MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages 4767-4772

# PEARSON CORRELATION COEFFICIENT AND XGBOOST MODEL FOR LIVE BIRTH PREDICTION

**Mrs. V.Kalaiselvi**
**Research Scholar, Department of Computer Science,**
**PSGR Krishnammal college for women,**
kalaiselvivelliangiri@gmail.com
**Dr. S.Poongodi**
**Associate Professor, Department of Computer Science,**
**PSGR Krishnammal college for women,**
poongodis@gmail.com

**ABSTRACT:** Male variables and the metabolic health of both couples are frequently under-represented in clinical prediction models for assisted reproductive technologies, which mostly concentrate on female ovarian reserve markers. Furthermore, nonlinear patterns in reproductive data may be difficult for conventional parametric models. In this paper, The Extreme Gradient Boosting (XGboost) model was developed to classify medical pregnancy result in couples after intracytoplasmic sperm injection (ICSI) or in vitro fertilization (IVF). Pearson Correlation Coefficient (PCC) is used for feature selection. Dataset is collected from https://www.kaggle.com/datasets/deepakloganathan/live-birth-dataset PCC calculates the two variables linear connection. It helps identify which clinical features are strongly associated with the target outcome (live birth).Gradient boosting decision trees are the basis of the XG Boost as ensemble learning method. It is widely used in healthcare prediction tasks due to its high accuracy and robustness. It enhances gradient boosting algorithms by combining weak learners (trees) sequentially to minimize errors, using advanced regularization and parallel processing to live birth prediction.. XG Boost is scalable method which enhances the prediction performance and speed of Gradient Boosting Machines (GBM). It accomplishes this by employing a novel tree learning technique using distributed and parallel computing to speed up model discovery. The XG Boost-based prediction model performed exceptionally for IVF/ICSI outcomes in male factor infertile couples. Metrics including precision, recall, f-measure, and accuracy are used to assess machine learning algorithms such as Logistic Regression (LR), Random Forest (RF), Light Gradient Boosting Machine (Light GBM), and extreme Gradient Boosting (XG Boost) based on the predictive elements.

**KEYWORDS:** Male factor infertility, In Vitro Fertilisation (IVF), Intracytoplasmic Sperm Injection (ICSI), Live Birth Prediction, Pearson Correlation Coefficient (PCC), Feature Selection, Extreme Gradient Boosting (XG Boost), Machine Learning, Reproductive Health Analytics, and Healthcare Data Mining.

## 1. INTRODUCTION

Infertility is a significant global health issue [1,2], impacting about 15% of couples in reproductive age, with male factors responsible for 40–50% of the cases [3]. IVF and ICSI are the most successful treatment options [4] due to male factors such as oligozoospermia, asthenozoospermia, teratozoospermia. Nevertheless, the scientific pregnancy rate per IVF/ICSI cycle remains basically 40%–60% [5]. Patients experiencing unsuccessful cycles often face considerable psychological distress and financial trouble, making precise pre-treatment predictions of pregnancy outcomes crucial for tailoring treatments and managing expectations [6].

A number of medical professionals have been using their knowledge to make trial-and-error predictions about the likelihood of pregnancy. Consequently, traditional prediction methods lack a systematic statistical methodology and depend on the expertise of a single medical professional. This makes them more subjective. In order to make decision about IVF treatment, patients and medical experts are anxiously awaiting a measurement. Recent technology advancements like artificial intelligence (AI), machine learning (ML), and deep learning (DL) have the possible to address some of the persistent issues with statistical data-driven methods.The study of machine learning enables computers and other systems to think similarly and produce predictions by learning from and training on previous experiences [7]. It uses meaningful, pattern-oriented data exploration to provide the systems the adaptability to replicate human decision-making. Deep learning is an area of machine learning based on the ideas of human brain networks [8]. Some important patterns in the data may be noticed by humans when evaluating large, complex data samples [9]. Feature selection is the process of identifying and selecting the optimal features (variables) from a dataset that significantly enhance a machine learning expected performance. It enhances interpretability, lowers computational cost, and boosts model accuracy by removing unnecessary, redundant, or noisy features. Prediction models based on machine learning are used in IVF/ICSI procedures to predict live births, feature selection is essential for identifying the most important clinical parameters linked to the intended outcome. By selecting only, the most important variables, the model becomes more efficient and avoids overfitting.

