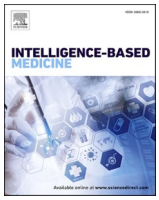
Intelligence-Based Medicine 11 (2025) 100204

Contents lists available at ScienceDirect

Intelligence-Based Medicine

journal homepage: www.sciencedirect.com/journal/intelligence-based-medicine



Comparing machine learning approaches for predicting the success of ICSI treatment: A study on clinical applications

Abrar Mohammad [a], Haneen Awad [a], Huthaifa I. Ashqar [a,b,*]

[a] Arab American University, Jenin, Palestine
[b] Columbia University, NY, USA

ARTICLE INFO

Keywords:
Intracytoplasmic sperm injection
Fertilization
Infertility treatment
Success rate
Classification

ABSTRACT

Intracytoplasmic Sperm Injection (ICSI) is widely used to treat almost all forms of male infertility and to overcome fertilization failure. While ICSI is a powerful procedure, it's also considered quite expensive, which means couples and clinicians have to make informed decisions about whether or not to proceed with this treatment. About 10,036 patient records, 46 attribute sets, and one label column that indicates the success or failure of pregnancy after the ICSI treatment were used to conduct this research. The data were gathered from Razan infertility center in Palestine. The ICSI dataset contains only clinical features that are known prior to deciding on ICSI treatment. The dataset contains 46 features, 5 of the independent features have categorical values, 12 are numerical, 3 are string, and 26 are binary. Based on the results, RF algorithm achieved the highest AUC score of 0.97, followed by the NN with a score of 0.95, and the RIMARC algorithm with a score of 0.92. AUC is a widely used metric for evaluating the performance of binary classification models. Therefore, judging by the AUC scores, it appears that RF algorithm outperformed the other two algorithms in terms of the evaluated metric. The method employed in our analysis demonstrates considerable promise, practicality, and generalizability, driving advancements in fertility treatments and ultimately improving the chances of couples achieving their desired family goals.

1. Introduction

Infertility is one of the globally rising social problems. It affects people physically and psychologically. It is defined as the failure to achieve pregnancy after 12 months of unprotected intercourse[1]. The infertile couple starts treatment with multiple tests and analysis. This requires money, time and health for a treatment that they may have a low chance of getting [2]. An infertile couple has a chance to have a baby with assisted reproductive technologies (ART). The first baby born using in vitro fertilization (IVF) was in 1978 [3]. New treatment options have been used by clinicians since 1978, including intracytoplasmic sperm injection, preimplantation genetic diagnosis, gamete, and embryo cryopreservation [2]. In the field of reproductive medicine, intracytoplasmic sperm injection (ICSI) has been hailed as a breakthrough [3].

Machine-learning methods can aid with the medical decision-making processes at both the clinical and diagnostic levels [4,5]. Analyzing a clinical database will help inducing useful relationships between the data such as patient characteristics, ovarian hyperstimulation data, male-related factors, and pregnancy outcome. In vitro fertilization (IVF) is a difficult and emotionally stressful technique for infertile couples. Predicting clinical pregnancy outcomes with classification algorithms can significantly improve IVF treatment success rates. By identifying key clinical indicators associated with successful pregnancies, healthcare professionals can make more educated treatment protocol decisions, increasing the likelihood of positive results for patients.

One of the most critical medical responsibilities for both healthcare providers and patients is for the clinicians to decide whether a treatment has a decent chance of success or not In order to start that treatment, this paper focuses on the domain of Intracytoplasmic Sperm Injection (ICSI), where estimating the success rate is a crucial step to aid in that decision. An infertile couple planning on undergoing ICSI could use the machine-learning methods learned by this method to estimate a success rate [4]. Models generated by this machine-learning methods could be used to produce scores that help in estimating a success rate for an infertile couple who are planning to go through with the ICSI procedure[3].

