) ratio.

Some studies have compared AI models with human predictions. A meta-analysis was not possible, but a systematic review indicated that AI-based prediction models were as good as embryologists, albeit marginally better [1]. It is also said that AI models have not yet surpassed clinically embryologists' predictive capability significantly.

Furthermore, one study [13] documented the lifestyle and demographic parameters of subfertile couples, together with the previous characteristics of IVF cycles. In addition, they measured and evaluated metabolomics, transcriptomics, and biomarkers by evaluating oocytes, sperm, and embryos using deep machine learning. This bundling of data collection and pre-processing was comprehensive enough to create artificial neural network models to increase the objectivity and accuracy of IVF success rate predictions.

In fact, some studies have investigated only a particular pre-processing technique, whereas others have focused on the need to establish the best pre-processing pipeline to follow before prediction. For example, [30] used an automated pre-processing model, referred to as a scenario-based model, in their study of construction accident severity prediction.

Compared to existing models, such as the protocol-based ANN framework by [13] and the XGBoost-based hierarchical model by [20], the proposed hybrid system significantly advances IVF outcome prediction. While prior studies lacked either empirical validation or comprehensive data integration, our model combined stacking, weighted ensembles, autoencoder-based fusion, and cascading strategies. This multi-stage approach addresses critical gaps such as limited data types, lack of personalization, and poor generalization, achieving superior accuracy (0.985) and AUC (0.999). Furthermore, statistical tests confirmed the model's significant improvement over traditional methods, establishing a robust and scalable predictive framework for clinical use.

To execute this model, several pre-processing steps are reviewed, including the processing of missing data, binned data, outlying data, scaling methods, and resampling data. The pre-processing pipeline plays a vital role, and we observed that in the most efficient scenario, we obtained the best out-of-prediction performance.

3 Methodology

3.1 Dataset

The dataset analyzed in this study was derived from anonymized registry data compiled by the Human Fertilization and Embryology Authority (HFEA), covering fertility treatments conducted between 2010 and 2016 [31]. It is publicly available and can be accessed via the direct download link provided in [31]. This dataset was previously used in a study by Goyal et al. [32]. It is the world's largest database of patient, donor, and offspring records that safeguards patient, donor, and offspring confidentiality to support patient care improvements. The dataset was filtered to include 2,000 samples with 95 attributes from actual in vitro fertilization (IVF) procedures. These attributes include age, infertility type, clinical treatment details, and embryonic data. Data were presented as numerical, categorical, or text. To ensure the same quality and consistency of the dataset, the missing values were considered through proper imputation. The numerical attributes were replaced with their median values, whereas the categorical attributes were set to their mode values. The dataset used did not contain attributes with more than 50% of the missing values. In total, 62 features were available after pre-processing and were chosen for further analysis.

The probability of a successful live birth with ART is a complex problem because it contains heterogeneous clinical, embryonic, and demographic factors. Therefore, we present a Hybrid Machine Learning Framework to improve the prediction accuracy and robustness by incorporating advanced data pre-processing techniques, utilizing a suite of machine learning models, and employing ensemble learning strategies. The IVF live birth prediction pipeline, consisting of data processing and the hybrid model development workflow, is illustrated in Figure 1. The proposed framework comprises four phases: data pre-processing, hybrid model development, model evaluation and comparison, and statistical significance testing. A detailed stepwise procedure of this Hybrid Machine Learning Framework for IVF Live Birth Prediction Algorithm has been systematically documented and is presented below to standardize a structured methodological approach to attain reproducibility.

Algorithm: Hybrid Machine Learning Framework for IVF Live Birth Prediction

Input: D = {X, Y}, where X represents patient demographics, clinical treatment details, and embryonic development data, and Y is the binary target variable indicating live birth outcomes.

Output: Predicted probability $\hat{Y}$ of live birth outcome.

Step 1: Data Preprocessing

1. Load dataset D.

2. Handle missing values:

a. Remove features F_i where $|F_i\text{missing}| / |F_i| > 0.5$.

b. Impute missing numerical values using the median: $Xnum(i) \leftarrow \text{median}(Xnum)$.

c. Impute missing categorical values using mode: $Xcat(i) \leftarrow \text{mode}(Xcat)$.

3. Encodes categorical variables using one-hot encoding.

4. Convert target variable: No live birth (Y = 0) and At least one live birth (Y = 1).

5. Perform stratified sampling: Split dataset into training and testing

6. Apply Synthetic Minority Over-Sampling Technique (SMOTE) to balance class distribution.

7. Perform feature selection using Recursive Feature Elimination with Cross-Validation (RFECV).

Step 2: Hybrid Model Development: Train machine learning models:

A. Stacking with Meta-Learners:

a. Train base models: Random Forest (RF), Support Vector Machine (SVM), Multi-Layer Perceptron (MLP).

b. Generate meta-features from out-of-fold predictions of base models using k-fold cross-validation to avoid data leakage

c. Use predictions $\hat{Y}base$ as input features for meta-learners: XGBoost, MLP, Gradient Boosting Machine (GBM).

B. Weighted Hybrid Ensemble:

a. Train models: RF, MLP, SVM, Naive Bayes (NB).

b. Assign optimal weights w_i via Bayesian Optimization: $\hat{Y} = \Sigma\, w_i\, \hat{Y}_i$.

C. Cascading Models:

a. Train a Decision Tree (SimpleCart) for handling easy cases.

b. Use MLP for uncertain cases.

c. Final refinement with Random Forest for cases with ambiguous probability outputs from the MLP.

D. Feature-Level Fusion with Autoencoder:

a. Train Autoencoder A(X) for feature compression: X' = A(X).

b. Train RF, SVM, and MLP on compressed features X'.

c. Combine predictions via stacking or voting.

E. SHAP-Enhanced XGBoost Model

a. Train an XGBoost classifier using optimized hyperparameters on the pre-processed dataset.

b. Compute SHAP values to assess feature importance and interpretability.

