Computer Methods and Programs in Biomedicine 271 (2025) 108979

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

An integrated optimization and deep learning pipeline for predicting live birth success in IVF using feature optimization and transformer-based models

Arezoo Borji [a,b,c], Hossam Haick [d,e], Birgit Pohn [b], Antonia Graf [b], Jana Zakall [b], S M Ragib Shahriar Islam [a,c], Gernot Kronreif [a], Daniel Kovatchki [f], Heinz Strohmer [f], Sepideh Hatamikia [b,a,c,*]

[a] *Austrian Center for Medical Innovation and Technology, Wiener Neustadt, Austria*
[b] *Department of Medicine, Danube Private University (DPU), Krems, Austria*
[c] *Department of Medical Physics and Biomedical Engineering, Medical University of Vienna, Vienna, Austria*
[d] *Laboratory for Life Sciences and Technology (LiST), Department of Medicine, Danube Private University, Krems, Austria*
[e] *Department of Chemical Engineering, Technion – Israel Institute of Technology, Haifa, Israel*
[f] *Kinderwunschzentrum an der Wien, Austria*

ARTICLE INFO

Keywords:
In-vitro-fertilization (ivf)
Live birth success
Transformer-based model
Particle swarm optimization (pso)
Tab_transformer
Feature Selection

ABSTRACT

The complicated interplay of clinical, demographic, and procedural factors makes it difficult to predict the success of in vitro fertilization (IVF), a commonly used assisted reproductive technology. The goal of this research was to create an artificial intelligence (AI) pipeline that could predict live birth outcomes in IVF treatments with high accuracy.

Design: We evaluated prediction performance by integrating different feature selection methods, such as principal component analysis (PCA) and particle swarm optimization (PSO), with different machine learning-based classifiers, including random forest (RF) and decision tree, as well as deep learning-based classifiers, including a custom transformer-based model and a Tab_transformer model with an attention mechanism. Additionally, this study analyzes confounding factors like patient age and previous IVF cycles and explores the influence of different perturbation and preprocessing techniques and validates the model's robustness under varied scenarios. In addition, Shapley Additive Explanations (SHAP) analysis was performed to enhance interpretability of methods.

Results: This research demonstrated that the best performance was achieved by combining PSO for feature selection with the Tab_transformer-based deep learning model, yielding an accuracy of 97 % and an AUC of 98.4 %, highlighting its significant performance in prediction live births. By identifying the most significant predictors of infertility and guaranteeing clinical significance, SHAP analysis significantly improved interpretability.

Conclusion: With the accuracy and interpretability, this study develops a strong AI pipeline for predicting live birth outcomes in IVF. This study establishes a highly accurate AI pipeline for predicting live birth outcomes in IVF, demonstrating its potential to enhance personalized fertility treatments.

1. Introduction

Assisted reproductive technologies (ART), particularly in vitro fertilization (IVF), have transformed the landscape of infertility treatment, offering hope to millions of couples worldwide [1]. Despite advancements in embryology and clinical practices, achieving consistent success in IVF remains a significant challenge [2]. Key results, like whether an embryo implants successfully and leads to a live birth, rely on many different factors; the complexity of IVF results comes from the need for many things to work together for the treatment to succeed [3, 4]. This includes patient age, hormonal profiles, clinical protocols, embryological characteristics, and even lifestyle or genetic factors, all of which contribute to the complexities that surround the process [5]. Each of these variables influences treatment success, making it challenging to

* Corresponding author at: Department of Medicine, Danube Private University (DPU), Krems, Austria
E-mail address: sepideh.hatamikia@dp-uni.ac.at (S. Hatamikia).

https://doi.org/10.1016/j.cmpb.2025.108979
Received 4 April 2025; Received in revised form 11 July 2025; Accepted 20 July 2025
Available online 21 July 2025

predict outcomes and optimize protocols. As illustrated in Fig. 1, the IVF process involves key stages, each contributing to these outcomes, from patient evaluation and ovarian stimulation to embryo selection and transfer. Traditional methods for embryo selection and live birth prediction are often unable to integrate and analyze these multidimensional data aspects effectively, as they primarily rely on static morphological grading systems, while foundational, are often subjective and limited in their ability to capture the complex dynamics of embryonic development and live birth outcomes [3].

Recent advancements in machine learning and artificial intelligence (AI) have introduced a paradigm shift in IVF, providing tools to analyze vast and complex datasets with unprecedented precision [4]. AI and ML have revolutionized IVF by automating embryo evaluation, predicting implantation potential, and enhancing live birth outcomes [5]. These technologies address many of the limitations of traditional methods, offering unprecedented precision, consistency, and scalability. They enable the analysis of large and complex datasets, offering predictive insights that surpass the capabilities of traditional statistical models [6].

One of the most promising applications of AI in IVF is embryo selection, where AI can predict the likelihood of a live birth for individual embryos [7]. Deep learning models, especially convolutional neural networks (CNNs), have shown success in automating embryo grading by analyzing time-lapse imaging data [8]. This technology helps identify embryos with a higher probability of resulting in a live birth, enhancing decision-making during the IVF process. Annotation-free scoring systems, such as those described by Ueno et al. [8], have further streamlined the embryo evaluation process by eliminating the need for extensive manual input while maintaining high predictive accuracy. These models analyze morphogenetic parameters, such as pronuclear fading, cleavage patterns, and blastulation timing, giving researchers dynamic perspectives on embryo development that were previously unattainable through static morphological assessments [9].

Beyond embryo grading, AI has been employed to predict implantation potential with notable success. Machine learning methods, like random forests and ensemble models, use information about embryo development and patient details to evaluate the chances of implantation. Research by Bamford et al. [10] and Uyar et al. [11] have demonstrated the ability of these models to achieve area under the curve (AUC) values exceeding 0.75 for implantation prediction. Furthermore, reinforcement learning-based systems like Dyn Score could dynamically update predictions in real time, offering clinicians actionable insights into embryo viability [12]. These adaptive models represent a significant step forward in IVF decision-making, allowing for more personalized and precise treatment strategies.

The goal of IVF is to achieve a live birth, making the prediction of live birth outcomes a critical focus of AI research [2]. AI models, which integrate patient demographics, clinical data, and simple quantitative features from imaging modalities, have shown promise in this domain. For instance, studies by Huang et al. [13] and Jiang et al. [14] utilized voting ensembles of CNNs to predict embryo ploidy, resulting in improvements in live birth rates. These models enable clinicians to optimize treatment protocols and maximize the likelihood of success.

Researchers have also proposed feature selection techniques to create efficient AI-based methods that support the IVF process. Kragh et al. [15]. explored distinctions between ranking embryos based on implantation potential and predicting probabilities of implantation success, as well as issues like dataset balancing, selection bias, and clinical applicability. By focusing on the most relevant features, these methods enhance model interpretability and reduce computational complexity without compromising performance. Studies by Ueno et al. [8] and Bamford et al. [10] highlighted the importance of feature selection in improving the efficiency and accuracy of IVF-based predictive models.

Several studies proposed promising AI methods for classifying live birth success as a binary outcome (success/failure). For example, Zhang et al. [16], utilizing 57,558 HFEA records, machine learning models such as artificial neural networks (ANN) and LogitBoost to predict live birth outcomes for natural-cycle IVF, with an AUC of approximately 0.79. Their method's ability to capture intricate feature interactions was limited, though, by the absence of sophisticated deep learning architectures like Transformer models. McLernon et al. [17]. applied a discrete-time logistic regression model, while Jones et al. [18] also utilized logistic regression. Sanders et al. [19] conducted a comparison of live birth rates using binary logistic regression. Raful Hassan et al. [20] used a hill-climbing wrapper algorithm for feature selection. Milewski et al. [21] employed the SIMBAF algorithm, a margin-based feature selection method that enhances classification performance. Finally, different approaches achieved an accuracy between 0.73 and

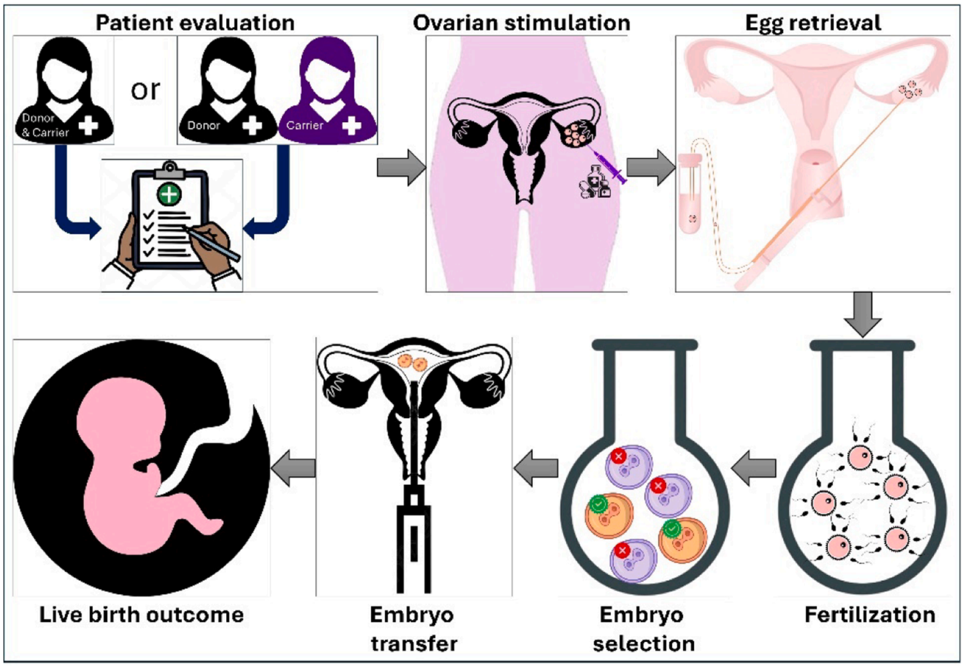

Fig. 1. Step-by-step process of in vitro fertilization (IVF).

0.96.

Despite the promising results achieved so far, the prediction of live birth outcomes using AI has not yet been integrated into clinical practice, meaning that it requires further innovation and development of more robust approaches in this area. Most prior relevant research on live birth prediction has primarily relied on traditional AI models, often overlooking the performance enhancements that advanced deep learning methodologies could offer [15–20]. Leveraging these cutting-edge deep learning techniques has the potential to refine predictive accuracy and enable more reliable, data-driven decision-making in clinical settings.

Our work tries to enhance previous works on live birth prediction by presenting a novel, integrated optimization and deep learning pipeline designed to predict live birth success in IVF with greater accuracy. This pipeline smoothly brings together Particle Swarm Optimization (PSO), a method used to choose important features, with a sophisticated deep learning model based on Tab_transformer, creating a new and effective way to handle the complicated data from IVF. While PSO has been widely utilized in other fields for optimizing feature subsets, its potential in IVF prediction tasks remains largely untapped. By incorporating PSO, the pipeline identifies the most influential features, streamlining the model and enhancing its interpretability. Simultaneously, transformers, originally developed for natural language processing, are adapted to capture intricate interactions between clinical and demographic variables, demonstrating superior predictive capabilities compared to traditional machine learning models. The use of transformer models for IVF prediction tasks, including live birth prediction, remains an unexplored area of research. This study shows a combination of PSO, and transformers provides a robust framework with significant performance for advancing IVF live birth prediction. We validated the proposed method using the open access dataset 2010–2018 HFEA. Additionally, this study used perturbation techniques (noise addition, outlier removal, and synthetic minority over sampling technique (SMOTE) to evaluate robustness, employed SHAP analysis for interpretability, and examined confounding variables such as patient age and prior IVF cycles to ensure accurate and bias-free predictions.

Fig. 1 shows the IVF procedure, starting with patient assessment for appropriateness and donor egg utilization. After hormone-stimulated ovarian stimulation, ultrasound-guided aspiration retrieves several eggs. The eggs are fertilized in the lab, then embryo selection selects the healthiest embryos for transfer [7]. Lastly, the uterus receives the selected embryos for successful implantation and a live birth. It shows the precision and complexity of assisted reproductive technology.

2. Methodology

2.1. General experimental design

In this work, we have applied inclusion and exclusion criteria to enhance the quality and relevance of the data, ensuring it was appropriate for our binary classification task. To reduce the dimensionality of the dataset and improve model performance, we utilized two feature selection and reduction techniques: Principal Component Analysis (PCA) and PSO. For classification, we evaluated the performance of four different classifiers: Random Forest (RF), Decision Tree (DT), a transformer-based model, and a Tab_transformer-based model. Finally, we designed different experimental setups: the first used PCA features as input to all classifiers (Method 1 and Method 3, Fig. 2), and the second used features provided by PSO as input to all these classifiers (Method 2 and Method 4, Fig. 2). In total, we have eight classification models including, PCA+RF, PCA+Decision Tree, PSO+RF, PSO+Decision Tree, PCA+Transformer-based model, PCA+ Tab_transformer-based model, PSO+Transformer based model, and PSO+Tab_transformer-based model.