This study undertakes a comparative analysis of various prediction models, evaluating their performance based on metrics such as

* Corresponding author. Arab American University, Jenin, Palestine.
E-mail address: huthaifa.ashqar@aaup.edu (H.I. Ashqar).

https://doi.org/10.1016/j.ibmed.2025.100204
Received 17 November 2024; Received in revised form 4 January 2025; Accepted 19 January 2025
Available online 20 January 2025

Accuracy, Precision, Recall, and the F1 score. These algorithms can identify patterns and correlations that might not be apparent to clinicians, leading to more tailored and effective estimation for the likelihood of successful pregnancy outcomes following ICSI treatment.

2. Literature review

Machine learning, a type of artificial intelligence, has gotten a lot of interest in a variety of fields. This study focuses on its usage in clinical decision-making, with a particular emphasis on infertility therapy and IVF. Large datasets can be analyzed using machine learning algorithms, and predictive models can be developed based on patterns and correlations between input and output variables. These models strive to increase the accuracy and efficiency of decision-making processes by using expertise and analyzing data. This section provides an overview of studies that have used machine learning approaches to predict fertility outcomes in the context of clinical decision-making.

A subset of artificial intelligence, machine learning, includes the use of algorithms to model the relationship between input and output data. Predictive models for data interpretation and understanding can be established by using experience. Machine learning methods include, among others, Random Forest, Support Vector Machines (SVM), Decision Tree, Artificial Neural Networks (ANN), and K-Nearest Neighbors (KNN). These approaches have showed promise in the solution of complicated problems [6,7].

Researchers have investigated the use of machine learning techniques in clinical decision-making. Several studies have been undertaken to use classification algorithms to predict outcomes in infertility therapies. For example, one study compared the clinical pregnancy rate (CPR) in infertility treatments using LR, RF, SVM, MLP, KNN, and GNB algorithms. The study included 1931 IVF or IUI patients. In terms of accuracy, sensitivity, F1 score, positive predictive value (PPV), and Matthews correlation coefficient (MCC), the Random Forest classifier surpassed other algorithms [8].

In another study, various classifiers such as Random Forest, recursive partitioning (RPART), adaptive boosting, SVM, and one-nearest neighbor were employed to predict IVF and ICSI outcomes. The results showed that Random Forest and RPART fared better than other strategies in terms of accuracy. Age, serum estradiol level, and the number of developing embryos were all important predictors [9].

Another study proposed a decision tree and SVM algorithm to predict ICSI results [10]. The model's input variables included information about the couple's medical history, hormone tests, and the reason for their infertility. The occurrence of a clinical pregnancy served as the output variable. The decision tree technique and SVMs had accuracy rates of 70.3 and 75.7 %, respectively. The SVM method outperformed the decision tree method in terms of performance.

Several studies argued that machine learning algorithm can predict conventional IVF success. One study used clinical information gathered from 58 randomly chosen couples during the IVF cycle to create a gradient-boosted machine classifier [4]. Additionally, sperm capacitation-related parameters like membrane potential, pHi, and hyperactivated motility were given to the classifier. In the majority of cases, the final algorithm accurately anticipated the outcome of fertilization. By enabling doctors to properly select between IVF and ICSI for each pair, a machine-learning method that incorporates clinical data and parameters linked to the capacitating status of sperm could be utilized to enhance and tailor IVF therapy for hundreds of thousands of couples each year.

Furthermore, machine learning models for predicting outcomes in male infertility therapies have been established. Using data from couples receiving testicular extracted sperm therapy, a logistic regression prediction model was developed. The model considered many parameters such as BMI, age, male follicle-stimulating hormone (FSH) levels, and uterine anomalies [11]. Machine learning algorithms have also been used to forecast embryo viability and clinical pregnancy outcomes. Models using computer vision image processing and deep learning approaches outperformed standard morphokinetic grading methods employed by embryologists in terms of prediction ability [12].

Furthermore, machine learning algorithms have shown promise in predicting results in infertility therapies [7,13]. These algorithms have been used to create predictive models based on a variety of characteristics including age, hormone levels, and treatment type [14,15]. Machine learning in the field of assisted reproductive technology has the potential to improve pregnancy success rates and optimize treatment decisions.