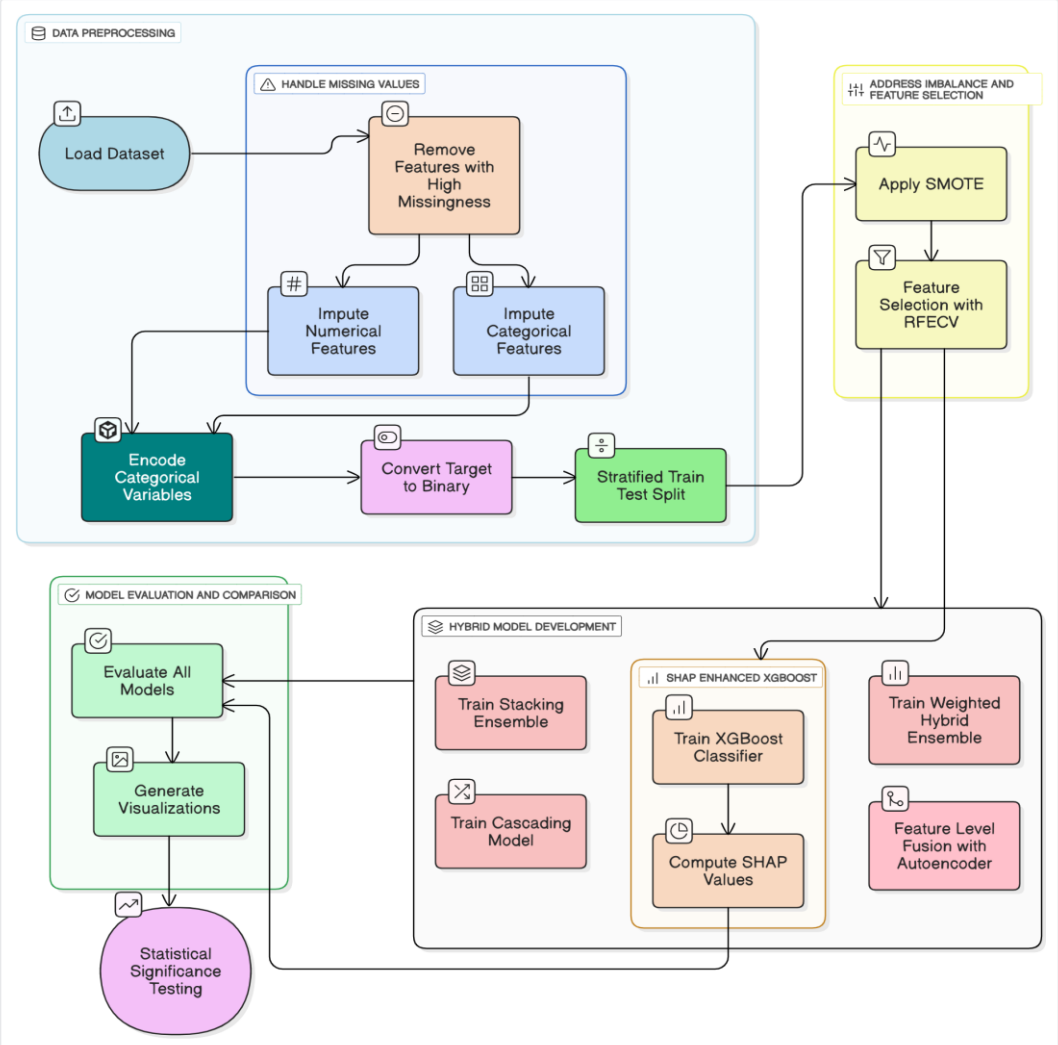

Figure 1: Data processing and model development flowchart for ivf live-birth prediction

c. Use SHAP visualizations (summary plot, force plot) to support model explainability for clinical insight.

Step 3: Model Evaluation and Comparison
1. Evaluate models using metrics: Accuracy, Precision, Recall, F1-Score, AUC
2. Generate ROC curves.
3. Visualize performance metrics.

Step 4: Statistical Significance Testing
1. Perform Wilcoxon Signed-Rank Test to compare predicted probabilities of the hybrid model vs. baseline Random Forest model.
2. Perform McNemar's Test to assess classification agreement and reduction in misclassification errors.
End of Algorithm

3.2 Data preprocessing

Data pre-processing is an essential component in building a strong predictive model that allows data quality, consistency of formats, and reliability for machine learning processes. The dataset was cleaned and pre-processed extensively to train the machine learning model. The target variable (i.e., the number of live births) was transformed into a binary classification, where 0 indicated no live births and 1 indicated at least one live birth. The categorical variables were one-hot encoded for use in different machine-learning algorithms. In addition, data stratification was performed to ensure an 80%-20% training-test split while maintaining a balanced class distribution.

The process starts with loading the IVF dataset; it then deals with missing values by either imputation or removal of feature(s). The approach to handling missing values involves a two-step process aimed at preserving data quality and model robustness. First, features with over 50% missing data were removed to avoid unreliable imputation of severely incomplete variables. Following this, the remaining missing values in the dataset were imputed using the median (for numerical features post-encoding) to ensure consistency and completeness before training. This combined strategy balances dimensionality control with the effective handling of data sparsity, aligning with the pre-processing goals outlined in this section. Missing values were identified with their treatment, keeping note of the loss of data with imputation; numerical features were assigned medians, while categorical features were assigned modes. Categorical variables were encoded, and the target variable was converted to binary for classification.

Moreover, through the synthetic minority oversampling technique, SMOTE provides the representation of minority classes owing to the class imbalance problem. To address RQ3, SMOTE was used to balance the dataset and improve the model fairness. Before SMOTE, the class distribution was 3.23:1; after SMOTE, it equalized to 1:1. Finally, in response to RQ2, Recursive Feature elimination with cross-validation (RFECV) was performed to retain the most informative predictors. This ensures that the input dataset is ready for machine learning models and improves predictive accuracy and generalizability.

3.3 Hybrid model development

The proposed methodology integrates multiple machine learning techniques to enhance the prediction accuracy of live birth outcomes in IVF treatments. By employing Stacking with Meta-Learners (Layered Learning), Weighted Hybrid Ensembles, Cascading Models (multi-stage refinement), Feature-Level Fusion with Autoencoder Networks, and SHAP-Enhanced XGBoost for post-hoc explainability, a comprehensive and robust predictive framework was developed to enhance both accuracy and interpretability in IVF live birth outcome prediction. This hybrid approach ensures improved model reliability and effectiveness in clinical decision-making. The algorithm outlined below outlines a hybrid machine learning framework that integrates ensemble learning, model stacking, and deep learning to improve live birth outcome predictions in IVF treatments. By using multiple predictive techniques, the proposed system ensures high accuracy and reliability in clinical decision making. A flow diagram of the Hybrid Model Development process is shown in Figure 2.

3.3.1 Model stacking with different meta-learners

To address RQ1, a hybrid ensemble approach was employed, particularly stacking with XGBoost as a meta-learner, to enhance the generalization across diverse IVF cases with non-linear patterns. Base learners such as Random Forest (RF), Support Vector Machine (SVM), and multilayer perceptron (MLP) were trained separately. Next, their predictions were used as input features for meta-learner feature construction using XGBoost, MLP, and Gradient Boosting equal to Gradient Boosting (GBM).

$$P_{final} = f_{meta}(P_{RF}, P_{SVM}, P_{MLP}) \quad (1)$$

where $\boldsymbol{P_{final}}$ is the final prediction, $\boldsymbol{P_{RF}, P_{SVM}, P_{MLP}}$ are the predictions from base the learners, and $\boldsymbol{f_{meta}}$ represents the meta-learner function.

A strict separation between the base learner training and meta-feature construction was implemented to ensure methodological rigor and prevent data leakage in the stacking ensemble. The full dataset was split into 80% training set and 20% held-out test set. Within the training data, the base models (Random Forest, SVM, MLP) were trained using 5-fold stratified cross-validation, and out-of-fold predictions were collected to construct meta-features. These meta-features, derived from unseen folds, were used to train the meta-learners (XGBoost, MLP, Logistic Regression), ensuring no overlap between the training and prediction phases.