Generally speaking, feature selection techniques fall into three categories: 1. Filter Methods: These techniques assess features using statistical metrics unrelated to the learning algorithm. The Pearson Correlation Coefficient (PCC), Mutual Information, and the Chi-square test are a few examples. PCC helps to preserve highly correlated predictors by measuring the linear connection between characteristics and the target variable. 2. Wrapper Methods: These techniques use a prediction model to assess feature subsets and choose the best-performing subset. This includes methods like recursive feature elimination (RFE), backward elimination, and forward selection. 3. Embedded Techniques: During dataset training, these techniques choose features. The model was built using algorithms including Random Forest (RF), Least Absolute Shrinkage and Selection Operator (LASSO), and XGBoost estimation.In general, feature selection improves prediction accuracy, model stability, and processing efficiency—especially when dealing with high-dimensional medical datasets. Machine learning algorithms, such random forest and gradient boosting trees, provide advantages over traditional statistical methods when modelling complicated, nonlinear, and high-dimensional data [10]. In the male factor infertility population, there are still a few high-accuracy machine learning models that take couple-level factors into account and are based on large samples. Extreme Gradient Boosting (XGBoost) model was developed to forecast clinical pregnancy outcomes in male factor infertile couples after ICSI/ IVF. Features are selected using Pearson Correlation Coefficient (PCC). Gradient boosting decision trees are the basis of the XGBoost ensemble learning method. Metrics including accuracy, f-measure, recall, and precision are used to assess the outcome.

## 2. LITERATURE REVIEW

Li et al., [11] validated a machine learning model and Shapley Additive explanations (SHAP) to forecast medical pregnancy result in couples' infertility going IVF/ICSI. This retrospective review analyzed 2,565 couples at Shanghai First Maternity and Infant Hospital from 2019 to 2025, using a training set of 70% of cases and a validation set of 30%. LASSO regression selected features, while five-fold cross-validation was employed to enhance LR, and Light GBM. SHAP visually clarified the best model, but potential studies are required to measure the impact of targeted therapies on clinical outcomes.

Liu et al., [12] developed prediction models that can forecast live births in women undergoing their first new or stationary IVF/ICSI cycles. The study analyzed a retrospective cohort of 1,857 women treated at Huizhou Municipal Central Hospital from 2019 to 2021. Variables were categorized based on data collected before and after the first cycle, followed by predictive tests. Four supervised machine learning algorithms—

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages 4767-4772

Logistic Regression (LR), Random Forest (RF), XGBoost, and LightGBM were utilized to create the models. Performance was assessed using metrics like AUC, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy.

Liu et al., [13] suggested classification methods for predicting live birth outcomes (LBO) based on historical data from 1405 IVF patients. Through univariate and multivariate analysis, significant factors were identified to build support vector machine (SVM) and artificial neural network (ANN) models. The model demonstrating the best performance was selected for real-world clinical application. This approach efficiently predicts LBO by considering important factors in IVF therapy, offering clinicians objective support for customizing treatment and embryo transfer plans.

Wu et al., [14] developed a prediction model for results following the transfer of fresh embryos. Between 2016 and 2023, the Shanghai First Maternity and Infant Hospital collected 51,047 papers related to assisted reproductive technologies (ARTs). In order to evaluate 11,728 samples and 55 pre-pregnancy features, the prediction model was constructed using RF, XGBoost, GBM, AdaBoost, LightGBM, and ANN after data collection. This study provides a significant increase in the live birth prediction by traditional evaluations, and patient counselling in ARTs can benefit from machine learning.