While the application of machine learning in predicting outcomes in infertility therapies is promising, there are gaps in the literature related to algorithm diversity, sample sizes, standardization, demographic considerations, clinical validation, model interpretability, and long-term predictions. Addressing these gaps could lead to more accurate, reliable, and widely applicable predictive models, ultimately improving pregnancy success rates and optimizing treatment decisions in assisted reproductive technology. On the other hand, there isn't enough research specifically looking at ICSI data as an infertility treatment, all previous studies include data on infertility treatment using IVF or IVF plus ICSI.

The paper contributes to the field of reproductive medicine by demonstrating the practical use of machine learning algorithms to predict the outcomes of fertility treatments. By showcasing the potential of predictive models to aid in clinical decision-making, the paper opens new avenues for research, collaboration, and the development of tools that can benefit both clinicians and patients in the realm of infertility treatment.

3. Dataset

Knowledge Discovery in Databases (KDD) requires well-defined and formal methods for managing data in order to be successful. In this study, the cross-industry standard process for data mining (CRISP-DM) was used. A six-phase methodology includes: 1) understanding the problem domain, 2) analyzing the data, 3) preparing the data, 4) modeling, 5) evaluating, and 6) deploying the solution (using the knowledge that has been discovered) [1]. For machine learning techniques to be successful, data preprocessing is a crucial step. As a final step leading to the modeling process, the final dataset is constructed from the original raw data set. We used a dataset that contains 10,036 patients' records, 46 attribute sets, and one label column that indicates the success or failure of pregnancy after the ICSI treatment. The data were gathered from Razan infertility center in Palestine.

To capture a wide range of clinical scenarios, we included all ICSI cycles performed at the center during the study period, regardless of the specific indications for the procedure. While this comprehensive approach provides a robust dataset for analysis, we acknowledge that it includes cases where ICSI was performed for reasons other than addressing infertility, such as to prevent DNA contamination during preimplantation genetic testing for aneuploidy (PGT-A). To maintain relevance, the success label in our dataset strictly refers to achieving clinical pregnancy, excluding intermediate outcomes like embryo quality or genetic testing results. The inclusion of a diverse range of cases enhances the generalizability of our findings but also highlights the need for caution when interpreting results in specific subpopulations. Future research may involve subgroup analyses to explore predictive factors and outcomes for distinct clinical indications, such as PGT-A cases, to provide more tailored insights.

The ICSI dataset contains only clinical features that are known prior to deciding on ICSI treatment. The dataset contains 46 features, 5 of the independent features have categorical values, 12 are numerical, 3 are string, and 26 are binary. The definition of success in IVF/ICSI treatment is multifaceted, with live birth rates widely regarded as the most comprehensive endpoint. However, in this study, we focus on clinical pregnancy as the primary outcome due to the nature of the dataset, which did not consistently include live birth data. While clinical

pregnancy may not capture the full picture, it often serves as the first critical decision point for families considering treatment, providing essential insights into the likelihood of initial success. Future research should aim to incorporate live birth rates and other factors, such as maternal health and pregnancy safety, to offer a more holistic evaluation of treatment outcomes. The ICSI dataset characteristics are summarized in Table 1.

The data cleaning process is an important step in classification modeling. Its function is to guarantee that the input data is in the correct format and free of errors, inconsistencies, and irrelevant information. Cleaning the data is critical since classification algorithms rely significantly on the accuracy and reliability of the input data to provide accurate and trustworthy predictions. The performed data cleaning process on the obtained dataset for this study contains five main processes to be performed on the various attributes we're handling. Using Jupiter Notebook, replace, transform, filter, categorize, and label encoding data processing techniques were performed. The data cleaning process prepares the data for further analysis and classification tasks by involving several steps to handle missing values, replace values, convert data types, and encode categorical variables.