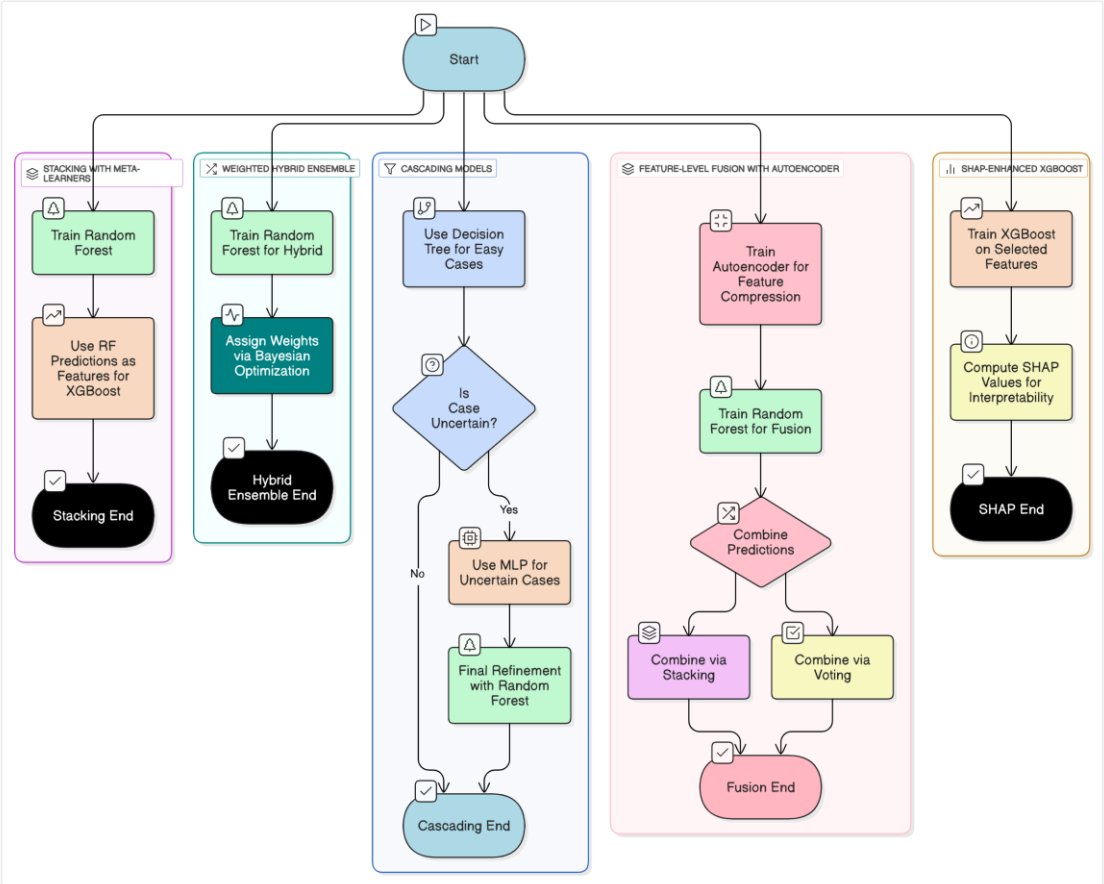

Figure 2: Flow diagram of hybrid model development step

For the final evaluation, base models were retrained on the complete resampled training data, and their predictions on the held-out test set were passed to the trained meta-learner, thereby enabling an unbiased assessment of the generalization performance on previously unseen data.

3.3.2 Weighted hybrid ensembles

A weighted ensemble approach was used to weigh the models based on their performances. Weight assignments were optimized using an optimization scheme such as Bayesian Optimization or Genetic Algorithms.

$$P_{final} = w_1 P_{RF} + w_2 P_{MLP} + w_3 P_{SVM} + w_4 P_{NBP} \quad (2)$$

where w1, w2, w3, and w4 are the optimized weights assigned to each model prediction.

3.3.3 Feature-level fusion with autoencoder networks

RQ5 is addressed using autoencoders for feature-level fusion, which captures non-linear relationships and compresses high-dimensional IVF data into informative latent features. This was used to perform the feature extraction. Random Forest, SVM, MLP were then run-on compressed feature representations to make predictions using stacking or weighted voting.

$$F_{compressed} = AE(X) \quad (3)$$

$$P_{final} = f_{meta}(P_{RF}, P_{SVM}, P_{MLP}) \quad (4)$$

where AE(X) denotes the autoencoder-transformed feature set.

3.3.4 Cascading models (multi-stage predictions)

To address RQ4, a two-stage cascading model was introduced to improve prediction reliability, particularly for borderline or uncertain cases. In the first stage, SimpleCart, a shallow Decision Tree that uses the Gini impurity criterion, was employed to classify cases that were easily separable. Predictions with low confidence were then escalated to include more complex classifiers. In the second stage, a Multilayer Perceptron (MLP) handled these uncertain cases, and samples with ambiguous probability scores (typically between 0.3 and 0.7) from the MLP were further passed to a Random Forest for final refinement. This cascading strategy ensures that uncertain predictions are progressively evaluated by increasingly powerful models, thereby enhancing classification robustness and clinical decision support.

$$P_{final} = f_{RF}\left(f_{MLP}\left(f_{DT}(X)\right)\right) \quad (5)$$

where, X denotes the input dataset.

3.3.5 SHAP-enhanced XGBoost

To improve both the predictive performance and interpretability, we implemented the SHAP-enhanced XGBoost model. XGBoost is a powerful gradient boosting algorithm widely used for structured data classification tasks. However, the decision-making process is often considered a "black box" because of the complexity of tree-based ensembles. To overcome this challenge, we integrated the SHAP into the XGBoost pipeline. SHAP is a unified framework based on cooperative game theory that attributes a model's prediction to each feature, thus offering instance-level interpretability.

For a specific prediction $\hat{y}_i$, the SHAP framework decomposes the output as:

$$\hat{y}_i = \phi_0 + \sum_{j=1}^{n} \emptyset_j^{(i)} \quad (6)$$

where ϕ_0 is the expected value of the model's output over the training data, and $\emptyset_j^{(i)}$ is the SHAP value representing the contribution of feature j for instance i.

Table 2 summarizes the internal architecture of the proposed hybrid machine learning models. It elucidates the model components, feature extraction, training, and optimization for all individual models. It also specifies the activation functions and decision logic at lower levels of detail and linkage, thereby illustrating the interactions of different models within each hybrid framework.

3.4 Hyperparameter tuning

The effectiveness of the proposed hybrid models is not solely determined by their architectural design, but also by the precision of the hyperparameter optimization, which directly addresses RQ6. To ensure robust generalization and predictive reliability, a systematic hyperparameter-tuning process was conducted using cross-validation of resampled training data. This process aims to strike a balance between model complexity, overfitting control, and computational efficiency.

Different tuning strategies were applied based on the model type and complexity. Grid Search Cross-Validation (GridSearchCV with 5-fold stratified CV) was employed for the baseline Logistic Regression model, stacking meta-learners (XGBoost, Neural Network, Logistic Regression), and an autoencoder-based classifier. For SHAP-enhanced XGBoost, a more efficient Random Search Cross-Validation (RandomizedSearchCV with 5-fold stratified CV) was adopted to explore a broader parameter space. The Weighted Hybrid Ensemble leveraged Bayesian Optimization (Bayesian optimization) was used to determine the optimal model weights, assessed via an inner 3-fold stratified cross-validation loop.

All tuning strategies prioritize the ROC AUC score as the primary evaluation metric to guide the selection of optimal parameter combinations. The final tuned hyperparameters for each model are summarized in Table 3, and collectively contribute to the superior predictive performance and reliability of the proposed hybrid framework.

This study utilized a modular implementation framework in Python 3.10, leveraging multiple machine-learning libraries to ensure consistency and reproducibility across models. Specifically, scikit-learn (version 1.6.1) was employed for traditional classifiers, ensemble techniques, and pre-processing tasks; TensorFlow (version 2.18.0) via the Keras API was used to design and train the autoencoder model; and XGBoost (version 2.1.4) was used to implement gradient boosting algorithms. These tools collectively support the development, tuning, and evaluation of proposed hybrid models in a unified pipeline.

3.5 Model evaluation and comparison

The various performance metrics, including accuracy, AUC, precision, recall and F1-score, were used for the comparative analysis of the proposed hybrid models. A ROC curve was constructed to assess the discrimination ability of the models, and radar charts and heat maps were constructed to compare their strengths and weaknesses. This evaluation framework enables a strong performance validation and interpretability for clinical deployment.

3.6 Statistical significance testing

To test the validity of the developed hybrid model, the significance test of Wilcoxon Signed-Rank Test was used to compare the predicted probabilities of the hybrid model with an existing baseline Random Forest model to evaluate the overall differences in performance. McNemar's test was applied to test the classification agreement, specifically to detect improvements in misclassification repair. These statistical tests confirmed that the observed performance improvements were not simply by chance.

With the methodology established for this study, the following section presents the empirical results and a comparative analysis of the proposed hybrid models. The comparative performance of these methods was evaluated based on evaluation metrics.

4 Result and discussion

4.1 Comparative performance of hybrid models

Once the machine learning models were built, the next critical step was their evaluation across multiple dimensions of performance: Accuracy, AUC, Precision, Recall, and F1-Score. Five models were evaluated in this study—Stacking with Meta-Learners (XGBoost), SHAP-enhanced XGBoost, Weighted Hybrid Ensemble, Autoencoder Fusion, and Cascading Model. The results are presented in Table 4, supplemented with ROC curves, radar charts, and heatmaps to allow for a comprehensive comparative analysis.