2.1.1. The dataset used

For this study, we utilized the Human Fertilization and Embryology Authority (HFEA) dataset, an anonymized registry dataset that encompasses fertility treatments conducted from 2010 to 2018. Designed to enhance patient care and maintain strict confidentiality for patients, donors, and offspring, this dataset stands as one of the most comprehensive and longest-running repositories of fertility treatment records globally. With 665,244 patient records and an initial set of 94 features, it provides a detailed account of fertility treatment cycles, covering patient demographics, treatment protocols, and infertility causes. These attributes present a comprehensive view of the factors influencing IVF outcomes during this period. The dataset includes both numerical and categorical variables, capturing a broad spectrum of critical factors. Key features encompass patient-specific details such as age at the time of treatment, number of prior IVF pregnancies, live birth outcomes, and specific infertility causes (e.g., tubal disease, ovulatory disorders, or male infertility factors). Additionally, we have meticulously recorded procedural details such as the type of eggs and sperm used (e.g., fresh, frozen, donor, or patient-derived), the number of eggs collected, and the number of embryos transferred. This level of granularity allows for an in-depth analysis of the variables affecting IVF success rates. This study assesses the prediction performance of live birth success as a binary outcome (success/failure).

To evaluate the generalizability of the proposed model, we tested its performance on an external temporal HFEA dataset from 2005 to 2009. Currently, there are no publicly available patient-level datasets that match the clinical and embryological depth of the HFEA registry. Well-known external resources, such as the CDC's NASS dataset, provide only clinic-level aggregate statistics and lack individual-level outcome and feature data required for machine learning–based prediction[1] (Similarly, datasets like OPTIMIST (Netherlands) and eIVF (USA) are not openly accessible and require formal data access requests or institutional collaborations. Therefore, we used a temporal split within the HFEA dataset to simulate out-of-sample generalization. While our original model was trained on data from 2010 to 2018, we additionally evaluated the performance of the proposed methods on earlier data from 2005 to 2009, which reflects temporally distinct clinical practice patterns.We note that the HFEA registry is inherently multi-center, as it includes data from all licensed fertility clinics across the United Kingdom. As mandated by UK regulationsall fertility clinics are required to submit treatment records to the HFEA, ensuring broad geographic, institutional, and clinical diversity within both training and validation cohorts.[2] As an additional evaluation, we also train and test our proposed models using the dataset from the same time period selected by Sadeghzadeh et al. [33] (train on dataset from 2010–2016 and test on dataset from 2017–2018), which represents the closest benchmark and demonstrates superior performance compared to earlier studies (see Table 15).

2.2. Data preprocessing pipeline for IVF data analysis

The preprocessing pipeline transformed raw IVF data (Section 2.1.1) into a clean and structured format for the AI pipeline's input. It began with column standardization, ensuring uniformity by converting names to lowercase and removing whitespace. We structurally aligned the datasets by reindexing and adding missing columns and then consolidated them into a single Data Frame for holistic analysis. We removed columns with <1 % non-null values to enhance data quality. We imputed missing values based on the data type. Finally, we numerically encoded categorical features and normalized numerical features to a [0, 1] range (Fig. 2). we used scikit-learn's train_test_split function to divide

[1] https://data.cdc.gov/Assisted-Reproductive-Technology-ART-/2022-Final-Assisted-Reproductive-Technology-ART-Su/cchw-gdwa/about_data

[2] https://ukhealthdata.org/members/human-fertilisation-and-embryology-authorit

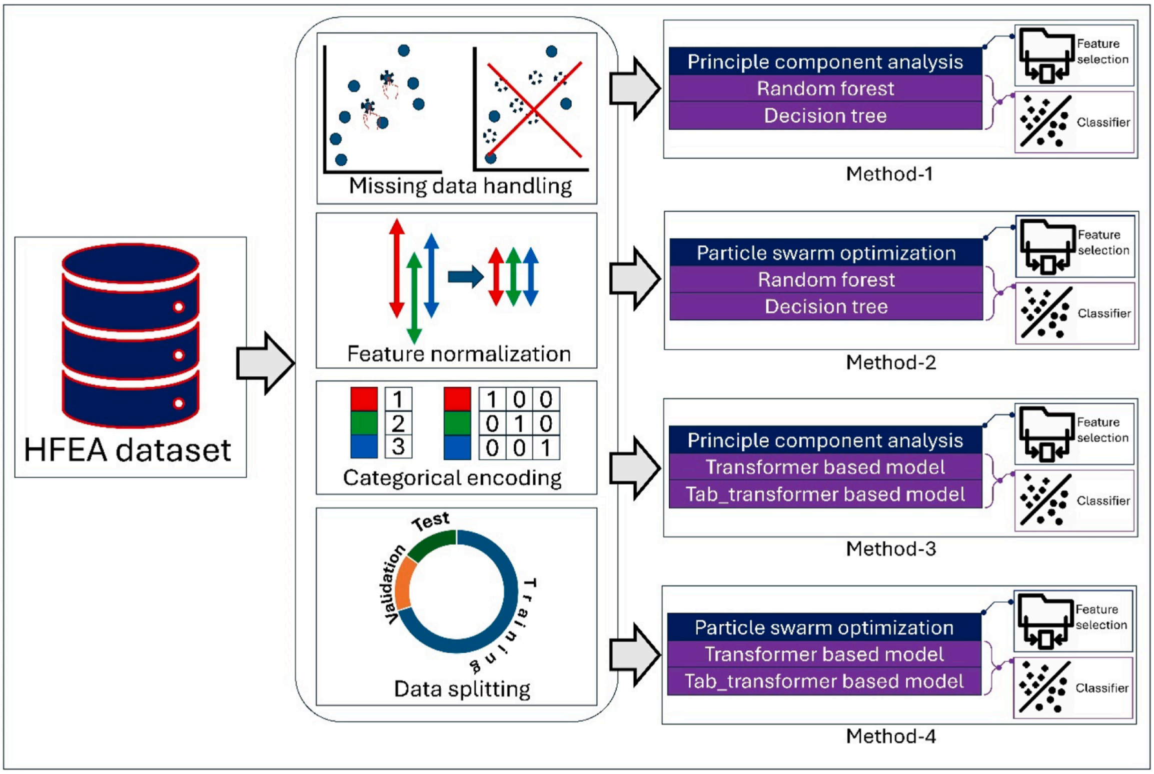

Fig. 2. Overview of preprocessing steps and classification methods used in this paper.

the dataset into three subsets: training (70 %), validation (15 %), and testing (15 %). This partitioning occurred prior to feature selection and model training. While the dataset lacks patient-level identifiers, we ensured that no samples were duplicated across splits. Each treatment cycle was associated with a unique patient ID, ensuring that no patient's records appeared more than one-fold during cross-validation. This approach effectively prevented patient-level data leakage between training and test sets.

2.3. Inclusion and exclusion criteria

Inclusion and exclusion criteria were defined to ensure a clean, complete, and relevant dataset for this study. These criteria were chosen based on the groundwork laid by Sadegh-Zadeh et al. [21], who used the same dataset and adhered to a set of inclusion and exclusion criteria conditions for data preparation and analysis. The dataset included people who met these conditions: (1) they had valid information for the target variable, "live birth occurrence"; (2) they reported at least one infertility-related cause, like "ovulatory disorder"; and (3) their cycle history showed they had not had negative records of previous treatment cycles. These criteria ensured the inclusion of relevant cases with sufficient data for analysis. We applied these inclusion criteria and then implemented exclusion criteria to improve the quality of the data. We excluded subjects with missing information for "elective single embryo transfer", as this variable was crucial for the analysis. Additionally, we removed entries with logical inconsistencies, such as negative treatment cycles or conflicting treatment-related dates, to ensure data validity. By adopting these inclusion and exclusion criteria, this study ensured a high-quality dataset suitable for robust predictive modeling of IVF live birth outcomes. The number of subjects included in this study after exclusion criteria is 115,012.

2.3.1. Feature selection using particle swarm optimization (PSO)

Feature selection reduces the number of predictors and focuses on the most relevant ones [22]. In this study we have employed PSO as a feature selection method due to its efficient search for optimal solutions in large and complex spaces [23]. PSO is a nature-inspired optimization technique, modeled after the social behavior of bird flocking or fish schooling. Each individual component, called a "particle," represents a candidate for a solution in the search space. These particles move through space by adjusting their positions based on their experiences and those of neighboring particles, mimicking how animals in groups share information to find food or navigate environments [20]. PSO works well for tackling complicated optimization problems, like choosing the best features, because it can quickly search through the large number of possible feature combinations. This study employs PSO to pinpoint the ideal feature subset for forecasting the success of live births. Each particle encoded a subset of features as a binary vector, where 1 indicated inclusion of a feature and 0 indicated exclusion of a feature.

Cost Function

The cost function in PSO evaluates the quality of each particle's solution (Eq. (1)). In this study, the cost function is defined as below:

$$C = -(\mathrm{F1} - P\cdot\mathrm{N}) \qquad (1)$$

Where C is the cost function value to be minimized by PSO, F1 is the F1-score of a logistic regression model trained on the selected features. We chose logistic regression as the surrogate model in the PSO fitness function because it is computationally efficient and robust in feature selection tasks. PSO necessitates evaluating the fitness of numerous feature subsets over hundreds of iterations, and using the final classifier for each evaluation would result in computational burden, especially given the model's complexity and training time. In contrast, logistic regression trains quickly and allows for the practical execution of the optimization process. when we use small feature subsets, logistic regression typically produces less variance across folds and is less prone to overfitting than high-capacity models, especially during the intermediate stages of feature selection[24]. This makes it a reliable estimator of generalizable performafnce, allowing PSO to prioritize feature sets that provide broad information rather than overfitting the training data.

The F1 score balances precision and recall, making it suitable for imbalanced datasets like IVF outcomes. P is the penalty weight, a parameter controlling the trade-off between model performance and simplicity. N is the number of features selected by the particle. The goal of PSO is to minimize C, which indirectly maximizes the F1 score while penalizing larger feature subsets. This process ensures the final feature set is both performant and understandable. The following steps outline how to select features using PSO:

Algorithm 1

2.3.2. Dimensionality reduction with principal component analysis (PCA)

Principal Component Analysis (PCA) is a statistical technique used to reduce the dimensionality of a dataset while retaining as much information as possible [24]. It does this by transforming the original data into a new set of orthogonal components, called principal components, which are ranked according to their ability to capture the variance within the data. In this study, we have applied PCA to the IVF dataset to reduce its dimensionality while retaining 95 % of the data's variance. This process can remove irrelevant variations and reduce computational complexity.

2.3.3. Random forest

Random Forest (RF) is an ensemble learning method that combines the outputs of multiple decision trees to improve predictive performance and reduce overfitting [25]. RF naturally evaluates feature importance by measuring the impact of each feature on prediction quality [26]. In this study, we have utilized an RF with 200 decision trees as estimators, each with a maximum depth of 10. Note that we restrict the depth of each tree to avoid overfitting and maintain interpretability. Moreover, the criterion='gini' used Gini impurity to evaluate split quality, random_state: 42, class_weight: 'balanced', min_samples_split: default, min_samples_leaf: default, and bootstrap: True.

2.3.4. Decision tree

A decision tree is a supervised learning algorithm used for classification and regression tasks [27]. It recursively splits the data into subsets based on feature thresholds, forming a tree-like structure where each internal node represents a decision based on a feature, and each leaf node represents an output prediction. Decision trees are highly interpreted, as they clearly outline the decision-making process, making them particularly useful for understanding feature importance and validating selected features. In this study, we used the feature-extracted PCA and fed it into the decision tree with the following parameters: max_depth=10, limited depth to maintain simplicity, criterion=gini, splitter: 'best', random_state: 42, min_samples_split: default, and min_samples_leaf: default.

2.3.5. Transformer-based model

A deep learning model based on the transformer architecture, known as a transformer-based model for classification, can solve classification tasks [28]. Vaswani et al. originally introduced transformer architecture in the "Attention Is All You Need" paper [29], and it has since become the foundation of many state-of-the-art models in natural language processing (NLP), computer vision, and other fields. The attention mechanism is a critical component of the transformer-based model, designed to analyze and interpret tabular data to predict IVF success. The attention mechanism enables the model to dynamically assign importance to specific features, capturing complex interactions among them and improving the model's decision-making process. In this work, the

Algorithm 1
Pseudo code for the feature selection using PSO.

```
Inputs:
      - Dataset with features: $F$
      - Cost function: -($F1$_score-$P$. $N$)
      - Parameters: swarm size $S = 20$, Maximum iterations $T = 1000$, Inertia weight $w = 0.7$, Cognitive coefficient $c_1 = 1.5$, social coefficient $c_2 = 2$, penalty factor $P$
Outputs:
      - Optimal feature subset $F_{\mathrm{optimal}}$
      - Best fitness value $C_{\mathrm{best}}$
Procedure:
  Initialization:
      For each particle $i = 1$ to $S$:
      a. Initialize binary position vector $x_i \epsilon \{0,1\}^n$
      b. Initialize velocity $v_i \epsilon R^n$
      c. Compute number of selected features: $N = \sum_j^n x_{ij}$
      d. Evaluate fitness: $C_i$ = -(F1_ $score_i$ - P. N)
      e. Set the personal best: $p_{best_i} = x_i$ , $C_{p-best_i} = C_i$
      End for
      Set the global best: $g_{\mathrm{best}} = \arg min_{i=1}^{S} C_i$ , $C_{g-best} = min_i (C_i)$
  Optimization:
      For iteration t= 1 to T:
        For each particle $i = 1$ to $S$:
          For each dimension (feature) $j \varepsilon 1$ $to$ $n$ :
            a. Generate random numbers $r_1, r_2 \sim \bigcup(0,1)$
            b. Update the velocity:
                 $v_{ij} = w. v_{ij} + c_1 . r_1 . (p\_ best_{ij} - x_{ij}) + c_2 . r_2 . (g\_best_j - x_{ij})$
            c. Update position using a sigmoid transfer:
                 $\mathbf{s}(v_{ij}) = \frac{1}{1 + e^{-v_{ij}}}$ , $x_{ij} = \{ \begin{matrix} 1, & if\ s(v_{ij}) > rand\ (0,1) \\ & 0,\ otherwise \end{matrix}$
          End for
          Compute $N = \sum_{j=1}^{n} x_{ij}$
          Evaluate fitness: $C_i = -(F1\_score_i - P.N)$
          If $C_i < C_{p\_best_i}$ : $Update\ p\_best_i = x_i$ , $C_{p\_best_i} = C_i$
        End for
         Update the global best:
      If $C_i < C_{g\_best_i}$ : $Update\ g_{\mathrm{best}} = x_i$ , $C_{best} = C_i$
      End for
      Return:
$F_{\mathrm{optimal}} = g_{\mathrm{best}}$, $C_{\mathrm{best}} = C_{g_{\mathrm{best}}}$
```

dataset includes features such as patient age, sperm quality, number of embryos transferred, and other clinical parameters. These features often interact in complex ways. The attention mechanism dynamically determines which features are most important for predicting IVF success and adjusts their importance based on the context of the input data for each individual case. For instance, the attention mechanism may prioritize features such as the quality of embryos for older patients. Younger patients may receive more emphasis on features like the number of eggs retrieved.