The data cleaning process has started by filtering out all rows that had either null or missing values in the attributes that would affect the results of this study. After discussing the attributes and their importance, irreplaceability, and their effect on the OPU/ICSI procedure, we concluded that ICSI/OPU outcome, F Age, M Age, F Weight, F blood type, and M blood type are the attributes that are considered irreplaceable due to several relevant reasons. Text areas-attributes including F surgical History, M surgical History, and F current Medications were transformed into new categorical attributes according to analysis and directions from doctors.

The missing values of all other numerical features are replaced with the median after converting them to be set within the normal or abnormal range related to that attribute, for example AFC LT normal range lies between 6 and 12 so all records with AFC LT $>$ 6 AND $<$12 is then replaced with 0, records with AFC $<$6 OR $>$ 12 is replaced with 1, others are replaced with 0 due to the fact that doctors rarely forget to not report abnormal cases.

Another cleaning process the data went through was Label Encoder, Label Encoder is a prominent technique for converting categorical or textual data into numerical labels in machine learning and data preprocessing applications. It is especially beneficial when dealing with non-numerical features or variables that must be represented numerically in order for machine learning algorithms to process them properly. All attributes are then discretized into categorical values (i.e., 0 is normal, 1 is abnormal). After that, the Data then was ready for modeling.

The distribution of the outcome variable indicates an imbalanced dataset with 4,29 instances categorized as 0 and 305 instances as 1. This uneven distribution may affect the performance of predictive models. The dataset, after data cleaning and preprocessing, consists of 4524 instances and 44 Attributes. Which would affect the results of performance metrics and the model's ability to correctly classify the minority class. To solve such problems Oversampling was used.

Oversampling is a strategy used in unbalanced learning scenarios in which the class distribution in the dataset is substantially skewed, with one class significantly outnumbering the others. Oversampling's purpose is to enhance the representation of the minority class by generating synthetic samples, which helps to balance the class distribution and reduces the influence of class imbalance on the learning algorithm.

Oversampling is the process of creating new synthetic samples for the dataset's minority classes. Typically, this is accomplished by developing synthetic instances that are comparable to existing minority class samples. SMOTE (Synthetic Minority Oversampling Technique) is a common oversampling technique that is widely utilized in practice. After applying the process of oversampling The Outcome variable exhibits a balanced dataset with an equal count of 4219 instances for both categories 0 and 1. This balanced distribution is conducive to more accurate model training and evaluation. Following data cleaning and preprocessing, the dataset comprises 8438 instances and 44 attributes.

The dataset initially contained 10,136 records ranging from 0 to 10,135 and comprised 49 attributes. These attributes consisted of 14 with data type as numerical and 35 with data type category. However, after the data cleaning process, the dataset was refined, resulting in 8438 entries spanning from 0 to 8437 and a reduction in the number of attributes to 44. The data types were also modified, with 32 attributes being represented as integers, enhancing the dataset's overall quality and usability.

We acknowledge the importance of providing additional details regarding the patients' treatment timeline and any changes in treatment protocols. Unfortunately, due to the retrospective nature of the data collection, the dataset did not consistently include specific information on the exact timeline of treatments or any changes in treatment protocols over the study period. As a result, this data could not be incorporated into the analysis. However, we recognize that such factors may influence the outcomes, and we discuss this limitation in the manuscript. While this limitation exists, the study provides valuable insights into the predictive modeling of ICSI treatment success.

Table 1
Features in ICSI dataset.