Among all models, SHAP-enhanced XGBoost demonstrated the best overall performance, achieving an accuracy of 0.997, AUC of 1.0, precision of 0.995, recall of 1.0, and F1-Score of 0.997. This clearly indicates that

the integration of SHAP explainability mechanisms not only preserves but also slightly improves the predictive capability compared to the original stacking approach. The Stacking (XGBoost) model followed closely, with an AUC of 0.999 and an F1-score of 0.985. The Weighted Hybrid model also performed well, while the Autoencoder Fusion model demonstrated clinically acceptable performance with high recall and reasonable precision. In contrast, the cascading model underperforms across most evaluation metrics.

Initially, some models exhibited signs of overfitting, as reflected by the unrealistically high AUC values during preliminary training. To correct this, multiple regularization techniques were applied: L2 regularization (for XGBoost, MLP, and Logistic Regression), reduced decision-tree depth, and early stopping in MLPs. Dropout was integrated into autoencoders to mitigate overfitting.

Table 2: Internal architecture details of hybrid models

Aspect	**Stacking with Meta-Learners**	**Weighted Hybrid Ensemble**	**Feature-Level Fusion**	**Cascading Model**
Model Type	Stacking Ensemble with Meta-Learners	Weighted Hybrid Ensemble	Autoencoder-based Feature Fusion with Classifier	Two-Stage Cascading Model
Architecture Details	Base models: Random Forest, SVM, MLP Meta-learners: XGBoost, MLP, Logistic Regression	Base models: Random Forest, MLP, SVM, Naïve Bayes Final output via Bayesian-weighted average	Autoencoder: Input → Encoding → Bottleneck → Decoding Classifier: Random Forest	Stage 1: Decision Tree (max_depth=3) Stage 2: MLP and Random Forest for refinement
Feature Engineering	RFECV for feature selection Standardization & SMOTE	RFECV for feature selection Standardization & Normalization	Encoded feature representation Random Forest trained on encoded outputs	Recursive Feature Elimination Stratified Data Splitting
Training Process	80–20 Train–Test Split 5-Fold Cross-Validation	Bayesian Optimization (10 initial pts, 20 iterations)	Autoencoder trained for dimensionality reduction Random Forest classifier trained separately	Stage 1 trained on full data Stage 2 trained on uncertain predictions (prob. = 0.3–0.7)
Optimization & Regularization	Early stopping (MLP) L2 regularization (α = 0.1 in MLP) XGBoost: learning_rate tuned (0.01–0.1), reg_lambda tuned (0.1–1)	L2 regularization in MLP Weights optimized via Bayesian search	Dropout (0.4 in encoding layer) ReLU for encoder, Sigmoid for decoder L2 regularization in RF	Dropout (0.5 in MLP) Early stopping in MLP
Activation Functions	MLP: ReLU (hidden), Logistic (output) SVM: Linear kernel Logistic Regression: Sigmoid	MLP: ReLU (hidden), Softmax (output) SVM: Linear kernel	Autoencoder: ReLU (encoding), Sigmoid (decoding)	MLP: ReLU (hidden), Logistic (output) Decision Tree: Gini criterion
Optimizers	MLP: Adam (default in sklearn) Logistic Regression: lbfgs XGBoost: Tree Booster (built-in)	MLP: Adam SVM and Naïve Bayes: Implementation-defined	Autoencoder: Adam Random Forest: Not Applicable	MLP: Adam Random Forest: Not Applicable Decision Tree: Not Applicable
Learning Rate	MLP: 0.001 (default) XGBoost: Tuned (0.01–0.1) Logistic Regression: Controlled by solver	MLP: 0.001 (default) SVM: Controlled by implementation	Autoencoder: 0.001 (default Adam) RF: Not Applicable	MLP: 0.001 (default) RF and DT: Not Applicable
Decision Flow	Base models predict first → meta-learners combine via stacking for final prediction	Individual model predictions weighted by performance → final prediction via weighted sum	Input passed through encoder-decoder → encoded output classified by Random Forest	Stage 1: Decision Tree filters easy predictions Stage 2: MLP and RF handle harder cases

Table 3: Hyperparameter tuning for hybrid models

Hybrid Model	Component	Hyperparameter	Value / Setting	Purpose
Model Stacking	Base Models	Random Forest – max_depth	5	Prevent overfitting, improve generalization
		SVM – C	0.1	Improve robustness, avoid excessive complexity
		MLP – hidden_layer_sizes	(20,)	Single hidden layer with 20 neurons
		MLP – max_iter	200	Limit training time
		MLP – alpha (L2 reg.)	0.1	Prevent overfitting
		MLP – early_stopping	Enabled (validation_fraction=0.2)	Stop training when validation performance stagnates
	Meta-Learners	XGBoost – n_estimators	25	Reduce overfitting and training time
		XGBoost – reg_lambda (L2 reg.)	1	Improve generalization
		Neural Network – hidden_layer_sizes	(10,)	Simpler meta-learner architecture
		Logistic Regression – penalty	L2	Improve model regularization
		Logistic Regression – C	0.1	Enhance generalization
Weighted Hybrid Ensemble	Base Models	Same as stacking model	Same as stacking model	-
	Weight Optimization	Search space for weights	(0,1)	Optimize ensemble performance
		Optimization method	Bayesian Optimization	Find optimal model weight distribution
		Initial points, iterations	10 initial points, 20 iterations	Improve accuracy
Cascadi-	Stage 1: Decision Tree	max_depth	3	Simplify early classification

	Stage 2: MLP, RF	Same as stacking model	Trained on uncertain cases only (p = 0.3 - 0.7)	Handle uncertain predictions efficiently
Feature-Level Fusion with Autoencoder	Autoencoder	encoded_dim	Min(30, half of input dim)	Reduce feature dimensionality
		Dropout rate	0.5	Prevent overfitting
		Training epochs, batch size	30 epochs, batch size = 32	Ensure stable training
	Random Forest	n_estimators, max_depth	25, 5	Train on compressed features

Table 4: Comparison of our approach with previous research

Study	Models Used	Accuracy	AUC	Precision	Recall	F1-Score
[13]	ANN + Deep Imaging	>0.75	N/A	N/A	N/A	N/A
[20]	XGBoost	N/A	0.839	N/A	N/A	N/A
[25]	Statistical Methods (Logistic Regression)	N/A	N/A	N/A	0.471	N/A
[26]	CNFE-SE (Ensemble)	0.87	0.87	N/A	0.82	0.92
[27]	Meta-analysis, Systematic Review	N/A	0.905	N/A	N/A	N/A
Proposed Approach	SHAP-enhanced XGBoost	0.997	1	0.995	1	0.997
	Stacking (XGBoost)	0.985	0.999	0.973	0.997	0.985
	Weighted Hybrid Ensemble	0.965	0.993	0.944	0.99	0.966
	Autoencoder Fusion	0.875	0.942	0.835	0.933	0.882
	Cascading Model	0.512	0.548	0.506	0.978	0.667

These corrections have led to more generalizable and realistic evaluation metrics.

Table 4 summarizes the performance of the proposed models and situates them within the context of the prior IVF prediction literature. It is evident that the SHAP-augmented and ensemble-based approaches significantly outperformed traditional statistical and individual machine learning methods.

4.2 Visual analysis of model discrimination power

The Receiver Operating Characteristic (ROC) curve depicted in Figure 3 visually assesses each model's classification capability. SHAP-enhanced XGBoost and Stacking (XGBoost) both achieved near-perfect AUCs (1.0 and 0.999, respectively), indicating excellent discriminative power. The Weighted Hybrid model performed slightly lower, with an AUC of 0.993, whereas the Autoencoder Fusion model reached 0.942. The Cascading model, with an AUC of 0.548, performed marginally better than the random guessing model.

These distinctions are critical in the context of IVF decision making. Higher AUC values support more confident treatment recommendations, especially in borderline cases, where accurate risk estimation is essential for guiding patients on whether to continue or adjust treatment strategies.