Dehghan et al., [15] create a contrasting ML approaches for predicting IVF outcome. RF, ANN, SVM, Recursive Partitioning and Regression Trees (RPART), and AdaBoost were developed to predict IVF success. The durability of the techniques is increased with the adoption of the Genetic Algorithm (GA).All classifiers' performance was much enhanced by GA, highlighting the significance of feature selection. These results demonstrate how ML and GA can help IVF clinicians make more accurate predictions, allowing for individualized treatment approaches for every patient. The usefulness and dependability of these predictive models in clinical IVF therapy can be further improved by more study and validation. AdaBoost attained the highest accuracy rate of 89.80%, especially when paired with GA feature selection. RF also performed well while using GA, attaining an accuracy rate of 87.40%.

Zhu et al., [16] conducted an internal validation study involving 1836 endometriosis patients who underwent IVF/ICSI fresh embryo transfers at Fujian Provincial Maternity and Children Hospital from 2018 to 2023. Participants were allocated in a 70:30 ratio to training (1285) and validation (551) sets. Independent variables were selected using LASSO and Recursive Feature Elimination (RFE). The XGBoost model was chosen for its superior prediction performance, with hyperparameters optimized through grid search. Additionally, feature importance and SHAP value plots were used to analyze the contributions and mechanisms of significant features in model predictions.

Wan et al., [17] developed a machine learning (ML) predictive model for clinical pregnancy outcomes prediction in endometriosis (EM) patients going fresh ET in a study spanning 2014 to 2024 with 1,752 participants. The model was based on 24 clinical and embryonic characteristics, utilizing algorithms such as Naïve Bayes, LR, RF, k-Nearest Neighbours (KNN), Neural Networks, and XG Boost. Feature selection employed LR and RFE with tenfold cross-validation, resulting in an XG Boost algorithm that effectively predicts clinical pregnancy in EM patients, demonstrating strong performance and interpretability.

## 3. PROPOSED METHODOLOGY

In this paper, the dataset is https://www.kaggle.com/datasets/deepakloganathan/live-birth-dataset. Pearson Correlation Coefficient (PCC) is measured between the linear relationship of two variables for optimal feature selection. Extreme Gradient Boosting (XG Boost) model was developed to forecast medical pregnancy outcomes in IVF/ICSI couples. The XG Boost is based on gradient boosting decision trees. By successively merging weak learners (trees) to reduce errors and employing sophisticated regularization and parallel processing to predict live births, it improves gradient boosting approaches. Precision, recall, f-measure, and accuracy are used to assess the results.

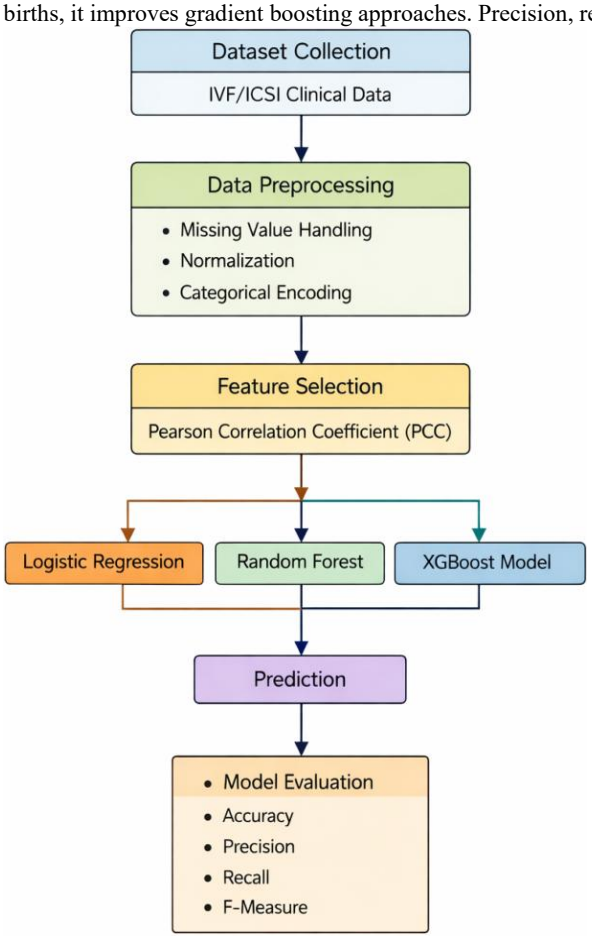



**FIGURE 1. PCC-XG Boost framework for IVF/ICSI prediction**

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages 4767-4772

### 3.1. DATASET COLLECTION

Kaggle Live Birth Dataset is collected from https://www.kaggle.com/datasets/deepakloganathan/live-birth-dataset. It contains clinical and demographic variables such as Male infertility indicators, Hormonal profiles, Embryo parameters, Treatment characteristics, and Live birth outcome (target variable).