The attention mechanism in this transformer-based model operates in the following steps:

Step 1: Input transformation

The input data is composed of tabular features, which are referred to as input_dim features after feature selection. We treat each feature as a component of the input vector. To make the features suitable for attention computation, we first project them into a higher-dimensional space using a dense layer.

$$X = \text{Dense}(x) \tag{2}$$

After this, a sequence dimension is added to simulate sequential processing.

Step 2: Scaled dot-product attention

The scaled dot-product attention mechanism computes the relationships between features:

$$\text{Attention }(Q,\ K, V) = \text{softmax}\left(\frac{Qk^T}{\sqrt{d_k}}\right)v \tag{3}$$

Where Query (Q) represents the feature being queried, key (K) represents the importance of each feature relative to the query, and value (V) contains the actual feature data.

Each feature attends to all other features, producing a matrix of attention scores that capture dependencies between them. The softmax function ensures that the attention scores sum to 1, creating a probabilistic weight for each feature.

Step 3: Multi-head attention

This work employs multi-head attention, dividing the input into multiple "heads." Each head learns to focus on different types of relationships. For instance, one individual might concentrate on the correlations between patient age and success. Another head might emphasize sperm quality or treatment type. We have concatenated and transformed the outputs from all heads into a single vector, combining multiple perspectives.

Step 4: Residual connection and layer normalization

The input is added back to the attention output:

$$x = \text{Add}(\text{x}, \text{AttentionOutput}) \tag{4}$$

Layer normalization: The output is normalized to stabilize gradients and ensure smooth learning.

The attention mechanism powers the Transformer model, offering a sophisticated approach to tabular data analysis for IVF success prediction. The architecture of this transformer-based model proposed in this study is explained in Table 1.

This work configures the transformer model with selected hyperparameters to optimize its performance on IVF success prediction. The number of selected features from PSO determines the input dimension (input_dim), representing the length of the reduced feature set (38 features). We set the number of attention heads (num_heads) to 4, which enables the model to learn diverse relationships between features through parallel attention mechanisms. The feed-forward network dimension (ff_dim) is 128, providing a hidden layer size that refines feature representations after attention. The model includes two transformer encoder layers (num_layers), each consisting of a multi-head attention block and a feed-forward network enabling hierarchical feature extraction.

The model applies to a dropout rate (dropout_rate) of 0.3 and L2 regularization (l2_reg) with a strength of 1e^{-4} to prevent overfitting. We set the batch size (batch_size) to 2048 during training to ensure efficient utilization of computational resources, and we set the learning rate (learning_rate) to a low value 1e^{-6} to ensure stable and gradual convergence. The binary crossentropy loss function optimizes the model for binary classification tasks, monitoring performance metrics such as accuracy and AUC throughout the training process. The model limits training to a maximum of 40 epochs and implements an early stopping patience of 2 epochs to halt learning if the validation loss does not improve, thereby preventing overfitting. These hyperparameters collectively ensure the model's ability to generalize well while capturing complex relationships within the IVF dataset. We have provided all these hyperparameter values using the grid search method. Table 2 displays all these hyperparameters and their corresponding values.

Table 1
The architecture of the proposed transformer-based classification model for predicting live birth success in IVF.

Layer	Output Shape	Explanation
Input Layer	(batch_size, 38)	Raw input features (38 selected features).
Dense Layer	(batch_size, 128)	Projects feature into 128 dimensions.
Sequence Expansion	(batch_size, 1, 128)	Adds a sequence dimension for Transformer processing.
Multi-Head Attention	(batch_size, 1, 128)	Learning relationships between features with 4 attention heads.
Residual + Normalization	(batch_size, 1, 128)	Preserves input information and normalizes activations.
Feed-Forward Network	(batch_size, 1, 128)	Further processes feature representations.
Residual + Normalization	(batch_size, 1, 128)	Adds stability and preserves the input.
Global Average Pooling	(batch_size, 128)	Aggregates the sequence into a single feature vector.
Dense (Output Layer)	(batch_size, 1)	Outputs a probability for the binary classification task.

2.3.6. *Tab_transformer-based model*

The Tab_transformer model is a deep learning approach that combines structured datasets with a mix of categorical and numerical features [30]. The Tab_transformer uses self-attention mechanisms to capture dependencies between features, particularly among categorical features. Instead of representing categorical data using traditional encoding methods, it maps each category to a learned embedding vector. These embeddings allow the model to capture semantic relationships between categories. The architecture starts by converting categorical features into embeddings and merging them with numerical features, either directly or via normalization layers. Transformer layers receive these inputs and use self-attention mechanisms to model the interactions between features. By doing so, the model identifies complex relationships between features that may be critical for the task, such as correlations between specific categories or numerical ranges. Table 3 summarizes the detailed architecture of the proposed Tab_transformer

Table 2
The information of all parameters used for transformer model in this study.

Parameter Name	Value
Input Dimension (input_dim)	38 features
Number of Heads (num_heads)	4
Feed-Forward Dimension (ff_dim)	128
Number of Layers (num_layers)	2
Dropout Rate (dropout_rate)	0.3
L2 Regularization (l2_reg)	1e^{-4}
Batch Size (batch_size)	2048
Learning Rate (learning_rate)	1e^{-6}
Loss Function	Binary_crossentropy
Number of epochs	40

Table 3
The architecture of the proposed Tab_transformer-based classification model for predicting live birth success in IVF.

Layer	Input Shape	Output Shape	Description
Numerical Input Layer	(batch_size, num_numerical)	(batch_size, num_numerical)	Input layer for scaled numerical features.
Categorical Input Layer	(batch_size, 1) (per feature)	(batch_size, embedding_dim)	Embedding layers convert categorical indices into dense vector representations.
Concatenation Layer	Combined inputs	(batch_size, total_dim)	Numerical features and categorical embeddings are concatenated.
Reshape Layer	(batch_size, total_dim)	(batch_size, 1, total_dim)	Reshapes the input for attention layers.
Multi-Head Attention	(batch_size, 1, total_dim)	(batch_size, 1, total_dim)	Self-attention layer captures feature dependencies and relationships.
Layer Normalization	(batch_size, 1, total_dim)	(batch_size, 1, total_dim)	Normalizes outputs from the attention mechanism for stable learning.
Dense Layer (ReLU)	(batch_size, 1, total_dim)	(batch_size, 1, 128)	Fully connected dense layer with ReLU activation to learn higher-level patterns.
Output Layer	(batch_size, 1, 128)	(batch_size, 1)	Sigmoid activation outputs probability for binary classification.

model in this study.

Here, like the transformer-based model proposed in Section 2.3.5, the attention mechanism plays a crucial role in learning complex relationships between the input features. It works by dynamically focusing on different parts of the feature set, which helps the model capture interactions between both numerical and categorical features. The multi-head attention mechanism is particularly powerful in this case because it allows the model to simultaneously attend to multiple aspects of the input data, learning different relationships in parallel.

The attention mechanism operates by computing attention scores that determine how much weight each feature should have in relation to others. Each feature is transformed into a query, key, and value, and the attention mechanism compares the query to all keys to compute a score. This score decides how much attention should be given to the corresponding value. By applying multi-head attention, the model can learn different types of relationships across features simultaneously. For instance, one head may focus on the interaction between numerical features, while another head could focus on categorical feature interactions. Fig. 3 illustrates all the steps in the proposed Tab_transformer model in detail.

In the Tab_transformer architecture, various parameters define the model's structure and behavior. The model uses ReLU activation for the

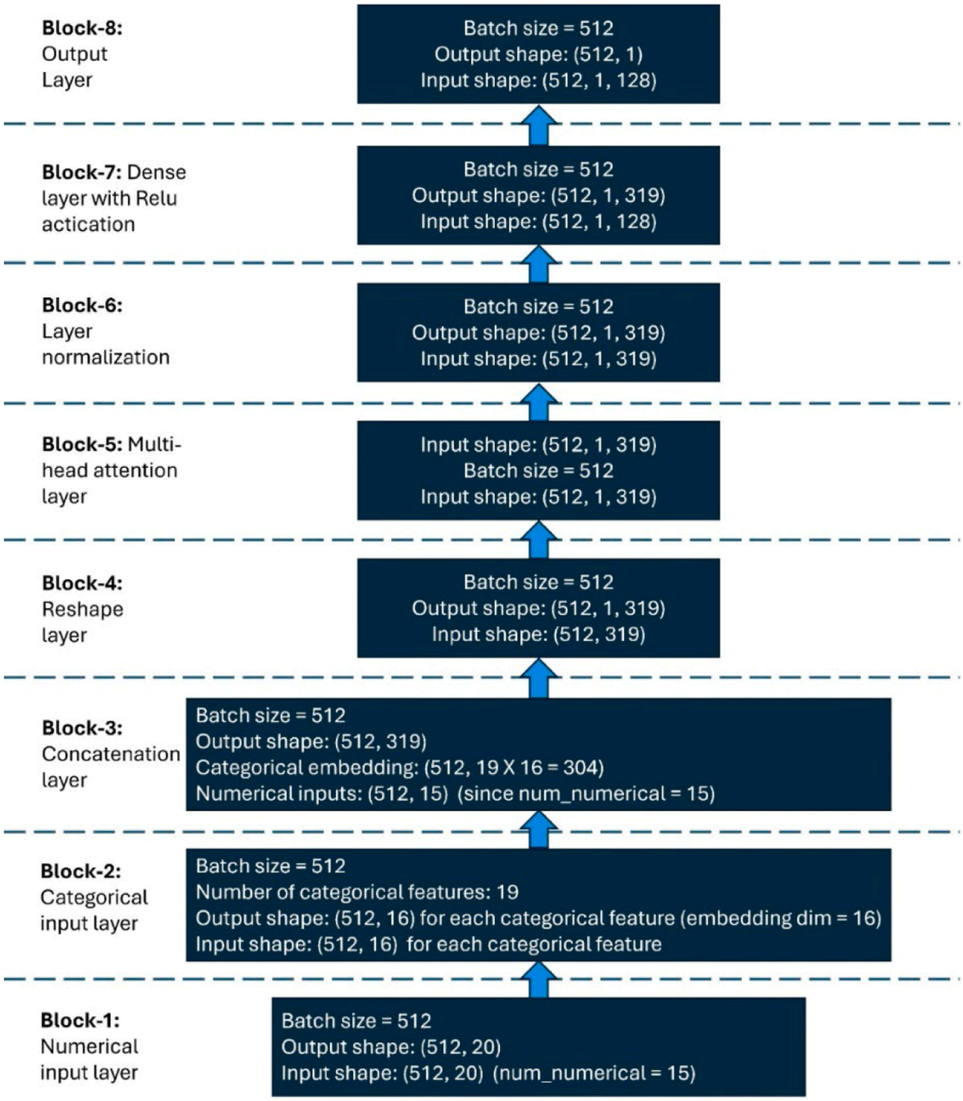
Block-8: Output Layer
Batch size = 512
Output shape: (512, 1)
Input shape: (512, 1, 128)
Block-7: Dense layer with Relu actication
Batch size = 512
Output shape: (512, 1, 319)
Input shape: (512, 1, 128)
Block-6: Layer normalization
Batch size = 512
Output shape: (512, 1, 319)
Input shape: (512, 1, 319)
Block-5: Multi-head attention layer
Input shape: (512, 1, 319)
Batch size = 512
Input shape: (512, 1, 319)
Block-4: Reshape layer
Batch size = 512
Output shape: (512, 1, 319)
Input shape: (512, 319)
Block-3: Concatenation layer
Batch size = 512
Output shape: (512, 319)
Categorical embedding: (512, 19 X 16 = 304)
Numerical inputs: (512, 15) (since num_numerical = 15)
Block-2: Categorical input layer
Batch size = 512
Number of categorical features: 19
Output shape: (512, 16) for each categorical feature (embedding dim = 16)
Input shape: (512, 16) for each categorical feature
Block-1: Numerical input layer
Batch size = 512
Output shape: (512, 20)
Input shape: (512, 20) (num_numerical = 15)

Fig. 3. Overview of the proposed Tab_transformer model.

feedforward layers to introduce non-linearity. In the output layer, the activation function is sigmoid for binary classification tasks and softmax for multi-class classification tasks. During training, we employed the Adam optimizer with a learning rate of 0.0000001. The input shape for the numerical features is (batch_size, input_dim), where input_dim corresponds to the number of numerical features selected from the dataset. For categorical features, we encode each categorical feature as an integer index before passing it through the embedding layers, resulting in an input shape of (batch_size, 1) for each categorical column.