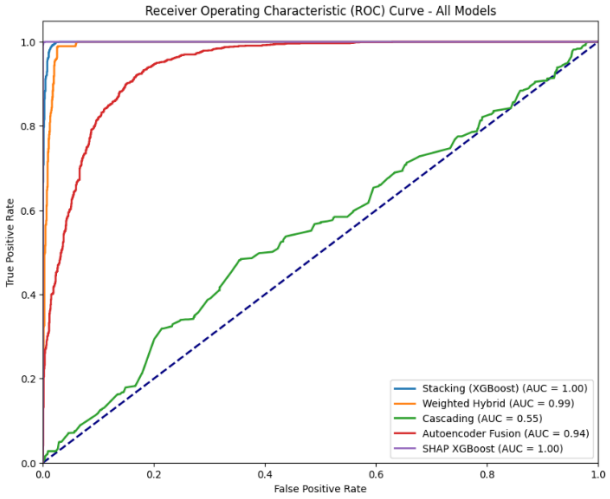

Figure 3: Evaluating model performance for ivf success prediction: roc curve analysis

4.3 Multi-Metric evaluation and comparative strengths

Figure 4 presents a radar chart that offers a simultaneous view of all the five-evaluation metrics across the models. The SHAP-enhanced XGBoost enclosed the widest area, confirming its balanced and robust predictive capability. Stacking (XGBoost) and Weighted Hybrid also showed excellent coverage. The moderate performance of the Autoencoder Fusion model is visible, whereas the cascading model reflects poor balance and lower values across most axes.

Figure 5, a heatmap, further confirms that the SHAP-enhanced XGBoost and Stacking models consistently outperform the others across metrics. Lighter shades in the

heatmap represent a stronger performance. The Cascading model's darker cells in Precision and Accuracy reaffirm its unsuitability for reliable clinical use.

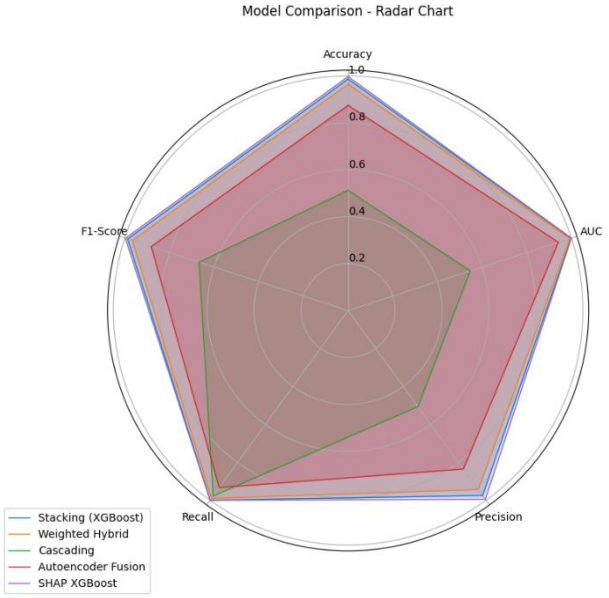

Figure 4: Assessing strengths and weaknesses of IVF prediction models – radar visualization

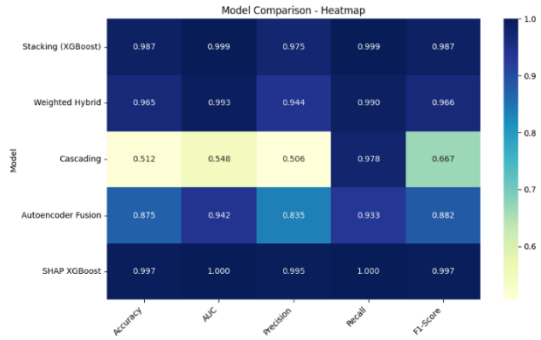

Figure 5: Performance heatmap of predictive models for IVF success prediction

In particular, SHAP-enhanced XGBoost excels in balancing high precision with perfect recall, making it ideal for clinical scenarios in which both false positives and false negatives must be minimized. Meanwhile, Autoencoder Fusion, despite good recall, suffers from potential information loss owing to aggressive feature compression. The Cascading model performs the weakest, likely because of the inability of the initial Decision Tree to filter uncertain cases effectively, leading to downstream overfitting and poor generalization.

Together, these hybrid models demonstrate clear advantages over the traditional standalone ML approaches. In particular, SHAP-enhanced XGBoost and Stacking models represent reliable, high-performance options for AI-assisted decision-making in IVF clinics, capable of providing explainable, patient-specific recommendations. Their high AUC and F1-scores make them suitable for real-world deployment, reducing the emotional and financial burden on patients through a more accurate prognosis.

4.4 Model interpretability with SHAP

To enhance transparency and enable the clinical interpretability of the predictions made by the stacking (XGBoost) model, we employed SHAP. SHAP provides a unified framework to quantify the contribution of each input feature to a model's prediction, making it particularly suitable for medical applications in which explainability is crucial.

Using TreeExplainer from the SHAP Python library, optimized for tree-based models such as XGBoost, we calculated the SHAP values on the meta-features derived from base learners (Random Forest, MLP, and SVM) in the stacking model. This allowed us to evaluate both the global feature influence and the local prediction explanations for individual patients.

The SHAP summary plot (Figure 6) illustrates the average magnitude and direction of the SHAP values for each meta-feature across all the predictions. It is evident that the Random Forest_Stacking output consistently contributes the most to the model's predictions, followed by MLP_Stacking, whereas SVM_Stacking has a minimal impact. This aligns with prior performance evaluations, confirming that Random Forest serves as the most informative base learner in the stacking ensemble.

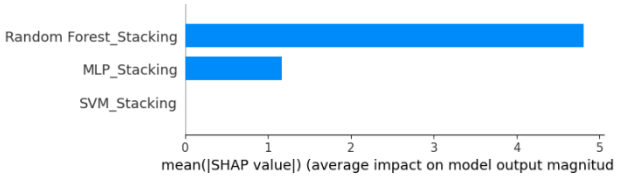

Figure 6: SHAP Summary Plot of Meta-Feature Contributions in Stacking (XGBoost)

To complement global interpretability, Figure 7 presents a SHAP force plot for a representative instance (instance 0). The base value of −0.01528 represents the average model output before any feature influence. In this instance, the prediction shifted to f(x) = 2.14, primarily driven by

- Random Forest_Stacking = 0.8134, and
- MLP_Stacking = 0.7444

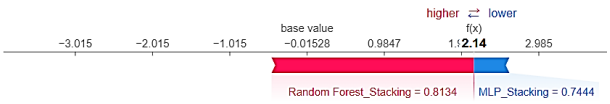

Figure 7: SHAP Force Plot of a Representative Patient Prediction

These values collectively pushed the model output toward a high probability of successful live births. The absence of the SVM_Stacking influence indicates its negligible role in this case.

Together, these SHAP visualizations bridge the gap between high-performance AI models and human decision-making.

- Highlighting the dominant drivers of model predictions.
- Enabling clinicians to interpret individual patient outcomes.
- Building trust in model outputs, particularly for borderline or high-risk cases.

This approach affirms that Stacking with XGBoost, when combined with SHAP-based post-hoc explainability, not only offers exceptional accuracy but also supports clinically meaningful, interpretable predictions for IVF success.

4.5 Generalization and overfitting analysis via learning curves

To evaluate the generalization capability of each proposed hybrid model and address potential overfitting, we analyzed learning curves plotting training and validation error (1 - accuracy) across incremental training set sizes (Figure 8).

The stacking model with the XGBoost meta-learner exhibited the most stable and lowest error rates, with training and validation curves converging closely throughout, thus validating its AUC of 0.999 and accuracy of 0.985. Similarly, SHAP-enhanced XGBoost showed robust generalization, confirming that integrating explainability did not compromise the predictive performance.

The cascading model also maintained tight training-validation alignment, reflecting its ability to progressively handle easy and difficult cases. In contrast, the autoencoder fusion and weighted hybrid models showed higher variance, especially at mid-level training sizes, but stabilized with full training data. These results demonstrate that although all models generalize well, ensemble-based stacking consistently outperforms the others in terms of reliability and predictive robustness. These findings confirm that the reported high performance is not a result of overfitting, but rather due to strong architectural generalization supported by regularization, SMOTE-based balancing, and effective feature engineering.