### 3.2. DATA-PROCESSING

The dataset is preprocessed to improve its quality. Among the procedures are categorical variable encoding, data normalization, and missing value management. This stage verifies that the dataset is appropriate for machine learning models. Initially, the mean imputation approach for numerical data and mode imputation for categorical variables are used to handle missing values. This method substitutes the most common category for missing categorical data and the mean value of the associated a feature for missing numerical values. This technique helps maintain the dataset with statistical features and avoids data loss that might happen if the dataset has incomplete records were removed in this step.

Data normalization is then applied using the Min–Max normalization technique, which modifies numerical features into a predetermined range between 0 and 1. By preventing variables with larger numerical ranges from controlling those with smaller ranges, this technique enhances machine learning model performance and convergence. The equation (1) for rescaling is described as follows [18],

$$v' = \frac{v - min_A}{max_A - min_A} \tag{1}$$

where $v$ is the original value, $min_A$ and $max_A$ is denoted as lowest and highest range of the feature , and $v'$ is the normalized value.

Encoding categorical variables using the One-Hot Encoding technique is another crucial step. Each categorical attribute in this method is converted into several binary variables that indicate whether a category is present or not.

### 3.3. FEATURE SELECTION USING PCC

The Pearson Correlation Coefficient (PCC) is used to assess how each attribute relates to the result of a live birth. PCC is computed using the covariance between each feature and the target variable, as determined by their respective standard deviations. By calculating correlation values between variables and the goal, strongly linked predictors can be identified in the PCC. It also removes redundant or irrelevant variables. For the classifier, this dimensionality is reduced and model efficiency is raised. It is estimated by dividing the product of two variables' standard deviations by their covariance. Using this method, each categorical attribute is transformed into multiple binary variables that represent the presence or absence of a category. Features with higher absolute correlation values are selected as significant predictors for training model. This allows ML algorithms to effectively process categorical information without introducing ordinal bias. It is described by equation (2) [19],

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2}$$

where $x_i$ is the feature input value, $y_i$ is defined as target variable (live birth outcome), $\bar{x}, \bar{y}$ is the mean of feature and target values, and n is defined as number of observations.

### 1.1. LIVE BIRTH PREDICTION USING EXTREME GRADIENT BOSSTING (XG BOOST)

The XGBoost model predicts the likelihood that IVF/ICSI procedures will result in a live birth or clinical pregnancy. A scalable technique called XG Boost enhances the prediction performance and speed of Gradient Boosting Machines (GBM). It manages by using distributed and parallel computing to speed up model discovery and a novel tree learning algorithm. XG Boost is well-known for its outstanding prediction performance and has been applied in many different fields. XG Boost is a boosting application that enhances prediction accuracy by combining multiple learning applications. It utilizes a decision tree-based ensemble machine learning method, commonly used in data science, integrating outcomes from various distinct trees. By employing a gradient descent optimization method, the XG Boost model seeks to minimize the loss function [20]. Boosting is an ensemble technique that merges numerous lower-performing forecasting models into high-performance model through continuous integration in allowed parameters. Let us consider that the live birth prediction dataset is denoted as $DS$ with m features and an n number of samples $DS = \{(x_i, y_i), i = 1,2,3 \ldots, n, x_i \in \mathbb{R}^m, y_i \in \mathbb{R}\}$. Let $\hat{y}_i$ be the predicted results of an XG Boost model created using the equations (3-4),

$$A_i = \sum_{k=1}^{K} f_k(X_i), f_k \in \mathcal{F} \tag{3}$$

Number of trees in the XG Boost model is denoted as K, $f_k$ is represented as the k[th] tree. It is used to find the greatest set of functions by reducing the loss and regularization objective by equation (4),

$$\mathcal{L}(\phi) = \sum_i l(y_i, A_i) + \sum_k \Omega(f_k) \tag{4}$$

where $l$ is represented as the loss function. It is computed based on the predicted output $\hat{y}_i$ and the actual output $y_i$, $\Omega$ is a measure the complexity of model, this helps in solving over-fitting of the model and it is calculated using equation (5),