The multi-head attention mechanism has four attention heads. This functionality allows the model to focus on multiple aspects of the data simultaneously, learning complex feature relationships. The feed-forward layers, which process the output of the attention mechanism, set the feed-forward dimension (ff_dim) to 128 units. The model consists of 1 layer of multi-head attention followed by feed-forward layers, with a dropout rate set to 0.2 to help prevent overfitting during training. To further reduce overfitting, we apply L2 regularization with a regularization strength of 0.01 to the weights in the feed-forward layers. To fine-tune the model, we set the batch size during training to 512 and the learning rate for the Adam optimizer to an extremely low value of 0.0000001. For binary classification tasks, we use binary cross-entropy as the loss function for training, while we typically use categorical cross-entropy for multi-class classification problems. The optimizer trains the model for 40 epochs, iteratively adjusting the weights to minimize the loss. For binary classification tasks, binary cross-entropy is used as the loss function. Table 4 summarizes the parameters, and their corresponding values used for the proposed Tab_transformer model.

Hyperparameter tuning was performed exclusively within the training folds using nested cross-validation. A separate grid search was conducted for each outer fold without referencing the test data. Furthermore, early stopping was employed based on validation loss during inner training, further safeguarding against overfitting. These steps collectively ensured a rigorous and fair evaluation of model performance.

2.4. Analyzing confounding factors in the modeling process

This study analyzed confounding factors, such as patient age at treatment and the number of previous IVF cycles, to evaluate their impact on model performance for the best-performing AI model. Patient ages are categorized into five subgroups across different age brackets: 18–25, 26–30, 31–35, 36–40, and 41–45 years. Similarly, the number of previous IVF cycles is divided into five categories (0, 1, 2, 3, and 4+ cycles), and computed metrics are used to assess the adaptability of the model for each category. A robust validation framework is employed to further validate the model's reliability. This framework included various experimental scenarios: (1) a baseline condition using the original dataset and (2) outlier removal by excluding extreme values for factors such as "Patient Age at Treatment" and "Previous IVF Cycles." The SMOTE algorithm was used to even out the data, which fixed the problem of class imbalance. Gaussian noise was added to simulate real-world variation, with situations ranging from moderate noise (σ=0.05) to high noise (σ=0.15). Additionally, an explainable AI technique, SHAP, is utilized to enhance model interpretability by quantifying the contribution of each feature to predictions and offering information about the importance of key factors in the decision-making process.

Table 4
The information of all key parameters used for the proposed Tab_transformer-based model and their respective values used in this study.

Parameter	Value
Activation Function (Feedforward)	ReLU
Activation Function (Output Layer)	Sigmoid (binary), Softmax (multi-class)
Optimizer	Adam
Learning Rate	0.0000001
Input Shape (Numerical Features)	(batch_size, input_dim)
Input Shape (Categorical Features)	(batch_size, 1)
Input Dimension (input_dim)	Number of selected numerical features
Number of Heads (num_heads)	4
Feed-Forward Dimension (ff_dim)	128 units
Number of Layers (num_layers)	1 layer of multi-head attention + feed-forward layers
Dropout Rate (dropout_rate)	0.2
L2 Regularization (l2_reg)	0.01
Batch Size (batch_size)	512
Loss Function	Binary cross-entropy (binary classification), categorical cross-entropy (multi-class classification)
Number of Epochs	40

2.5. Explainable AI through SHAP: feature importance and clinical alignment

The SHAP method was chosen because it has a strong theoretical base in cooperative game theory. This advantage makes it an important tool for making complex machine learning models understandable both globally and locally. The SHAP approach was chosen because of its solid mathematical underpinnings in cooperative attribution modeling [30]. Assigning a Shapley value to each feature, which measures its contribution to individual predictions, improves the interpretability of the model. To ensure that importance scores are distributed fairly among all inputs, SHAP calculates these values by methodically comparing model outputs with and without each feature. Both local interpretability (explaining specific predictions) and global interpretability (identifying the most important features throughout the dataset) are made possible by this method. By ranking feature importance, SHAP helps make AI predictions clearer and aligns them with clinical reasoning by showing the key factors that affect model decisions [31].

2.6. Robustness assessment

To account for the real-world variability in IVF data, a systematic preprocessing strategy was implemented to guarantee model reliability and robustness. This strategy included Gaussian noise addition, outlier removal, and class balancing. The interquartile range (IQR) method was employed to identify and eliminate extreme values. Patients who were either younger than 18 or older than 45 were excluded, as they are outside the typical range of IVF treatment. Furthermore, the rarity of cases with >15 IVF cycles may have distorted the model's performance, leading to their exclusion. We implemented SMOTE to balance class distributions and ensure more equitable model training. Lastly, to test how well the model works when faced with real-world changes in data, two types of Gaussian noise were added: a moderate noise (σ = 0.05) to mimic small changes in clinical data and a high noise (σ = 0.15) to see how the model performs with larger changes. Table 5 summarizes the preprocessing steps and their corresponding experimental conditions.

To clarify, Gaussian noise was added exclusively to the training data, not to the validation or test sets. This decision was made to prevent data leakage and ensure the integrity of model evaluation. The rationale for selecting σ = 0.05 and σ = 0.15 was twofold:1) Clinical Realism: These values simulate realistic clinical variability and minor input inaccuracies in features such as patient age at treatment and Previous IVF Cycles. For example, slight variations in reported patient age or cycle counts may occur due to rounding, reporting delays, or data entry inconsistencies. Based observed data distributions, σ = 0.05 represents a moderate level of clinical uncertainty, while σ = 0.15 simulates more extreme cases, such as recording errors or borderline patient eligibility. 2) Robustness Testing: To check how strong our model is, we used these two levels of

Table 5
An overview of data preprocessing techniques for noise addition and outlier removal.

Action	Feature Affected	Details	Reason
Noise Addition	Patient Age at Treatment	- Gaussian noise added with a mean of 0 and: • Moderate Noise (σ = 0.05) • High Noise (σ = 0.15)	- Simulates slight inaccuracies (moderate noise) and extreme data perturbations (high noise). - Tests model robustness under imperfect data conditions.
	Previous IVF Cycles	- Gaussian noise added with a mean of 0 and: Moderate Noise (σ = 0.05) High Noise (σ = 0.15)	- Mimics variability in clinical reporting and tests resilience of the model predictions.
Outlier Removal	Patient Age at Treatment	- Removed ages below 18 years or above 45 years. - Values outside the interquartile range (IQR) were flagged and removed.	- Ages below or above the typical range for IVF treatments are unrealistic or non-clinical.
	Previous IVF Cycles	- Removed cases with IVF cycle counts above **15.** - IQR-based outliers identified and removed.	- Extremely high IVF cycle counts represent rare cases, which could skew the training data.

changes to see how it performs in noisier situations. The findings of this sensitivity analysis demonstrated that σ = 0.05 resulted in minor modifications that mirrored mild real-world noise, such as minor variations in patient age or number of treatments. A stress-test level of σ = 0.15 was used to capture larger perturbations while maintaining significant input structure. These two levels, which provide a useful compromise between realism and robustness assessment, were chosen as representative thresholds for moderate and high noise. Higher values (e.g., σ = 0.20 or above) significantly distorted feature distributions and decreased model reliability, whereas lower values (e.g., σ = 0.01) had no effect on performance. Thus, we deduced that the best values for evaluating model stability under realistic input were σ = 0.05 and σ = 0.15.

These procedures were integrated within a broader preprocessing pipeline that also included outlier removal using IQR filtering and SMOTE-based class balancing, further enhanced model robustness. All modifications were carefully limited to the training data to ensure fair and unbiased evaluation.

2.7. Performance evaluation metrics

Computational resources used in this study included an Intel(R) Core (TM) i7–10,700 K CPU running at 3.80 GHz, 32 GB of RAM, and an NVIDIA GeForce RTX 3080 GPU with 10 GB of VRAM. This hardware setup enabled efficient implementation of PSO and the transformer-based model, ensuring rapid experimentation and testing. We evaluated our machine learning models using several metrics, each providing distinct insights into model performance. These metrics include accuracy, precision, recall, and F1-score, which are defined as follows:

Accuracy: Accuracy measures the proportion of correctly predicted instances out of the total instances. It is calculated as:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{5}$$

Where:
TP: True Positives (correctly predicted positive cases)
TN: True Negatives (correctly predicted negative cases)
FP: False Positives (incorrectly predicted positive cases)
FN: False Negatives (incorrectly predicted negative cases)

Precision: Precision quantifies the proportion of correctly predicted positive cases out of all predicted positives. It is defined as:

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

Recall: Recall, also known as sensitivity or true positive rate, measures the proportion of actual positives correctly identified by the model. It is calculated as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

F1-Score: The F1-score is the means of precision and recall, offering a single metric that balances the two. It is calculated as:

$$F_{1}\_score = 2 \times \frac{precision \times recall}{precision + recall} \tag{8}$$

By employing these evaluation metrics, we obtained a comprehensive understanding of the models' strengths and weaknesses [32]. For all experiments, 10-fold cross-validation was used to reduce overfitting and to ensure that the model generalizes well to unseen data, and all reported evaluation metrics represent the average performance across the 10 cross-validation folds.

3. Results

3.1. Classification results

Table 6 presents the validation results of eight classification models (Section 2.1) designed to predict live birth success in IVF. The performance of each model is reported by five performance metrics including, accuracy, precision, recall, F1-score, and AUC.

Our results show that the PCA+Decision Tree model shows the least performance across all five-performance metrics, with the recall value as the lowest value, indicating that the model misses some truly positive situations. When using the Random Forest (RF) model, especially in combination with Particle Swarm Optimization (PSO), performance improves in most metrics compared to when the Decision Tree model was used. We observed that the combination of both feature selection methods, especially the PSO and transformer-based classifiers, outperformed the classification performance compared with traditional machine learning classifiers such as RF and decision tree. The PSO + Tab_transformer-based model has achieved the best accuracy (97 %), precision (95.2 %), recall (96.1 %), F1-score (95.6 %), and AUC (98.4 %). This model outperformed the other seven models in all performance metrics. The visualization of all evaluation results for eight different methods compared to each other is shown in Fig. 4.

Fig. 5 shows the attention weight matrix extracted from the trained Tab_transformer for a representative test case with a positive outcome (live birth success). The matrix shows how the model's attention heads dynamically assigned importance to input features based on contextual interactions between them. The darker red squares along the diagonal and upper-left part show that the model paid a lot of attention to important features like "tubal disease," "fresh transfer," "endometriosis," and "partner sperm morphology." Both the attention mechanism and the SHAP analysis consistently show that these features are very important, confirming their role in predicting successful IVF outcomes. The lower portion of the matrix (represented with faded colors and ellipses "…") denotes additional features with minimal attention contribution and is truncated for clarity in visualization. The color scale reflects the magnitude of attention weights, with warmer colors indicating a greater influence of one feature over another during the decision process. This visualization demonstrates that the Tab_transformer not only identifies individual important features but also models complex dependencies between them using its attention layers, which goes beyond what linear models or SHAP summary plots can capture.

Table 6
Performance evaluation of all designed classification models on the (2010–2018) dataset for predicting live birth success in IVF. PCA: Principal Component Analysis, PSO: Particle Swarm Optimization, RF: Random Forest.