4.6 Statistical validation and clinical relevance

To further validate that the proposed Stacking with Meta-Learners (XGBoost) hybrid model significantly outperforms traditional machine learning approaches, a comparative analysis was performed, which resulted in a baseline Random Forest model. To assess the statistical significance and validate the observed enhancements, two non-parametric statistical tests were employed: the Wilcoxon Signed-Rank Test, and McNemar's test. The Wilcoxon Signed-Rank Test was used to investigate the differences in the predicted probabilities between the hybrid and baseline models, which allowed us to evaluate their relative performance independent of the distribution. The output showed a test statistic of 20.021 and an extremely low p-value ($p = 1.69 \times 10^{-11}$), with strong evidence (overwhelming or red) of the prediction differences between hybrid and baseline models being significant, suggesting improved predictive power with the hybrid model.

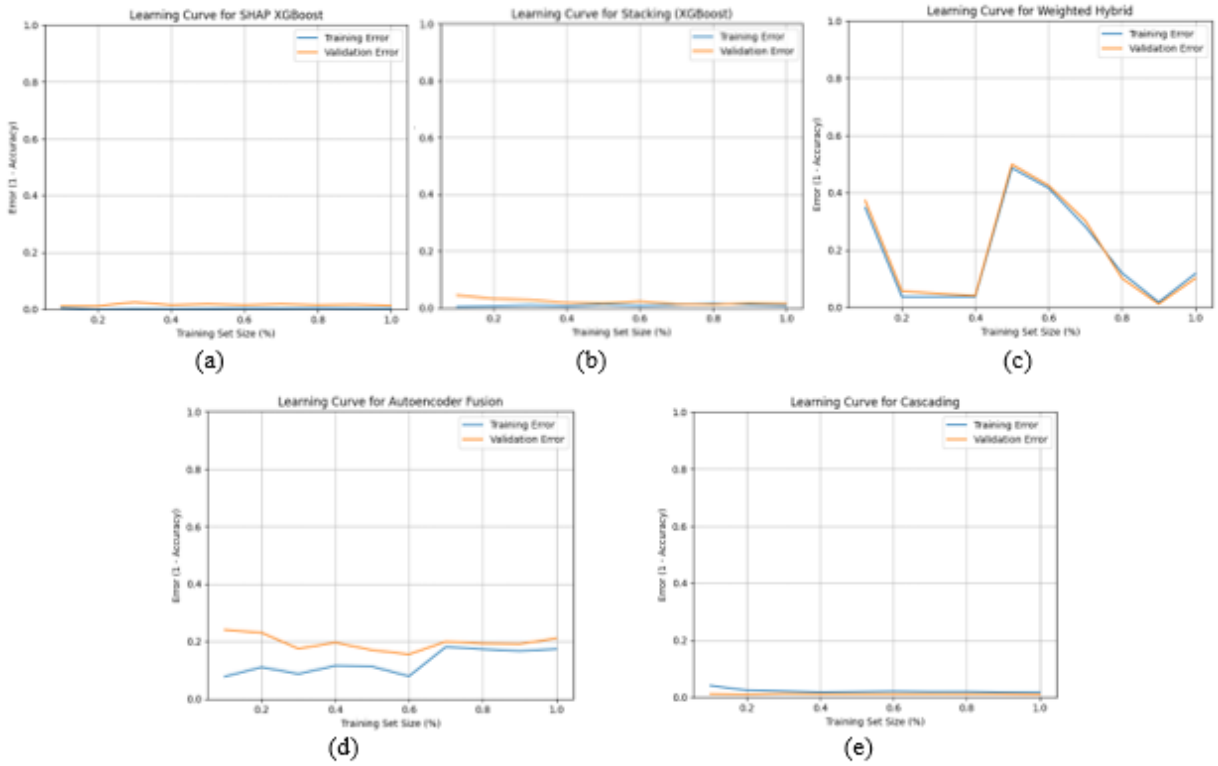

Figure 8: Learning curves showing training vs. validation error across five hybrid models. (a) SHAP-enhanced XGBoost (b) Stacking (XGBoost) (c) Weighted Hybrid Ensemble (d) Autoencoder Fusion (e) Cascading Model.

Furthermore, McNemar's test, which is limited to misclassified cases, was performed to assess the classification agreement between the two models. The test statistic was 20.02 with p-value 7.66×10^{-6} confirming that a significant number of misclassifications from the baseline Random Forest model were corrected by the hybrid model. This indicates a significant improvement in the classification performance, especially when differentiating between successful and unsuccessful IVF cases.

Thus, utilizing these comparative statistical tests provides powerful evidence that stacking with the Meta-Learners (XGBoost) hybrid methodology is superior to conventional machine learning approaches, and enables us to extrapolate this finding with a high level of statistical confidence, further promoting the reliability and clinical applicability of the proposed hybrid model for IVF success prediction. The hybrid model's outperformance further confirms its predictive power and underscores its potential to inform personalized ART treatment.

To further validate the robustness of our hybrid framework, an ablation study was conducted to systematically quantify the contribution of each model component. As presented in Table 5, the removal of individual components resulted in measurable performance degradation, with more pronounced drops observed when multiple components were simultaneously excluded. In particular, the absence of SMOTE and stacking consistently led to lower F1-scores, reflecting their critical roles in handling class imbalance and enabling meta-level learning. Although the removal of RFECV and Bayesian optimization resulted in smaller declines, their contributions to feature selection stability and model tuning remained evident. Notably, the complete removal of all core components led to a breakdown in classification performance, confirming the necessity of each module. Overall, the ablation results demonstrate that the hybrid architecture is not only modular but also synergistic, and each element significantly enhances the model's robustness and predictive reliability.

Table 5: Ablation study evaluating the contribution of individual components to the performance of the hybrid IVF outcome prediction model

Configuration	Accuracy	AUC	F1-Score
Full Model (Stacking + Bayes Opt)	0.989	0.9991	0.9893
- SMOTE	0.9845	0.9987	0.9681
- RFECV	0.9879	0.9993	0.9880
- Stacking (Weighted Hybrid)	0.9813	0.9939	0.9817
- Bayesian Optimization (Stacking + Simple LR)	0.9810	0.9941	0.9813
- SMOTE, - RFECV	0.988	0.9996	0.9751
- SMOTE, - Stacking (Weighted Hybrid)	0.9755	0.9930	0.9505
- SMOTE, - Bayesian Optimization (Stacking + Simple LR)	0.9755	0.9903	0.9504
- RFECV, - Stacking (Weighted Hybrid)	0.9840	0.9948	0.9842
- RFECV, - Bayesian Optimization (Stacking + Simple LR)	0.9826	0.9937	0.9829
- Stacking, - Bayesian Optimization (Simple Averaging)	0.9473	0.9933	0.9496
- SMOTE, - RFECV, - Stacking (Weighted Hybrid)	0.9645	0.9898	0.9298
- SMOTE, - RFECV, - Bayesian Optimization (Stacking + Simple LR)	0.9755	0.9911	0.9499
- SMOTE, - Stacking, - Bayesian Optimization (Simple Averaging)	0.973	0.9891	0.9458
- RFECV, - Stacking, - Bayesian Optimization (Simple Averaging)	0.9456	0.9924	0.9481
- All Components (Simple Averaging)	0.7635	0.9898	0.0000

4.7 Addressing research gaps through hybrid machine learning model

To systematically illustrate how the proposed hybrid models address the primary research challenges present in IVF prediction, a mapping between the defined research questions and how hybrid models contribute to resolving them is shown in Table 6.