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\|w\|^2 \tag{5}$$

where, T is denoted as the total number of leaves in a tree, while w represents the weight of those leaves. To minimize the objective function in decision trees, the model utilizes function boosting, introducing a new function f as training process. Consequently, in the t[th] iteration a new function f is computed using equations (6-9),

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, A_i^{(t-1)} + f_t(X_i) + \Omega(f_t)\right) \tag{6}$$

$$\mathcal{L}_{\text{split}} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in i} h_i + \lambda}\right] - \gamma \tag{7}$$

$$g_i = \partial_{A^{(t-1)}} l\left(y_i, A_i^{(t-1)}\right) \tag{8}$$

$$g_i = \partial^2_{A^{(t-1)}} l\left(y_i, A_i^{(t-1)}\right) \tag{9}$$

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages 4767-4772

### 3.4. PERFORMANCE EVALUATION

Precision, recall, f-measure, and accuracy have been used to assess the efficiency of prediction methods.

## 4. RESULTS AND DISCUSSION

The purpose of the simulation study was to evaluate how well classification techniques predicted clinical pregnancy outcomes (live birth) comparing couples by male factor infertility utilizing IVF/ICSI treatment. The information, which included pertinent clinical, demographic, and metabolic characteristics of both spouses, was taken from the publicly accessible Kaggle repository (Live Birth information). Preprocessing steps including missing value imputation, data normalization, and categorical feature encoding are performed to ensure data dependability and model compatibility. Prediction outcomes are assessed using metrics such as precision, recall, f-measure, and accuracy, derived from the Confusion Matrix which includes True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The dataset is divided into 80:20 training and testing groups for applying prediction techniques. For prediction, LR, RF, Light GBM, and XG Boost were used. Grid search is also used for hyperparameter tuning in order to optimize the model. Python was used for all simulations, along with tools like XG Boost and Scikit-learn

The accuracy of positive predicts is measured by precision. It is computed using equation (10),

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

Recall measures the ability of a model to detect every significant instance in a dataset using equation (11),

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

The harmonic mean of precision and recall is known as the F-measure. It provides a single score that achieves a balance between the two metrics. It is calculated using equation (12),

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{12}$$

The ratio of exactly predicted observations to total samples is known as accuracy by equation (13),

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

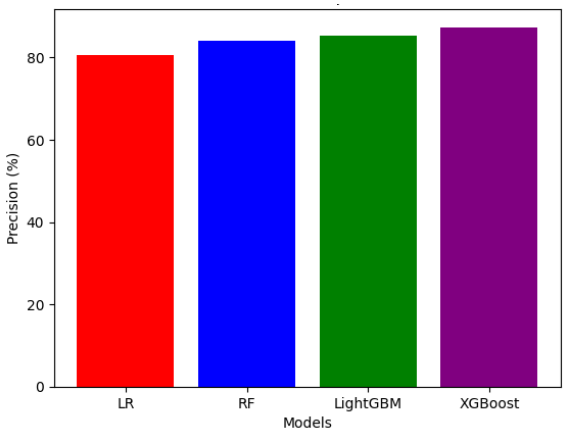
Table 1 demonstrate that XG Boost performs best across all evaluation metrics, followed by Light GBM, RF, and LR. The results validate the effectiveness of XG Boost in improving prediction accuracy and reliability for live birth outcomes in assisted reproductive technology as 87.36%, 85.93%, 86.63%, and 88.61% for precision, recall, f-measure, and accuracy.