The number of experiments	Classification Method	Accuracy	Precision	Recall	F1-Score	AUC
1	PCA + RF	90.2 %	91 %	90 %	92.4 %	93
2	PCA + decision tree	89.8 %	88 %	86 %	90.4 %	90
3	PSO + RF	90.5 %	91.9 %	89.6 %	93.8 %	94
4	PSO + decision tree	88.9 %	89 %	90.6 %	93.5 %	93.4
5	PCA + Transformer-based model	92.5 %	92.7 %	92.4 %	94.5 %	95
6	PCA + Tab_transformer-based model	94.1 %	95.2 %	90.1 %	91.3 %	94.7
7	PSO + Transformer-based model	96.3 %	95.7 %	96.4 %	94.8 %	98 %
8	PSO + Tab_transformer-based model	97 %	95.2 %	96.1 %	95.6 %	98.4 %

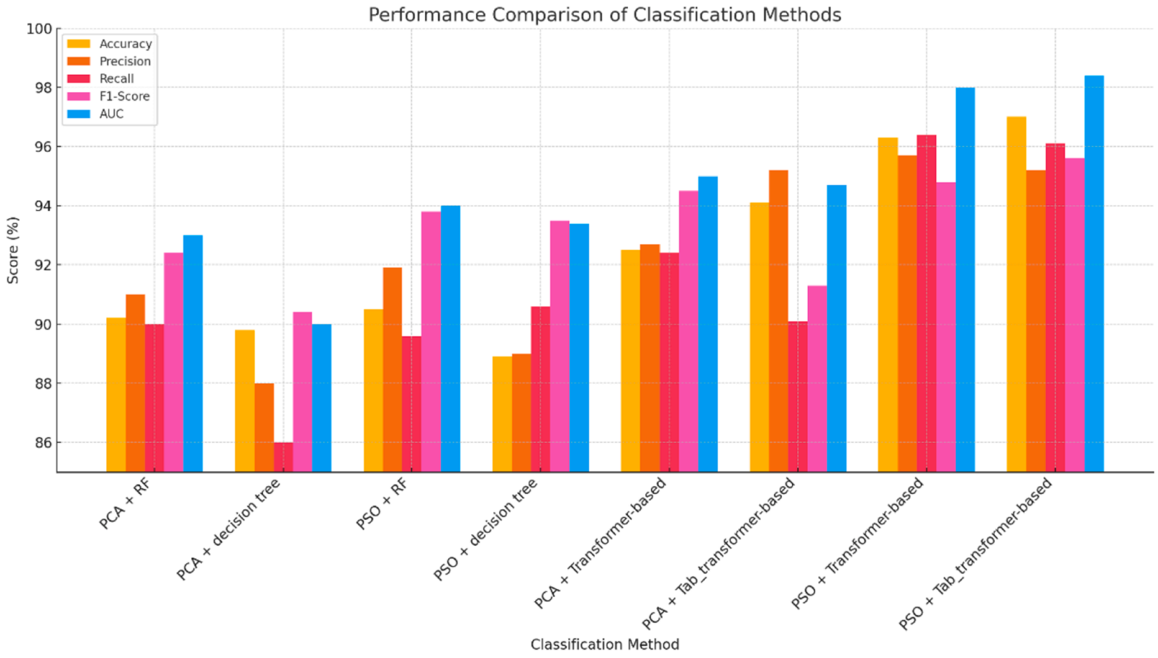

Fig. 4. Visualization of the binary classification results in this paper for all eight experiments and using five performance metrics including accuracy, precision, recall, F1-score and AUC.

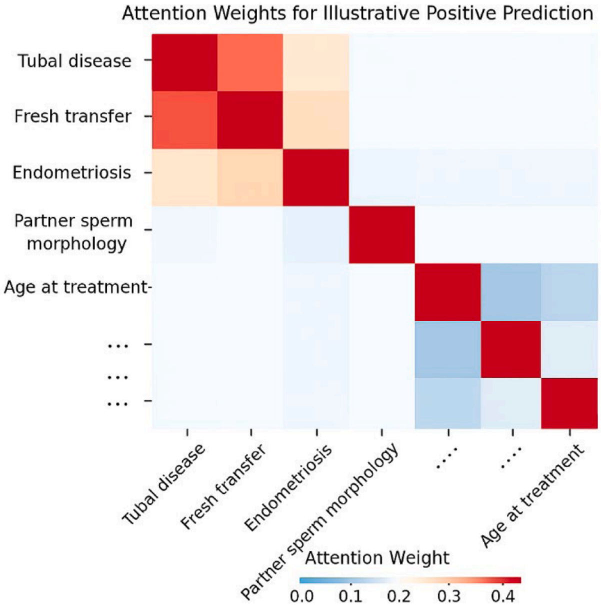

Fig. 5. Attention weight matrix for a positive IVF outcome (live birth).

3.2. *Analyzing details of the best performing model*

As mentioned in Section 3.1, the best prediction results were achieved by the PSO + Tab_transformer-based model. The outcome of PSO feature selection using this model is a reduced set of 38 features, selected based on their relevance to birth prediction and their ability to improve model performance. Each feature represents a critical aspect of the IVF dataset, categorized into groups such as Infertility Cause, Procedural Detail, Patient History, and Outcome. The selected features and their explanation are described in Table 7.

The training and validation performance of the Tab_transformer model when combined with PSO, our top-performing model in this study, is also shown in Fig. 6.

Fig. 6 shows that over 40 epochs, the Tab_transformer model showed steady improvements in both training and validation metrics. Throughout the training process, both training and validation loss decreased gradually and stayed closely aligned, as seen in Fig. 6's left panel. The close tracking between losses effectively minimizes overfitting while maintaining strong generalization to unknown data. The validation loss by the last epoch was 0.0716, which was very close to the training loss of 0.0798. The right panel displays the training and validation accuracy, demonstrating a steady improvement and convergence to >97 % by the end of the epoch. These outcomes show how well the Tab_transformer generalizes and how robust its learning ability is.

Table 7
Explanation of each of the selected features using PSO.

Feature Name	Category	Description
Cause of infertility - tubal disease	Infertility Cause	Indicates whether the patient's infertility is due to tubal disease.
Cause of infertility - partner sperm immunological factors	Infertility Cause	Refers to immunological issues with the partner's sperm that may affect fertility.
Cause of infertility - partner sperm morphology	Infertility Cause	Indicates abnormalities in sperm shape that may contribute to infertility.
Cause of infertility - endometriosis	Infertility Cause	Indicates whether the patient's infertility is caused by endometriosis.
Cause of infertility - female factors	Infertility Cause	Covers a range of female infertility factors not otherwise specified.
Cause of infertility - ovulatory disorder	Infertility Cause	Infertility is due to ovulatory disorders in the patient.
Cause of infertility - patient unexplained	Infertility Cause	Referring to cases where the cause of infertility is unknown or unexplained.
Date of egg mixing	Procedural Detail	The date when eggs were mixed with sperm during the IVF procedure.
Date of embryo thawing	Procedural Detail	The date when frozen embryos were thawed for use in the IVF cycle.
Eggs mixed with donor sperm	Procedural Detail	Indicates whether donor sperm was used to mix with eggs during the IVF cycle.
Eggs mixed with partner sperm	Procedural Detail	It indicates whether the partner's sperm was used to mix with eggs during the IVF cycle.
Eggs thawed	Procedural Detail	Indicates the number of eggs that were thawed during the IVF procedure.
Embryos transferred	Procedural Detail	The number of embryos transferred to the uterus during the IVF cycle.
Embryos transferred from eggs micro-injected	Procedural Detail	Indicates embryos transferred that were developed from micro-injected eggs.
Embryos stored for use by patient	Procedural Detail	The number of embryos stored for future use by the patient.
Frozen cycle	Procedural Detail	Indicates whether the cycle involved frozen embryos.
Stimulation used	Procedural Detail	Indicates the type of hormonal stimulation protocol used during the cycle.
Total embryos created	Procedural Detail	The total number of embryos created during the IVF cycle.
Total number of previous treatments, both IVF and DI	Patient History	Total number of previous treatments (IVF or donor insemination) conducted at the same clinic.
Total number of previous IVR pregnancy	Patient History	Total number of previous pregnancies provided through in vitro fertilization (IVF).
Total number of previous DI pregnancy	Patient History	Total number of previous pregnancies achieved through either IVF or donor insemination.
Total number of previous pregnancies- IVF and DI	Patient History	Combination of total number of pregnancies provided through either IVF or donor insemination (DI).
Donated embryo	Procedural Detail	Indicates whether a donated embryo was used during the IVF cycle.
Total number of previous DI cycles	Patient History	Total number of donor insemination cycles performed prior to the current treatment.
Type of infertility - female secondary	Infertility Cause	Indicates secondary infertility in the female patient.
Type of infertility - male primary	Infertility Cause	Indicates primary infertility in the male partner.
PGD (Preimplantation Genetic Diagnosis)	Procedural Detail	Indicates whether PGD was performed to screen embryos for genetic abnormalities.
PGT-A treatment	Procedural Detail	Indicates whether Preimplantation Genetic Testing for Aneuploidy (PGT-A) was used.
PGT-M treatment	Procedural Detail	Indicates whether Preimplantation Genetic Testing for Monogenic disorders (PGT-M) was used.
Total eggs mixed	Procedural Detail	The total number of eggs mixed with sperm during the IVF cycle.
Fresh eggs stored	Procedural Detail	Indicates whether fresh eggs were stored for future use.
Fresh eggs stored (0/1)	Procedural Detail	Binary indicator for whether fresh eggs were stored.
Patient Age at treatment	Patient Demographics	Age of the patient at the time of treatment, a known factor influencing success.
Previous IVF cycles	Patient History	Number of IVF cycles previously undergone by the patient.
Total number of IVF cycles	Patient History	Cumulative number of IVF attempts by the patient.
Embryos transferred during fresh cycle	Procedural Detail	Number of embryos transferred during a fresh (non-frozen) IVF cycle.
Number of embryos developed from ICSI	Procedural Detail	Total embryos that successfully developed after Intracytoplasmic Sperm Injection.
Total number of live births - conceived through IVF or DI	Outcome	Total number of live births achieved from either IVF or donor insemination cycles.

3.3. Performance evaluation across subgroups

This study looked at the performance of the modeling of different groups of patients based on their age and the number of IVF cycles they had previously. Table 8 shows how well the model did with people of different ages. The groups with the best results were those between the ages of 26 and 30 (accuracy: 96.8 %, AUC: 96.9 %) and 31 to 35 (accuracy: 95.8 %, AUC: 96.4 %), which shows how important these reproductive times are in a medical sense. The performance went down a little for the youngest (18–25) and oldest (41–45) groups, with accuracy rates of 94.9 % and 93.5 %, respectively. However, the model kept performing well across all age groups, proving its strong performance across different subgroups.

Table 9 shows how well different subgroups did based on previous IVF cycles. The subgroup with 4+ cycles had the best accuracy (96.6 %) and AUC (96.1 %), showing that the model can handle cases with more clinical data. Also, the 2-cycle and 3-cycle subgroups did very well, with 96.5 % and 99.3 % accuracy rates. The 0-cycle subgroup did a little worse (accuracy: 93.9 %, AUC: 93.7 %) Overall, while the model performed best for cases with long treatment histories, the model showed consistent reliability across all subgroups, proving its performance across all subgroups.

3.4. Impact of data preprocessing techniques on model performance

Table 10 assesses the model's performance under various conditions and investigates the effects of outlier removal, data balancing (SMOTE), and Gaussian noise addition on critical metrics, including accuracy, recall, F1-score, and AUC. These experimental conditions are consistent with the data preprocessing steps delineated in Table 5, which involved the removal of outliers using the IQR method, the balancing of the data, and the introduction of Gaussian noise at varying levels. The negative impact of extreme values on model performance was indicated by the increase in AUC to 97.5 % and the improvement in accuracy from 96.5 % to 97.2 % because of the removal of outliers. The addition of moderate

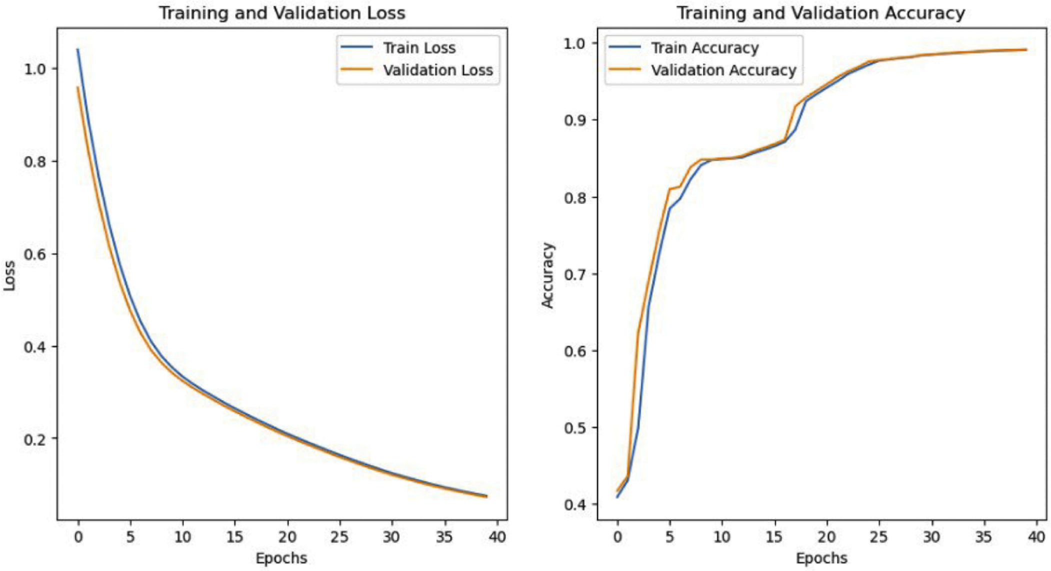

Fig. 6. Training and validation performance of the Tab_transformer model Over 40 Epochs.

Table 8
Subgroup metrics for patient age at treatment.