These models overcome the individual challenges of generalization, feature selection, class imbalance, uncertainty quantification, deep feature representation, and hyperparameter tuning by capturing the complementary strengths of different machine-learning integration of diverse classifiers, with optimal feature selection from within the weighted hybrid ensemble using Bayesian Optimization. Autoencoder fusion plays a role in the extraction of deep feature representations as well as in improving model interpretability and performance. Predictive performance is improved by flexibility in a cascade model, where predictions are iteratively fine-tuned based on uncertainty estimates; thus, cases that are harder to predict receive more attuned processing. The structured hybrid modelling approach adopted in this study provides a robust and scalable predictive framework for assisted reproductive technologies. The proposed methodologies holistically address existing research gaps in reproductive management, thereby enabling the development of accurate and trustworthy decision-support systems in reproductive medicine.approaches. Stacking with XGBoost uniquely generalizes across complex IVF datasets owing to the seamless

Table 6: Mapping research questions to hybrid models

Research Gap Questions	Stacking with XGBoost	Weighted Hybrid Ensemble	Autoencoder Fusion	Cascading Model
RQ1: How can predictive models generalize better for complex IVF data?	Combines diverse classifiers to improve generalization.	Uses weighted voting to optimize predictions across models.	Extracts deep feature representations to enhance generalization.	Refines predictions through cascading decision layers.
RQ2: How can feature selection be optimized for IVF prediction?	Uses RFECV for selecting the most relevant features.	Bayesian Optimization fine-tunes feature importance.	Autoencoder extracts hidden features and reduces dimensionality.	Feature selection is refined at multiple classification stages.
RQ3: How can we handle class imbalances in IVF datasets?	Applies SMOTE to rebalance the dataset.	Uses SMOTE-based weighting in hybrid ensemble learning.	Learns balanced representations using autoencoder transformations.	Adjusts for class imbalance dynamically at different stages.
RQ4: How can we quantify the uncertainty in IVF predictions?	Uses confidence scores from multiple models to assess uncertainty.	Weighted ensemble predictions provide confidence estimates.	Autoencoders identify ambiguous cases based on representation patterns.	Uses cascading classifiers to handle uncertain predictions adaptively.
RQ5: How can deep feature representations be utilized for improved prediction?	Meta-learner integrates deep patterns learned by base models.	Feature-weighted hybrid models leverage complex feature interactions.	Uses autoencoder for deep feature extraction and fusion.	Identifies key deep features dynamically across classification stages.
RQ6: How can hyperparameter tuning be improved to achieve better performance?	Applies Bayesian Optimization to fine-tune meta-learner parameters.	Uses Bayesian search to optimize ensemble weight distribution.	Optimizes feature learning through autoencoder parameter tuning.	Cascading logic ensures best-performing hyperparameters at each stage.

5 Conclusion

This study introduced novel hybrid machine-learning methodologies to enhance the predictability, reliability, and interpretability of live birth outcomes in ART. By integrating multiple predictive paradigms, including Stacking with Meta-Learners (XGBoost), Weighted Hybrid Ensembles, Cascading Models, Feature-Level Fusion using Autoencoder Networks, and SHAP-Enhanced XGBoost, the proposed framework effectively addressed key challenges such as data heterogeneity, nonlinearity, limited interpretability, and class imbalance inherent in IVF prediction tasks. Among the models evaluated, Stacking with Meta-Learners (XGBoost) achieved the highest performance (AUC = 0.999, accuracy = 0.985), substantially outperforming traditional statistical and standalone machine learning approaches.

The robustness and generalizability of the proposed models were further validated through a rigorous comparative analysis against a baseline Random Forest model utilizing RFECV. Statistical validation using the Wilcoxon Signed-Rank Test and McNemar's test (both $p < 0.05$) confirmed the significant performance gains of the hybrid models, underscoring their potential to support personalized IVF treatment planning and clinical decision-making.

As an avenue for future research, further exploration of the latent feature space of the Autoencoder Fusion model using dimensionality reduction techniques such as t-distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) may offer deeper insights into feature separability, thereby enhancing both model interpretability and predictive accuracy. Additionally, we aim to perform prospective validation using real-time patient data and assess the integration of the proposed hybrid models into a Clinical Decision Support System (CDSS). This will facilitate deployment feasibility evaluations and strengthen the clinical applicability of AI-driven systems in IVF treatment workflows. Ultimately, the widespread adoption of such intelligent systems will require careful attention to explainable AI (XAI) integration, clinical multicenter validation, and adherence to ethical and regulatory standards to ensure trustworthy, patient-centered reproductive care.

Ethics approval

No medical intervention was performed in couples' biomedical or behavioral routines. Since this study was only about data analysis and no human subjects were involved, an Institutional Review Board (IRB) was not considered necessary. All procedures were strictly performed in accordance with the ethical guidelines.

Data availability

It is publicly available and can be accessed via the direct download link provided in [31]. This dataset was previously used in a study by Goyal et al. [32].

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] K. Sfakianoudis, E. Maziotis, S. Grigoriadis, A. Pantou, G. Kokkini, A. Trypidi, P. Giannelou, A. Zikopoulos, I. Angeli, T. Vaxevanoglou, K. Pantos, M. Simopoulou, Reporting on the Value of Artificial Intelligence in Predicting the Optimal Embryo for Transfer: A Systematic Review including Data Synthesis, Biomedicines. vol. 10, no. 3, p. 697, Mar. 2022. https://doi.org/10.3390/biomedicines10030697

[2] L. Sun, J. Li, S. Zeng, Q. Luo, H. Miao, Y. Liang, L. Cheng, Z. Sun, W.H. Tai, Y. Han, Y. Yin, K. Wu, K. Zhang, Artificial intelligence system for outcome evaluations of human in vitro fertilization-derived embryos, Chin Med J (Engl). vol. 137, no. 16, p. 1939–1949, Aug. 2024. https://doi.org/10.1097/CM9.0000000000003162

[3] L. Rienzi, D. Cimadomo, A. Vaiarelli, G. Gennarelli, J. Holte, C. Livi, M. Aura Masip, P. Uher, G. Fabozzi, F.M. Ubaldi, Measuring success in IVF is a complex multidisciplinary task: time for a consensus?, Reprod Biomed. vol. 43, no. 5, p. 775–778, Nov. 2021. https://doi.org/10.1016/j.rbmo.2021.08.012

[4] S.M. Diakiw, J.M.M. Hall, M. VerMilyea, A.Y.X. Lim, W. Quangkananurug, S. Chanchamroen, B. Bankowski, R. Stones, A. Storr, A. Miller, G. Adaniya, R. van Tol, R. Hanson, J. Aizpurua, L. Giardini, A. Johnston, T. Van Nguyen, M.A. Dakka, D. Perugini, M. Perugini, An artificial intelligence model correlated with morphological and genetic features of blastocyst quality improves ranking of viable embryos, Reprod Biomed. vol. 45, no. 6, p. 1105–1117, Dec. 2022. https://doi.org/10.1016/j.rbmo.2022.07.018

[5] E. Taeidi, A. Ranjbar, F. Montazeri, V. Mehrnoush, and F. Darsareh, Machine Learning-Based Approach to predict Intrauterine Growth Restriction, Cureus, vol 15, no. 7, p. e41448, Jul. 2023, https://doi.org/10.7759/cureus.41448

[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, Feb. 2022, https://doi.org/10.1371/journal.pone.0262661

[7] Z. Li, K. Kamnitsas, B. Glocker, Analyzing Overfitting Under Class Imbalance in Neural Networks for Image Segmentation, IEEE Trans Med Imaging. vol. 40, no. 3, p. 1065–1077, Mar. 2021. https://doi.org/10.1109/TMI.2020.3046692

[8] T. Gong, K. Chen, L. Zhang, J. Wang, Debiased Contrastive Curriculum Learning for Progressive Generalizable Person Re-Identification, IEEE Transactions on Circuits and Systems for Video Technology vol. 33, no. 10, p. 5947–5958, Oct. 2023. https://doi.org/10.1109/TCSVT.2023.3262832

[9] D. Velasquez, E. Perez, X. Oregui, A. Artetxe, J. Manteca, J.E. Mansilla, M. Toro, M. Maiza, B. Sierra, A Hybrid Machine-Learning Ensemble for Anomaly Detection in Real-Time Industry 4.0 Systems, IEEE Access. vol. 10, p. 72024–72036, 2022. https://doi.org/10.1109/ACCESS.2022.3188102