**TABLE 1. RESULTS COMPARISON OF PREDICTION METHODS**

| Methods/ metrics | LR | RF | LightGBM | XGBoost |
|---|---|---|---|---|
| **Precision(%)** | 80.65 | 83.96 | 85.25 | 87.36 |
| **Recall(%)** | 79.47 | 81.47 | 83.68 | 85.93 |
| **F-measure(%)** | 80.06 | 82.69 | 84.45 | 86.63 |
| **Accuracy(%)** | 81.68 | 83.55 | 86.02 | 88.61 |

The precision comparison with respect to prediction techniques such as LR, RF, Light GBM, and XG Boost is displayed in Figure 2. The number of correctly identified positive cases (live births) with all anticipated positive outcomes is used to calculate it.LR has lowest precision results of 80.65%, RF improves precision to 83.96%, LightGBM further improves performance to 85.25%, and XG Boost achieves the highest precision of 87.36%, indicating its superior ability to precisely classify true positive cases through minimal false results. This suggests that models based on boosting are more accurate in forecasting successful clinical pregnancy outcomes. The recall comparison with respect to prediction techniques such as LR, RF, Light GBM, and XG Boost is displayed in Figure 3. The capability of models to precisely recognize real positive cases (live births) is well-known as recall. LR achieves a recall of 79.47%, indicating some missed positive cases. While Light GBM achieves 83.68%, demonstrating superior sensitivity, RF boosts recall to 81.47%. With the highest recall of 85.93%, XG Boost demonstrates a strong ability to capture most real live birth cases. This demonstrates the sophistication ensemble techniques can effectively lower false negatives in clinical prediction tasks.

**Figure 2. Precision comparison of prediction methods**



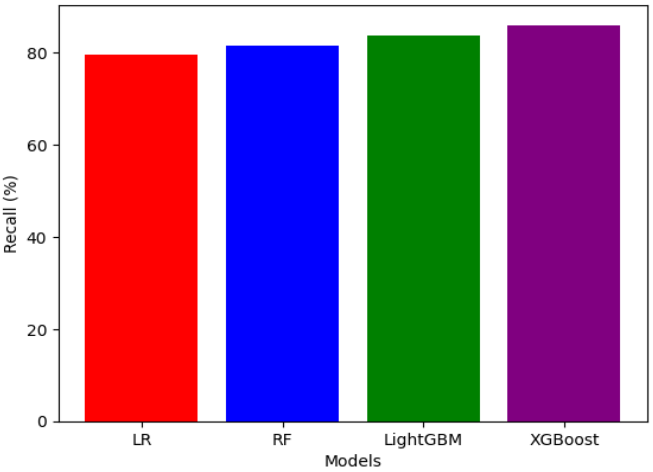




**Figure 3. Recall comparison of prediction methods**

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages 4767-4772

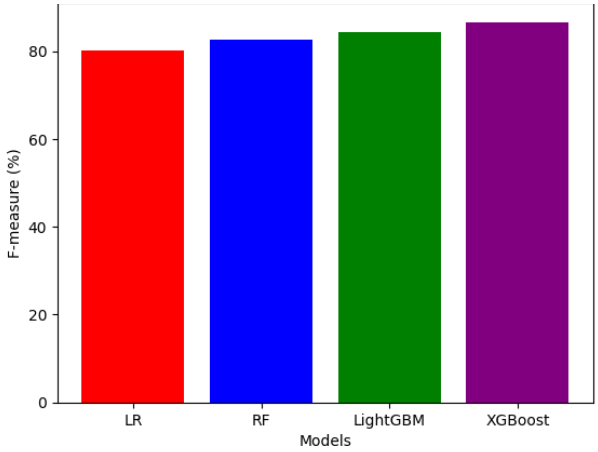



**Figure 4. F-measure comparison of prediction methods**

F-measure comparison of all prediction methods like LR, RF, Light GBM, and XG Boost are illustrated in figure 4. LR gives the minimum results (80.06%), RF has slightly increased than the LR (82.69%). Light GBM achieves increased results of 84.45%, indicating an improved trade-off among precision and recall. XG Boost has maximum F-measure of 86.63% when compared to all models demonstrating its robustness and balanced results. XG Boost is extremely appropriate for managing imbalanced and complex datasets.