Age Group (Years)	Accuracy	Precision	Recall	F1-Score	AUC
18–25	94.9 %	93.2 %	94.1 %	95.0 %	94.7 %
26–30	96.8 %	96.6 %	95.9 %	95.7 %	96.9 %
31–35	95.8 %	95.0 %	95.5 %	94.6 %	96.4 %
36–40	94.2 %	93.5 %	93.0 %	94.7 %	94.3 %
41–45	93.5 %	92.1 %	92.9 %	92.5 %	92.7 %

Table 9
Subgroup metrics for previous IVF cycles.

IVF Cycles	Accuracy	Precision	Recall	F1-Score	AUC
0	93.9 %	94.3 %	95.0 %	93.1 %	93.7 %
1	95.8 %	95.5 %	94.9 %	94.7 %	94.4 %
2	96.5 %	96.3 %	95.6 %	95.4 %	95.2 %
3	96.3 %	96.0 %	95.5 %	95.2 %	95.6 %
4+	96.6 %	96.1 %	94.9 %	94.7 %	96.1 %

Table 10
Performance metrics across different experimental scenarios.

Metric	Baseline (No Changes)	Outlier Removed	Balanced Data	Moderate Noise	High Noise
Accuracy	96.5 %	97.2 %	97.1 %	95.2 %	93.0 %
Precision	96.3 %	96.9 %	96.8 %	94.8 %	92.6 %
Recall	96.7 %	97.3 %	97.3 %	95.4 %	93.2 %
F1-Score	96.5 %	97.1 %	97.0 %	95.1 %	92.9 %
AUC	96.8 %	97.5 %	97.4 %	95.5 %	93.1 %

Gaussian noise ($\sigma = 0.05$) had a minor impact (95.2 % accuracy, 95.5 % AUC), whereas high noise ($\sigma = 0.15$) resulted in a substantial decrease in accuracy to 93.0 %. These findings indicate that the model maintains a high predictive capability while experiencing some performance degradation under extreme noise conditions, despite remaining robust against minor variations.

3.5. SHAP analysis of feature importance

The SHAP feature importance values are illustrated in Fig. 7, which ranks key predictors according to their contribution to the model's decision-making process for IVF success prediction. Tubal disease, partner sperm issues, and sperm shape problems are the most important factors, followed by endometriosis, ovulation problems, and unexplained infertility, which match what is already known in medicine. Furthermore, predictability is influenced by embryo handling factors, including the date of egg mixing, embryo thawing, and the presence of donor or partner sperm in the eggs, which emphasizes the value of laboratory procedures. Stimulation type, frozen cycles, and embryo storage are additional pertinent predictors, which underscores the significance of treatment strategies. Overall, the SHAP analysis shows that the AI model is important for clinical use because it highlights key factors related to infertility and matches medical knowledge, making it easier to understand and trust decisions about fertility treatments.

The SHAP analysis was performed on a test set that was not used during training or validation. This ensures that the importance of rankings reflects the model's interpretability and behavior on previously unseen data, resulting in a strong and unbiased explanation of the model's decision-making procedure. After training the final model with the best features selected by PSO, we calculated SHAP values on the test set to see how each feature influenced the model's predictions about live birth outcomes. As explained in Section 3.5 and shown in Fig. 7 and Table 11, the highest-ranked features—like "tubal disease," "partner sperm immunological factors," and "embryo handling procedures"—were not only important but also made sense in a clinical context, matching what we already know in reproductive medicine.

Table 11 shows the SHAP feature importance rankings, which reveal the most important variables in predicting live birth outcomes.

Features associated with infertility causes, particularly "tubal disease" (0.088), "partner sperm immunological factors" (0.084), and "sperm morphology" (0.081), dominate the top rankings. Furthermore, procedural details such as "date of egg mixing" and "embryo thawing" make a significant contribution, highlighting the model's reliance on both clinical history and treatment process variables.

Fig. 8 depicts a concrete example of a negative prediction case in which the model predicted a low chance of a live birth after an IVF cycle. This visualization depicts how individual features influenced the model's output by displaying the cumulative effect of each feature's SHAP value on the overall prediction. The prediction starts with a baseline probability (usually the average prediction across all patients, around 0.5), and features are added or removed based on their contribution. In this case, several clinical factors—most notably tubal disease, the lack of a fresh transfer, and the presence of endometriosis—all had a

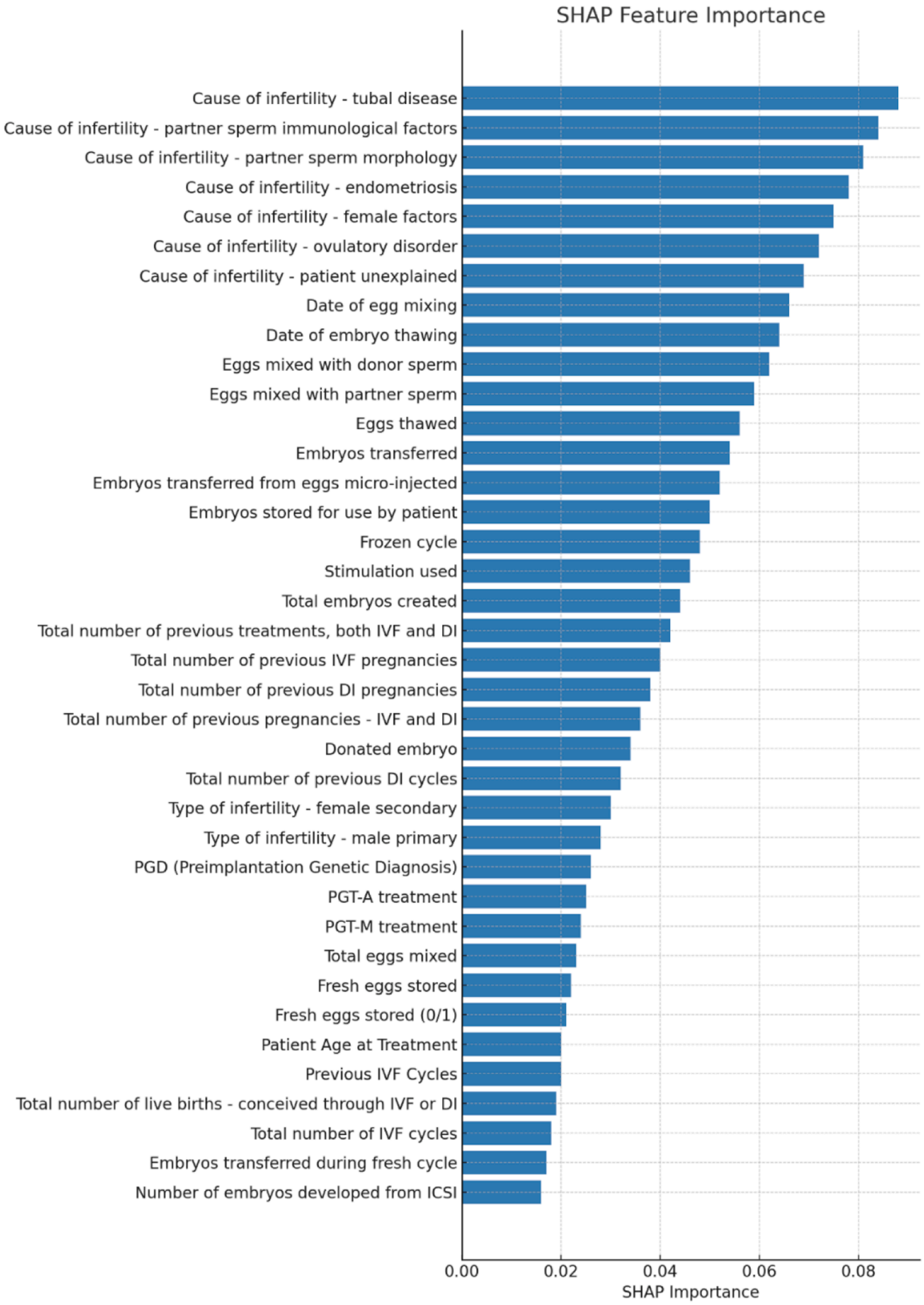

Fig. 7. SHAP feature importance for selected IVF predictors.

negative impact, decreasing the predicted likelihood of a live birth. Large red bars reflect these factors, causing the prediction to fall. Although partner sperm morphology made a small positive contribution (green bar), it was insufficient to counteract the overall negative influence. The patient's age at treatment had a minor negative impact. Overall, these effects produced a final predicted probability of approximately 0.23, prompting the model to classify this instance as a negative outcome. This chart makes it clearer and helps doctors understand the specific reasons behind the model's decision, showing how important SHAP explanations are in real-life fertility predictions.

Fig. 9 depicts a positive prediction case in which the model predicted a high chance of a live birth after IVF treatment. The plot begins with a baseline prediction, which typically represents the average model output across the population, it then shows how individual features shift the prediction upward or downward using their SHAP values. In this case, five features had a significant impact on the model's decision. The presence of tubal disease, a fresh embryo transfer, endometriosis, and favorable sperm morphology (as evidenced by partner sperm morphology) all helped improve prediction. These factors are represented by a green bar, which incrementally increases the predicted probability. Although "age at treatment" had a slight negative contribution (represented by a small red bar), it was insignificant when compared to the other features' strong positive influences. The cumulative impact of these factors increased the prediction from the baseline to a final probability of around 0.88, allowing us to confidently classify the outcome as a successful live birth. This type of SHAP visualization provides an understandable breakdown of the model's reasoning, making it ideal for clinical decision support. By clearly identifying which

Table 11
SHAP feature importance data.

Feature	SHAP Importance
Cause of infertility - tubal disease	0.088
Cause of infertility - partner sperm immunological factors	0.084
Cause of infertility - partner sperm morphology	0.081
Cause of infertility - endometriosis	0.078
Cause of infertility - female factors	0.075
Cause of infertility - ovulatory disorder	0.072
Cause of infertility - patient unexplained	0.069
Date of egg mixing	0.066
Date of embryo thawing	0.064
Eggs mixed with donor sperm	0.062
Eggs mixed with partner sperm	0.059
Eggs thawed	0.056
Embryos transferred	0.054
Embryos transferred from eggs micro-injected	0.052
Embryos stored for use by patient	0.050
Frozen cycle	0.048
Stimulation used	0.046
Total embryos created	0.044
Total number of previous treatments, both IVF and DI	0.042
Total number of previous IVF pregnancies	0.040
Total number of previous DI pregnancies	0.038
Total number of previous pregnancies - IVF and DI	0.036
Donated embryo	0.034
Total number of previous DI cycles	0.032
Type of infertility - female secondary	0.030
Type of infertility - male primary	0.028
PGD (Preimplantation Genetic Diagnosis)	0.026
PGT-A treatment	0.025
PGT-M treatment	0.024
Total eggs mixed	0.023
Fresh eggs stored	0.022
Fresh eggs stored (0/1)	0.021
Patient Age at Treatment	0.020
Previous IVF Cycles	0.020
Total number of live births - conceived through IVF or DI	0.019
Total number of IVF cycles	0.018
Embryos transferred during fresh cycle	0.017
Number of embryos developed from ICSI	0.016

clinical variables most contributed to the predicted outcome, it provides clinicians with actionable insights for tailoring treatment plans.

3.6. *Temporal feature interpretation*

Date of embryo thawing" and "Date of egg mixing" were identified as moderately important features. The model can detect latent patterns related to the timing and manner of treatments by incorporating these precise procedural dates. By encoding the temporal alignment of clinical interventions—which can indirectly influence biological processes—these features play a significant role in model predictions. For example, the "Date of embryo thawing" indicates the exact day a frozen embryo was prepared for transfer. Even small deviations in thawing timing can have clinical implications, particularly in relation to endometrial receptivity. Similarly, the "Date of egg mixing"—the day on which oocytes and sperm are combined—can influence the timing of fertilization and subsequent developmental kinetics, both of which are associated with embryo quality. These temporal variables function as anchors, capturing the interaction between laboratory procedures and biological readiness during treatment, despite not being biomarkers themselves. Beyond their biological implications, such temporal features also act as indirect proxies for cohort effects and institutional protocols. IVF procedures evolve over time, and different clinics or time periods may follow distinct freezing techniques, laboratory practices, embryo handling methods, and regulatory standard [33].

3.7. *Effectiveness of feature selection methods*

To understand how each group of features selected by PSO affects the results, we analyzed our internal dataset from 2010 to 2018 by removing

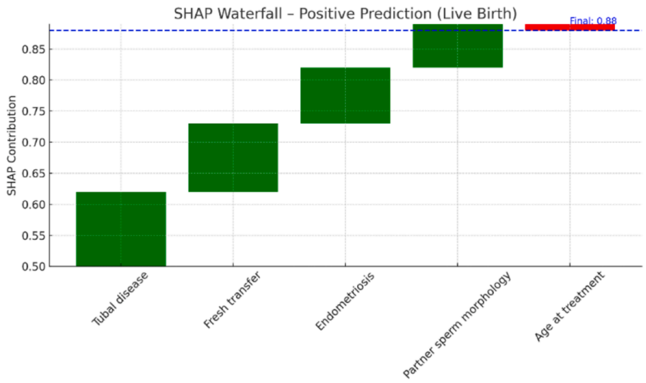

Fig. 9. The SHAP waterfall plot shows the top contributing features for the positive prediction (live birth).