[10] M. Casalechi, M. Reschini, M.C. Palermo, G. Di Stefano, P. Vercellini, L. Benaglia, E. Somigliana, P. Viganò, Is endometrial receptivity affected in women with endometriosis? Results from a matched pair case-control study of assisted reproductive technology treatments, Reprod Biomed. vol. 47, no. 6, p. 103414, Dec. 2023. https://doi.org/10.1016/j.rbmo.2023.103414

[11] A. Goyal, M. Kuchana, K.P.R. Ayyagari, Machine learning predicts live-birth occurrence before in-vitro fertilization treatment, Sci Rep. vol. 10, no. 1, p. 20925, Dec. 2020. https://doi.org/10.1038/s41598-020-76928-z

[12] Y.O. Martirosyan, D.N. Silachev, T.A. Nazarenko, A.M. Birukova, P.A. Vishnyakova, G.T. Sukhikh, Stem-Cell-Derived Extracellular Vesicles: Unlocking New Possibilities for Treating Diminished Ovarian Reserve and Premature Ovarian Insufficiency, Life. vol. 13, no. 12, p. 2247, Nov. 2023. https://doi.org/10.3390/life13122247

[13] C. Siristatidis, S. Stavros, A. Drakeley, S. Bettocchi, A. Pouliakis, P. Drakakis, M. Papapanou, N. Vlahos, Omics and Artificial Intelligence to Improve In Vitro Fertilization (IVF) Success: A Proposed Protocol, Diagnostics. vol. 11, no. 5, p. 743, Apr. 2021. https://doi.org/10.3390/diagnostics11050743

[14] S. Dehghan, R. Rabiei, H. Choobineh, K. Maghooli, M. Nazari, and M. Vahidi-Asl, Comparative study of machine learning approaches integrated with genetic algorithm for IVF success prediction, PLoS ONE, vol. 19, no. 10, p. e0310829, Oct. 2024, doi: https://doi.org/10.1371/journal.pone.0310829

[15] Z. J. Pavlovic, V. S. Jiang, and E. Hariton, Current applications of artificial intelligence in assisted reproductive technologies through the perspective of a patient's journey, Current Opinion in Obstetrics & Gynecology, vol. 36, no. 4, pp. 211–217, Apr. 2024, https://doi.org/10.1097/gco.0000000000000951

[16] Y. Hew, D. Kutuk, T. Duzcu, Y. Ergun, and M. Basar, Artificial intelligence in IVF Laboratories: Elevating outcomes through precision and efficiency, Biology, vol. 13, no. 12, p. 988, Nov. 2024, https://doi.org/10.3390/biology13120988

[17] A. Almomani, K. Nahar, M. Alauthman, M.A. Al-Betar, Q. Yaseen, B.B. Gupta, Image cyberbullying detection and recognition using transfer deep machine learning, International Journal of Cognitive Computing in Engineering. vol. 5, p. 14–26, 2024, doi: 10.1016/j.ijcce.2023.11.002. https://doi.org/10.1016/j.ijcce.2023.11.002

[18] P.H. Vuong, L.H. Phu, T.H. Van Nguyen, L.N. Duy, P.T. Bao, T.D. Trinh, A bibliometric literature review of stock price forecasting: From statistical model to deep learning approach, Sci Prog. vol. 107, no. 1, p. 1-31, Jan. 2024. https://doi.org/10.1177/00368504241236557

[19] J. Yao, X. Zhang, W. Luo, C. Liu, L. Ren, Applications of Stacking/Blending ensemble learning approaches for evaluating flash flood susceptibility, International Journal of Applied Earth Observation and Geoinformation. vol. 112, p. 102932, Aug. 2022. https://doi.org/10.1016/j.jag.2022.102932

[20] Q. Xi, Q. Yang, M. Wang, B. Huang, B. Zhang, Z. Li, S. Liu, L. Yang, L. Zhu, L. Jin, Individualized embryo selection strategy developed by stacking machine learning model for better in vitro fertilization outcomes: an application study, Reproductive Biology and Endocrinology. vol. 19, no. 1, p. 53, Dec. 2021. https://doi.org/10.1186/s12958-021-00734-z

[21] Z. Han, J. Zhao, H. Leung, K.F. Ma, W. Wang, A Review of Deep Learning Models for Time Series Prediction, IEEE Sens J. vol. 21, no. 6, p. 7833–7848, Mar. 2021. https://doi.org/10.1109/JSEN.2019.2923982

[22] C. Luo, J. Zheng, T. Li, H. Chen, Y. Huang, X. Peng, orthogonally constrained matrix factorization for robust unsupervised feature selection with local preserving, Inf Sci (N Y). vol. 586, p. 662–675, Mar. 2022. https://doi.org/10.1016/j.ins.2021.11.068

[23] X. Mi, B. Zou, F. Zou, J. Hu, Permutation-based identification of important biomarkers for complex diseases via machine learning models, Nat Commun. vol. 12, no. 1, p. 3008, May 2021. https://doi.org/10.1038/s41467-021-22756-2

[24] A.T. Tredennick, G. Hooker, S.P. Ellner, P.B. Adler, A practical guide to selecting models for exploration, inference, and prediction in ecology, Ecology. vol. 102, no. 6, p. e03336, Jun. 2021. https://doi.org/10.1002/ecy.3336

[25] J. Diao, G. Gao, Y. Zhang, X. Wang, Y. Zhang, Y. Han, A. Du, H. Luo, Caesarean section defects may affect pregnancy outcomes after in vitro fertilization-embryo transfer: a retrospective study, BMC Pregnancy Childbirth. vol. 21, no. 1, p. 487, Dec. 2021. https://doi.org/10.1186/s12884-021-03955-7

[26] S. Ranjbari, T. Khatibi, A. Vosough Dizaji, H. Sajadi, M. Totonchi, F. Ghaffari, CNFE-SE: a novel approach combining complex network-based feature engineering and stacked ensemble to predict the success of intrauterine insemination and ranking the features, BMC Med Inform Decis Mak. vol. 21, no. 1, p. 1, Dec. 2021. https://doi.org/10.1186/s12911-020-01362-0

[27] F. Del Giudice, F. Belladelli, T. Chen, F. Glover, E.A. Mulloy, A.M. Kasman, A. Sciarra, S. Salciccia, V. Canale, M. Maggi, M. Ferro, G.M. Busetto, E. De Berardinis, A. Salonia, M.L. Eisenberg, The association of impaired semen quality and pregnancy rates in assisted reproduction technology cycles: Systematic review and meta-analysis, Andrologia. vol. 54, no. 6, p. e14409, Jul. 2022. https://doi.org/10.1111/and.14409

[28] S. Wu, Y. Zhang, X. Wu, G. Hao, H. Ren, J. Qiu, Y. Zhang, X. Bi, A. Yang, L. Bai, J. Tan, Association between exposure to ambient air pollutants and the outcomes of in vitro fertilization treatment: A multicenter retrospective study, Environ Int. vol. 153, p. 106544, Aug. 2021. https://doi.org/10.1016/j.envint.2021.106544

[29] Y. Zou, Y. Pan, N. Ge, Y. Xu, R. Gu, Z. Li, J. Fu, J. Gao, X. Sun, Y. Sun, Can the combination of time-lapse parameters and clinical features predict embryonic ploidy status or implantation? Reprod Biomed. vol. 45, no. 4, p. 643–651, Oct. 2022. https://doi.org/10.1016/j.rbmo.2022.06.007

[30] K. Koc, A.P. Gurgun, Scenario-based automated data pre-processing to predict severity of construction accidents, Autom Constr. vol. 140, p. 104351, Aug. 2022. https://doi.org/10.1016/j.autcon.2022.104351

[31] hfea.gov.uk, 2025. https://www.hfea.gov.uk/media/2667/ar-2015-2016-xlsb.xlsb (accessed Jan 15, 2025)

[32] A. Goyal, M. Kuchana, and K. P. R. Ayyagari, Machine learning predicts live-birth occurrence before in-vitro fertilization treatment, Scientific Reports, vol. 10, no. 1, Dec. 2020, https://doi.org/10.1038/s41598-020-76928-z