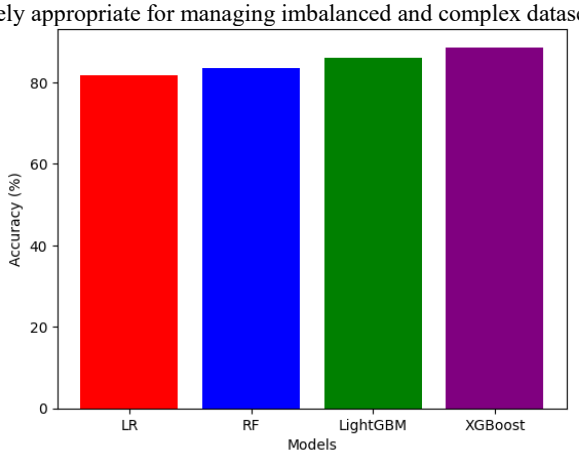



**Figure 5. Accuracy comparison of prediction methods**

Figure 5, accuracy shows the overall comparison of all prediction methods like LR, RF, Light GBM, and XG Boost in predicting together positive and negative cases. RF boosts the accuracy to 83.55%, while LR yields the lowest accuracy of 81.68%. XG Boost has the highest accuracy of 88.61%, indicating its superior analytical ability, while Light GBM has the highest accuracy of 86.02%. It comes to the conclusion that, in terms of capturing nonlinear relationships identified in IVF/ICSI clinical data, the proposed approach performs better than other methods.

## 5. CONCLUSION AND FUTURE WORK

In this work, Pearson Correlation Coefficient (PCC) is introduced for feature selection that which evaluates the linear relationship among two features in the dataset. PCC is calculated by evaluating the covariance among every feature and the target class, normalized depending on their standard deviations which discover the robustly correlated predictors. These values are ranged among −1 and +1, where maximum absolute values specify stronger linear relationships. Based on these values, features which show strong correlation by the target class are chosen, at the same time as weakly correlated or redundant features are separated, thus decreasing dimensionality and increasing accuracy of model. Extreme Gradient Boosting (XG boost) model was created to predict medical pregnancy results in IVF/ICSI couples by male factor infertility. Gradient boosting decision trees are a basis of the XG Boost method. It enhances gradient boosting algorithms by combining weak learners (trees) sequentially to minimize errors, using advanced regularization and parallel processing to live birth prediction. Due to its capability to confine complex nonlinear associations and interactions among clinical variables, XG Boost is best for accurate live birth prediction in assisted reproductive technology datasets. LR, RF, Light GBM, and XG Boost models were assessed using Precision, Recall, F-measure, and Accuracy metrics. XG Boost achieved scores of 87.36% for precision, 85.93% for recall, 86.63% for f-measure, and 88.61% for accuracy. The model is also extended to include deep learning and ensemble methods for classification. Hybrid ensemble strategies, including stacking and voting mechanisms, are also explored to leverage the complementary strengths of individual classifiers.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages 4767-4772