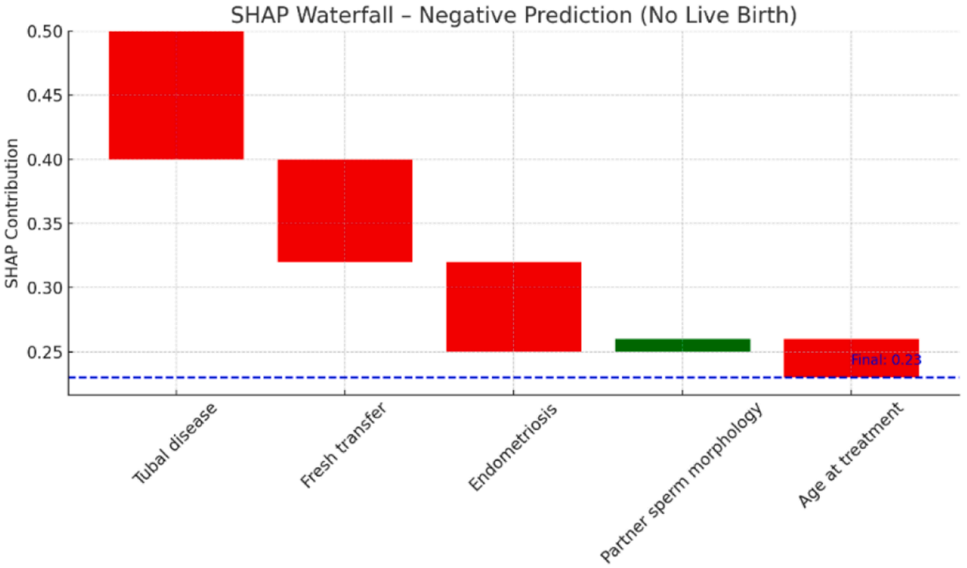

Fig. 8. The SHAP waterfall plot shows the top contributing features for the negative prediction (no live birth).

each of the top seven features ranked by SHAP (which are all important causes of infertility) and checking how this change impacted the model's performance using the PSO + Tab_Transformer method. The results are shown in Table 12. The findings are summarized in Table 12.

These findings indicate that if we take away any of the most important features, the performance decreases, especially for tubal disease, which shows how crucial they are for the model's decisions and emphasizes the clinical significance of the PSO-selected features as shown in the SHAP analysis.

To validate PSO's effectiveness, we compared it to two conventional feature selection techniques: mutual information (MI) and LASSO (L1-regularized logistic regression). Each method was used to determine the best feature subset, and identical Tab_transformer models were trained and tested with the same train/test splits. The comparative results are presented in Table 13.

These findings highlight PSO's superior performance in selecting features that align with domain expertise while also resulting in higher model accuracy and generalizability. MI and LASSO, on the other hand, had lower discriminative power, as evidenced by low AUC scores and imbalanced precision-recall behavior.

3.8. Evaluation of the proposed model on an external dataset

Table 14 shows the performance of the proposed PSO + Tab_transformer-based model on both internal (2010–2018) and external (2005–2009) datasets: internal dataset from 2010 to 2018 and another from 2005 to 2009. The results show that the model can generalize well across dataset from different time points. On the internal dataset, the model performed well across all metrics, with an accuracy of 97 % and an AUC of 98.4 %, indicating high classification capability. On the external dataset, which included previously unseen data, the model performed almost equally well, with 96.1 % accuracy and 97.2 % AUC. This consistency across datasets supports our proposed pipeline's reliability and stability in predicting IVF live birth outcomes across different temporal cohorts.

4. Discussion and conclusion

In this study, we explored various machine learning and deep learning models with a combination of two feature selection techniques for predicting live birth success in IVF using the comprehensive HFEA dataset. Our study could achieve a very high performance for five different evaluation metrics by utilizing PSO for feature selection combined with Tab_transformer, an advanced deep learning model. The AI pipeline is designed by integrating PSO for feature selection, the Tab_transformer for tabular data classification and attention mechanism, balanced datasets to address class imbalance, cross-validation to prevent overfitting, and robust regularization techniques to enhance model stability. The proposed model then has the potential to deal with common problems like overfitting, inconsistent patient data, and uneven datasets, thus showing promise for a clinically applicable tool for predicting live birth success in IVF. With an accuracy of 97 %, precision of 95.2 %, recall of 96.1 %, F1-score of 95.6 %, and AUC of 98.4, the PSO + Tab_transformer-based model produced exceptional results, making it the most successful model for forecasting the success of live births. In contrast, the excellent accuracy and recall offered by the transformer-based techniques could not be achieved by models like PCA + Decision Tree and PCA + Random Forest, despite their effectiveness. These findings demonstrate that deep learning-based transformer models can enhance the prediction of IVF outcomes. These deep learning models provide an advantage over conventional classifiers in terms of their ability to recognize relevant features and to capture intricate relationships within the data.

Table 12
sensitivity analysis of top SHAP-Ranked feature groups on model performance (PSO + Tab_Transformer model).

Feature group removed	Accuracy (%)	AUC (%)	F1-Score (%)
Full PSO feature set (no removal)	97	98.4	95.6
Cause: tubal disease	94.4	96.1	93.2
Cause: partner sperm immunological factors	95.2	96.9	94.1
Cause: sperm morphology issues	95.0	96.5	93.8
Cause: endometriosis	95.4	96.7	94.2
Cause: female factors	95.7	96.8	94.6
Cause: ovulatory disorder	95.8	96.9	94.7
Cause: unexplained infertility	96.1	97.2	95.0

Table 13
Comparison of PSO, Mutual Information, and LASSO Feature Selection Methods.

Feature Selection Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
PSO (Proposed)	97	95.2	96.1	95.6	98.4
Mutual information	65.5	65.6	77.9	71.1	70.5
LASSO	65.0	64.9	79.9	71.5	62.8

Particularly, the Tab_transformer offers several advantages over traditional models. First, it can efficiently handle the high-cardinality categorical features by learning embedding instead of one-hot encoding, which can lead to representations that are sparse and high-dimensional. Second, it captures complex interactions between features using self-attention, which traditional models might overlook. Third, it reduces the need for extensive manual feature engineering, enabling end-to-end learning directly from raw tabular data. This model is particularly effective in domains where meaningful relationships between features play a crucial role. When compared to traditional machine learning models, the Tab_transformer excels with larger datasets and high-cardinality features, offering state-of-the-art performance. Historically, the challenges in IVF outcome prediction also including live birth prediction included limited application of advanced deep learning models for tabular data. The application of these advanced deep learning techniques is explored in this study for the first time. We also have surveyed some of previous efforts that used the HFEA dataset for the classification of live birth success as a binary outcome (success/failure), similar goal to our study (Table 14).

SHAP analysis was implemented to enhance the interpretability of AI-driven predictions and to gain a more profound understanding of the model's decision-making process. This analysis uncovered the primary features that influence model predictions, thereby improving clinical relevance and transparency. We carefully tested the PSO-enhanced Tab_transformer model to ensure it was strong and dependable by using various methods to change the data and analyze other influencing factors. The model showed it can work well for different groups of patients because it performed consistently well, even when we divided the data into smaller groups based on age, gender, and health conditions.

These results confirm that the PSO-enhanced Tab_transformer is a dependable tool in clinical decision-making across a variety of scenarios, as it is not only highly accurate but also resilient to data variability.

AI models that predict IVF live birth outcomes could improve clinical workflows and patient care. A clinical decision support tool (CDST) can use the model to help fertility specialists calculate treatment success rates based on demographic, clinical, and embryological factors. Patients can have transparent, data-driven consultations, optimized embryo selection, and customized treatment protocols. Artificial intelligence can help patients make informed decisions by providing probabilistic success rates, promoting informed consent, and setting realistic expectations. AI in high-risk reproductive medicine raises ethical concerns despite benefits. First, data bias and algorithmic fairness must be addressed to reduce health disparities, especially when models are trained on small or non-representative datasets. Second,

Table 14
Proposed method's results on both internal and external datasets.

Proposed method	Training/Internal dataset (2010–2018)					Test/External dataset (2005–2009)				
	Accuracy	Precision	Recall	F1-Score	AUC	Accuracy	Precision	Recall	F1-Score	AUC
PSO + Tab_transformer-based model	97 %	95.2 %	96.1 %	95.6 %	98.4 %	96.1 %	94.3 %	95.1 %	94.7 %	97.2 %

clinicians and patients must understand AI-derived predictions, especially in emotionally and financially sensitive settings like IVF. Overusing opaque models can impair clinical judgment and patient autonomy. Third, low predicted success rates may harm patient choices or mental health, while high predictions that failure can cause false hope and distress. Finally, delicate reproductive health data requires strict ethical and legal standards, so data privacy, regulatory oversight, and legal accountability must be considered. AI in IVF has many clinical benefits, but it needs a strong ethical framework to ensure fairness, transparency, and patient-centered care.

As shown in Table 15, the results of this research surpass all previous studies utilizing HFEA datasets. Notably, compared to the study by Sadegh-Zadeh et al. [34] achieving an accuracy of 96.35 %, which adhered to same inclusion and exclusion criteria as used in our study, our study could improve upon these results with an accuracy of 97 %, precision of 95.2 %, recall of 96.1 %, F1-score of 95.6 %, and AUC of 98.4 %.

Among the related recent studies, only Sadegh-Zadeh et al. (2024) [34] reported their result on both an internal and external dataset related to two separate time points. They achieved a classification accuracy of 96.35 % using ensemble models such as AdaBoost and LogitBoost on HFEA dataset (2010–2016) and validated their model on the other temporal dataset of HFEA from 2017–2018 and achieved an accuracy of 95.78 %. As an additional evaluation to the dataset used in this study (dataset from (2010–2018) as training and dataset (2005–2009) as test), we also trained and test our proposed models using the HEEA dataset from the same time period selected by Sadeghzadeh et al. [34] (dataset from (2010–2016) as training and dataset from (2017–2018) as test) as test in order to enable direct comparison of the performance of our proposed method with their results. These results are summarized in Table 16.

The results show our model outperforms on multiple evaluation metrics across both internal and external datasets. On the internal dataset, our model achieved an accuracy of 97.3 %., precision of 95.6 %, F1-score of 96 %, and AUC of 98.9 %. Compared to the Sadeghzadeh model's 96.35 %, 87.29 %, 92.96 % and 98 %, respectively. On the external dataset, our method was better at generalizing, achieving higher accuracy of 96.5 % and almost similar AUC, while the Sadeghzadeh et al. results had a lower precision of 85.75 % and a lowerF1-score of 92.06 % compared to our methods. These findings highlight our pipeline's robustness, particularly in maintaining high precision and balanced performance when tested on different temporally distinct unseen data. By applying the same pipeline of our proposed method to the 2010–2016 dataset, we found that 33 out of the 38 features (88 %) were also selected in the 2010–2016 training dataset, indicating a high degree of stability. The five features not selected in the 2010–2016 subset were: PGT-M treatment, fresh eggs stored (binary), total number of previous DI cycles, total number of previous pregnancies (IVF and DI), and embryos stored for use by patient. These excluded features were among the lower-ranked in the full-dataset selection and may not have provided sufficient predictive value within the smaller, temporally constrained dataset. The high overlap between the two features selected using the two training datasets related to different time points supports our proposed method consistently identifies core predictive features while allowing for minor, context-specific adjustments.

For embryo selection, researchers have used embryo morphological grading systems that assess features like fragmentation, cell symmetry, and development stage [33]. However, these evaluations are subjective and limited in predictive power, with reported AUCs typically ranging from 0.60 to 0.70 [34]. Likewise, statistical models such as the Templeton score rely on simple features like patient age, infertility duration, and number of prior IVF attempts, achieving AUCs around 0.68–0.72 [35]. These models, though interpretable, lack the flexibility to capture complex, non-linear patterns in modern IVF datasets. In contrast, our method uses a PSO-based feature selection to identify key clinical variables and a Tab_transformer architecture to model interdependence using attention mechanisms, which enhances the ability to capture important patterns [35]. This enables our model to outperform prior approaches, achieving an AUC of 96.5 % on the internal dataset (2010–2018) and an AUC of 91.7 % on the external validation set (2005–2009). These results not only surpass those of conventional scoring systems but also exceed strong ML baselines (e.g., AdaBoost, LogitBoost with AUC ≈ 98 % on internal test set but not validated externally [32]). In summary, while traditional embryo grading and early predictive models have clinical familiarity, our method provides a quantitatively superior and generalizable approach, demonstrating enhanced robustness, interpretability (via SHAP), and clinical utility.

Future work could focus on integrating additional domain-specific features related to IVF treatments and patient characteristics to further improve model performance. Exploring the use of other advanced deep learning models, including those that account for sequential or temporal data, will be also explored for their potential to enhance prediction

Table 15
Related works Using the HFEA Dataset for Predicting Live Birth Success.