## REFERENCES

1.Serafini, S. and O'Flaherty, C., 2025. Dysregulation of sphingolipid and cholesterol homeostasis imposes oxidative stress in human spermatozoa. Redox Biology, 84, pp.1-16.
2.Jin, Z.R., Fang, D., Liu, B.H., Cai, J., Tang, W.H., Jiang, H. and Xing, G.G., 2021. Roles of CatSper channels in the pathogenesis of asthenozoospermia and the therapeutic effects of acupuncture-like treatment on asthenozoospermia. Theranostics, 11(6), pp.2822-2844.
3.Salonia, A., Capogrosso, P., Boeri, L., Cocci, A., Corona, G., Dinkelman-Smit, M., Falcone, M., Jensen, C.F., Gül, M., Kalkanli, A. and Kadioğlu, A., 2025. European Association of Urology guidelines on male sexual and reproductive health: 2025 update on male hypogonadism, erectile dysfunction, premature ejaculation, and Peyronie's disease. European Urology, 88(1), pp.76-102.
4.Goyal, A., Kuchana, M. and Ayyagari, K.P.R., 2020. Machine learning predicts live-birth occurrence before in-vitro fertilization treatment. Scientific reports, 10(1), pp.1-12.
5.Adeniyi, T., Horne, G., Ruane, P.T., Brison, D.R. and Roberts, S.A., 2021. Clinical efficacy of hyaluronate-containing embryo transfer medium in IVF/ICSI treatment cycles: a cohort study. Human reproduction open, 2021(1), pp.1-10.
6.Zou, K., Wang, J., Bi, H., Zhang, Y., Tian, X., Tian, N., Ma, W. and Wu, J., 2019. Comparison of different in vitro differentiation conditions for murine female germline stem cells. *Cell Proliferation*, *52*(1), pp.1-12.
7.Rowe, M., 2019. An introduction to machine learning for clinicians. *Academic Medicine*, *94*(10), pp.1433-1436.
8.El Bouchefry, K. and de Souza, R.S., 2020. Learning in big data: Introduction to machine learning. In *Knowledge discovery in big data from astronomy and earth observation* (pp. 225-249). Elsevier.
9.Cui, S., Tseng, H.H., Pakela, J., Ten Haken, R.K. and El Naqa, I., 2020. Introduction to machine and deep learning for medical physicists. Medical physics, 47(5), pp.e127-e147.
10. Huang, S., Tuerganbayi, K., Wang, J., Saad, S.H., Zhang, J., Zou, J., Yan, X. and Huang, K., 2025. Machine learning-based preliminary screening tool for clinical pregnancy prediction: towards management of IVF/ICSI stages. *Annals of Medicine*, *57*(1), pp.1-17.
11. Li, H., Gao, J. and Li, Y., 2026. Machine learning–based prediction of IVF/ICSI outcomes in male factor infertility highlighting couple-level BMI. Frontiers in Endocrinology, 17, pp.1-11.
12. Liu, X., Chen, Z. and Ji, Y., 2023. Construction of the machine learning-based live birth prediction models for the first in vitro fertilization pregnant women. *BMC Pregnancy and Childbirth*, *23*(1), pp.1-12.
13. Liu, L., Liang, H., Yang, J., Shen, F., Chen, J. and Ao, L., 2024. Clinical data-based modeling of IVF live birth outcome and its application. *Reproductive biology and endocrinology*, *22*(1), pp.1-12.
14. Wu, S., Wang, X., Liu, Y., Ren, Y., Zhao, M., Song, H., Shen, H., Wu, Y., Wei, Z., Lu, H. and Li, K., 2025. Predictive models for live birth outcomes following fresh embryo transfer in assisted reproductive technologies using machine learning. Journal of Translational Medicine, 23(1), pp.1-13.
15. Dehghan, S., Rabiei, R., Choobineh, H., Maghooli, K., Nazari, M. and Vahidi-Asl, M., 2024. Comparative study of machine learning approaches integrated with genetic algorithm for IVF success prediction. *Plos one*, *19*(10), pp.1-19.
16. Zhu, S., Xu, H., Li, R., Chen, X., Jiang, W., Zheng, B. and Sun, Y., 2025. Development and validation of a machine learning-based predictive model for live birth outcomes following fresh embryo transfer in patients with endometriosis. *Journal of Assisted Reproduction and Genetics*, *42*(11), pp.3853-3867.
17. Wan, X., Yu, M., Wu, X., Huang, Z. and Tan, J., 2025. Machine learning prediction of clinical pregnancy in endometriosis patients following fresh IVF/ICSI-ET. *European Journal of Medical Research*, *30*(1), pp.1-10.
18. Shantal, M., Othman, Z. and Abu Bakar, A., 2025. Missing data imputation using correlation coefficient and min-max normalization weighting. *Intelligent Data Analysis*, *29*(2), pp.372-384.
19. Dufera, A.G., Liu, T. and Xu, J., 2023. Regression models of Pearson correlation coefficient. *Statistical Theory and Related Fields*, *7*(2), pp.97-106.
20. Noorunnahar, M., Chowdhury, A.H. and Mila, F.A., 2023. A tree based eXtreme Gradient Boosting (XGBoost) machine learning model to forecast the annual rice production in Bangladesh. *PloS one*, *18*(3), pp.1-15.