Study	Dataset Used	Key Features/Methods	Model Used	Performance Metrics
Zhang et al. [16]	57,558 NC-IVF cycles (2005–2016)	Patient demographics, hormonal profiles, cycle history, treatment outcomes; data balancing (SMOTE), SHAP, cross-validation	Artificial Neural Network (ANN)	F1-score: 70.87 %, AUC: 0.7939
Sadeghzadeh et al. [34]	495,630 IVF cycles (2010–2018)	Clinical and demographic data; temporal validation, feature normalization, interpretability	Ensemble models (AdaBoost, LogitBoost)	Accuracy: 96.35 % F1-score:92.96
McLernon et al. [17].	113,873 women, 184,269 cycles (1999–2008)	Multiple complete IVF/ICSI cycles; pre- and post-treatment analysis	Discrete-time Logistic Regression	C-index: 0.73 (pre-treatment), 0.72 (post-treatment)
Jones et al. [18]	93,495 women, 174,418 IVF cycles (1991–1998)	Focused on the likelihood of live birth success	Logistic Regression	AUC: 0.635
Sanders et al. [19]	190,010 IVF cycles (2016–2018)	Comparison of PGT-A and non-PGT-A cycles; odds ratios (ORs), descriptive statistics	Binary Logistic Regression	Focused on ORs instead of AUC
This Research	**2010–2018 IVF dataset**	**Advanced ML techniques, IVF-specific preprocessing**	**PSO + Tab_transformer**	Accuracy: 97 % precision 95.2 %, Recall 96.1 % F1-score :95.6 % AUC : 98.4 %.

Table 16
Comparison of the performance of our proposed model with study done by Sadeghzadeh et al. [34].

Model	Training/Internal dataset (2010–2016)					Test/External dataset (2017–2018)				
	Accuracy	Precision	Recall	F1-Score	AUC	Accuracy	Precision	Recall	F1-Score	AUC
The proposed model	97.3 %	95.6 %	96.4 %	96 %	98.9 %	96.5 %	94.7 %	95.8 %	95.2 %	97.9 %
Sadeghzadeh et al. [33] model	96.35 %	87.29 %	99.41 %	92.96 %	98 %	95.78 %	85.75 %	99.38 %	92.06 %	98 %

accuracy. Expanding the dataset to include more diverse populations and treatment types would help improve model generalizability and applicability in broader IVF contexts. In addition, expanding the dataset to include more diverse populations also including data from different geographic locations and treatment types will be explored to help further improve model generalizability and applicability in broader IVF contexts. In addition, expanding the dataset to include more diverse populations also including data from different geographic locations and treatment types will be explored to help further improve model generalizability and applicability in broader IVF contexts.'

Ethics statement

The UK Human Fertilisation and Embryology Authority (HFEA) database, which contains fully anonymised data from 2010 to 2018, was used in this study. Since there is no identifiable personal information in the dataset and no direct human experimentation is involved, ethical approval was not needed for this study. Data protection standards, institutional guidelines, and applicable laws were followed throughout the entire process.

Funding for this study

This study was funded by ACMIT COMET Module FFG project (FFG number: 879,733, application number: 39,955,962).

Data availability

The original contributions presented in the study are included in the supplementary material, further inquiries can be directed at the corresponding author. Additionally, The code used to implement the PSO-based feature selection and Tab_transformer model is publicly available at the following GitHub repository: https://github.com/arezooborji/IVF_PSO_TabTransformer.

CRediT authorship contribution statement

Arezoo Borji: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Hossam Haick:** Writing – review & editing, Conceptualization. **Birgit Pohn:** Investigation, Conceptualization. **Antonia Graf:** Investigation, Conceptualization. **Jana Zakall:** Investigation, Conceptualization. **S M Ragib Shahriar Islam:** Visualization, Investigation. **Gernot Kronreif:** Writing – review & editing, Supervision, Funding acquisition. **Daniel Kovatchki:** Investigation, Conceptualization. **Heinz Strohmer:** Investigation, Conceptualization. **Sepideh Hatamikia:** Writing – review & editing, Supervision, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that none of their personal relationships or known competing financial interests could have appeared to have influenced the work described in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2025.108979.

References

[1] Morse, K. (2022). ni ve rs ity of e to w n ve rs ity e to w, 213.
[2] L. Liu, H. Liang, J. Yang, F. Shen, J. Chen, L. Ao, Clinical data-based modeling of IVF live birth outcome and its application, Reproductive biology and endocrinology 22 (1) (2024) 1–12, https://doi.org/10.1186/s12958-024-01253-3.
[3] A. Uyar, Y. Sengul, A. Bener, Emerging technologies for improving embryo selection: a systematic review, Adv. Health Care Technol. (2015) 55, https://doi.org/10.2147/ahct.s71272.
[4] S. Hanassab, A. Abbara, A.C. Yeung, M. Voliotis, K. Tsaneva-Atanasova, T. W. Kelsey, W.S. Dhillo, The prospect of artificial intelligence to personalize assisted reproductive technology, NPJ. Digit. Med. 7 (1) (2024), https://doi.org/10.1038/s41746-024-01006-x.
[5] D.J. Patel, K. Chaudhari, N. Acharya, D. Shrivastava, S. Muneeba, Artificial intelligence in obstetrics and gynecology: transforming care and outcomes, Cureus. 16 (7) (2024), https://doi.org/10.7759/cureus.64725.
[6] K. Vassakis, E. Petrakis, I. Kopanakis, Big data analytics: applications, prospects and challenges, Lecture notes on data engineering and communications technologies 10 (2018) 3–20, https://doi.org/10.1007/978-3-319-67925-9_1.
[7] T.S. Chen, P.L. Kuo, T. Yu, M.H. Wu, IVF and obstetric outcomes among women of advanced maternal age (≥45 years) using donor eggs, Reprod. Biomed. Online 49 (4) (2024) 1–8, https://doi.org/10.1016/j.rbmo.2024.104291.
[8] S. Ueno, J. Berntsen, M. Ito, T. Okimura, K. Kato, Correlation between an annotation-free embryo scoring system based on deep learning and live birth/neonatal outcomes after single vitrified-warmed blastocyst transfer: a single-centre, large-cohort retrospective study, J. Assist. Reprod. Genet. 39 (9) (2022) 2089–2099, https://doi.org/10.1007/s10815-022-02562-5.
[9] G. Coticchio, C. Lagalla, M. Taggi, D. Cimadomo, L. Rienzi, Embryo multinucleation: detection, possible origins, and implications for treatment, Human reproduction 39 (11) (2024) 2392–2399, https://doi.org/10.1093/humrep/deae186.
[10] T. Bamford, C. Easter, S. Montgomery, R. Smith, R.K. Dhillon-Smith, A., …. Barrie, A. Coomarasamy, A comparison of 12 machine learning models developed to predict ploidy, using a morphokinetic meta-dataset of 8147 embryos, Human reproduction 38 (4) (2023) 569–581, https://doi.org/10.1093/humrep/dead034.
[11] A. Uyar, A. Bener, H.N. Ciray, Predictive modeling of implantation outcome in an in vitro fertilization setting, Medical decision making 35 (6) (2015) 714–725, https://doi.org/10.1177/0272989x14535984/.
[12] S. Giscard d'Estaing, E. Labrune, M. Forcellini, C. Edel, B. Salle, J. Lornage, M. Benchaib, A machine learning system with reinforcement capacity for predicting the fate of an ART embryo, Syst. Biol. Reprod. Med. 67 (1) (2021) 64–78, https://doi.org/10.1080/19396368.2020.1822953.
[13] B. Huang, S. Zheng, B. Ma, Y. Yang, S. Zhang, L. Jin, Using deep learning to predict the outcome of live birth from >10,000 embryo data, BMC. Pregnancy. ChildBirth 22 (1) (2022) 1–7, https://doi.org/10.1186/s12884-021-04373-5.
[14] V.S. Jiang, H. Kandula, P. Thirumalaraju, M.K. Kanakasabapathy, P. Cherouveim, I., …. Souter, H. Shafiee, The use of voting ensembles to improve the accuracy of deep neural networks as a non-invasive method to predict embryo ploidy status, J. Assist. Reprod. Genet. 40 (2) (2023) 301–308, https://doi.org/10.1007/s10815-022-02707-6.
[15] M.F. Kragh, H. Karstoft, Embryo selection with artificial intelligence: how to evaluate and compare methods? J. Assist. Reprod. Genet. 38 (7) (2021) 1675–1689, https://doi.org/10.1007/s10815-021-02254-6.
[16] Y. Zhang, L. Shen, X. Yin, W. Chen, Live-birth prediction of natural-cycle In vitro fertilization using 57,558 linked cycle records: a machine learning perspective, Front. Endocrinol. (Lausanne) 13 (April) (2022) 1–12, https://doi.org/10.3389/fendo.2022.838087.
[17] D.J. McLernon, E.W. Steyerberg, E.R. Te Velde, A.J. Lee, S. Bhattacharya, Predicting the chances of a live birth after one or more complete cycles of in vitro fertilisation: population based study of linked cycle data from 113 873 women, BMJ (online) (2016) 355, https://doi.org/10.1136/bmj.i5735.
[18] C.A. Jones, A.L. Christensen, H. Salihu, W. Carpenter, J. Petrozzino, E., …. Abrams, L.G. Keith, Prediction of individual probabilities of livebirth and multiple birth events following in vitro fertilization (IVF): a new outcomes counselling tool for IVF providers and patients using HFEA metrics, Journal of experimental and clinical assisted reproduction 8 (2011) 1–10.
[19] K.D. Sanders, G. Silvestri, T. Gordon, D.K. Griffin, Analysis of IVF live birth outcomes with and without preimplantation genetic testing for aneuploidy (PGT-A): UK Human Fertilisation and Embryology Authority data collection 2016–2018, J. Assist. Reprod. Genet. 38 (12) (2021) 3277–3285, https://doi.org/10.1007/s10815-021-02349-0.

[20] M.R. Hassan, S. Al-Insaif, M.I. Hossain, J. Kamruzzaman, A machine learning approach for prediction of pregnancy outcome following IVF treatment, Neural computing and applications 32 (7) (2020) 2283–2297, https://doi.org/10.1007/s00521-018-3693-9.

[21] R. Milewski, the usage of margin-based feature selection algorithm, Ivf Icsi /Et 21 (34) (2010) 35–46.

[22] D. Theng, K.K. Bhoyar, Feature selection techniques for machine learning: a survey of more than two decades of research, 66, in: Knowledge and Information Systems, 66, Springer London, 2024.

[23] B. Sowan, M. Eshtay, K. Dahal, H. Qattous, L. Zhang, Hybrid PSO feature selection-based association classification approach for breast cancer detection, Neural computing and applications 35 (7) (2023) 5291–5317, https://doi.org/10.1007/s00521-022-07950-7.

[24] O.S. Qasim, Z.Y. Algamal, Feature selection using particle swarm optimization-based logistic regression model, Chemometrics and intelligent laboratory systems 182 (2018) 41–46, https://doi.org/10.1016/j.chemolab.2018.08.016.

[25] L. Zhou, D. Xu, Y. Yuan, L. Wang, Research on transformer fault intelligent diagnosis technology based on improved random forest algorithm, Journal of physics: conference series 2728 (1) (2024), https://doi.org/10.1088/1742-6596/2728/1/012056.

[26] A. Borji, A. Seifi, T.H. Hejazi, An efficient method for detection of Alzheimer's disease using high-dimensional PET scan images, Intelligent decision technologies 17 (3) (2023), https://doi.org/10.3233/IDT-220315.

[27] B. Charbuty, A. Abdulazeez, Classification based on decision tree algorithm for machine learning, Journal of applied science and technology trends 2 (01) (2021) 20–28, https://doi.org/10.38094/jastt20165.

[28] S. Tabinda Kokab, S. Asghar, S. Naz, Transformer-based deep learning models for the sentiment analysis of social media data, Array 14 (October 2021) (2022) 100157, https://doi.org/10.1016/j.array.2022.100157.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N., … Gomez, I. Polosukhin, Attention is all you need, *Adv Neural Inf Process Syst, 2017-Decem* (Nips) (2017) 5999–6009.

[30] E. Albini, J. Long, D. Dervovic, D. Magazzeni, Counterfactual Shapley additive explanations, in: ACM international conference proceeding series, 2022, pp. 1054–1070, https://doi.org/10.1145/3531146.3533168.

[31] J. Allgaier, L. Mulansky, R.L. Draelos, R. Pryss, How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare, Artif. Intell. Med. 143 (February) (2023) 102616, https://doi.org/10.1016/j.artmed.2023.102616.

[32] Borji, A., Hejazi, T.-H., & Seifi, A. (2024). Introducing an ensemble method for the early detection of Alzheimer's disease through the analysis of PET scan images, 1–22.

[33] G. Coticchio, B. Behr, A. Campbell, M. Meseguer, D.E. Morbeck, V. Pisaturo, K. Lundin, Fertility technologies and how to optimize laboratory performance to support the shortening of time to birth of a healthy singleton: a Delphi consensus, J. Assist. Reprod. Genet. 38 (5) (2021) 1021–1043, https://doi.org/10.1007/s10815-021-02077-5.

[34] S.A. Sadegh-Zadeh, S. Khanjani, S. Javanmardi, B. Bayat, Z. Naderi, A. M. Hajiyavand, Catalyzing IVF outcome prediction: exploring advanced machine learning paradigms for enhanced success rate prognostication, Front. Artif. Intell. 7 (November) (2024) 1–18, https://doi.org/10.3389/frai.2024.1392611.

[35] A. Borji, G. Kronreif, B. Angermayr, S. Hatamikia, Advanced hybrid deep learning model for enhanced evaluation of osteosarcoma histopathology images, Front Med (Lausanne) 12 (April) (2025) 1–19, https://doi.org/10.3389/fmed.2025.1555907.