ICSI Dataset (N: numerical, C: categorical, S: string, B: binary)		
Patient Key (N)	Male factor (B)	Hysterosalpingography Negative Left Tube (HSG N LT) (B)
Outcome (B)	Gender selection (B)	Hysterosalpingography Negative Left Tube (HSG N RT) (B)
Male Age (N)	Ovarian insufficiency (B)	HSG Tubes Hydrosalpinx LT (B)
Female Age (N)	Polycystic Ovary Syndrome (PCOS) (B)	HSG Tubes Hydrosalpinx RT (B)
Infertility Type (C)	Preimplantation Genetic Diagnosis - Polymerase Chain Reaction (PGD-PCR) (B)	Laparoscopy Tube LT status (B)
Female Blood Type (C)	Preimplantation Genetic Diagnosis - Fluorescence In Situ Hybridization) (PGD-FISH) (B)	Laparoscopy Tube RT status (B)
Male Blood Type (C)	Tubal (B)	Laparoscopy Ovary LT status (B)
Female weight (N)	Unexplained (B)	Laparoscopy Ovary RT status (B)
Kinship (C)	Female surgical History (S)	Semen Volume (N)
Female Height (N)	Female current medications (S)	Semen Count (N)
Female BMl (N)	Female surgical history (S)	Semen Motility (N)
Female cycle status (C)	Antral Follicle Count Left (AFC LT) (N)	Semen Morphology (N)
Anovoluation (B)	Ovary Antral Follicle Count Right Ovary (AFC RT) (N)	Testicular Sperm Extraction Positive (TESE + ve) (B)
Endomitriosis (B)	Uterus status (B)	Testicular Sperm Extraction Negative (TESE -ve) (B)
Hypo Female (B)	Hystroscopy Cavity status (B)	
Hypo Male (B)	HSG Cavity status (B)	

4. Methods

The paper addresses the challenge of deciding whether to proceed with ICSI treatment, which is an expensive procedure used to overcome various forms of couples' infertility. The paper emphasizes the importance of accurately predicting the success rate of ICSI treatments to guide clinical decision-making. The goal is to apply and compare classification algorithms that can predict the success or failure of the ICSI procedure based on patient data. This prediction will assist clinicians and patients in making informed decisions about whether to proceed

with the treatment. It should be noted that the approval of Institutional Review Board (IRB) or the Institutional Animal Care and Use Committee (IACUC) is not applicable in our case. Three classification algorithms were used for the prediction task:

Random Forest (RF): Random Forest is an ensemble learning algorithm that constructs multiple decision trees during training and combines their predictions to improve accuracy and reduce overfitting. It is known for its ability to handle high-dimensional data, capture complex relationships, and provide feature importance rankings.

Neural Networks (NN): Neural Networks are a type of machine learning model inspired by the human brain's structure. They consist of interconnected layers of artificial neurons that learn to recognize patterns and relationships in data through training. Neural Networks can capture both linear and nonlinear relationships, making them suitable for complex classification tasks.

RIMARC (Ranking by Imputation and Multiple Additive Regression Classification): RIMARC is a supervised ranking algorithm designed to maximize the Area Under the ROC Curve (AUC) by learning a scoring function for ranking instances. It constructs nonlinear score functions for each feature and combines them to achieve optimal ranking. RIMARC aims to provide accurate ranking of instances based on their likelihood of positive outcomes.

5. Analysis and results

We applied 3 classification models including Random forest, Neural Networks, and RIMARC. These prediction models are compared to each other in terms of performance indicators including area under the ROC curve, accuracy, precision, recall, and F1 score. These performance measures are extensively used in classification model evaluation and aid in assessing various elements of the model's performance [16]. The best appropriate metric is determined by the specific problem, dataset features, and classification task purpose [16].

The Area Under the Receiver Operating Characteristic Curve (AUC) is a performance metric often used to evaluate the performance of a binary classification model. It assesses the model's ability to discriminate between positive and negative classes. The AUC-ROC value ranges between 0 and 1, with a greater value indicating better classification performance. AUC-ROC of 0.5 indicates a random classifier, while a number closer to 1 indicates a better-performing model [17].

Accuracy is a regularly used evaluation metric that gauges a classification model's overall correctness. It computes the proportion of correct guesses to total forecasts made. Accuracy is calculated as number of correct guesses divided by the total number of forecasts. When the classes are balanced, which means that the quantity of samples in each class is about equal, accuracy is appropriate. When dealing with skewed datasets, however, it may not be an effective statistic [16].

Precision is a statistic that quantifies the proportion of accurately anticipated positive cases in comparison to all positive instances forecasted. To put it another way, it assesses the model's capacity to avoid false positives. Precision is calculated as the true positives divided by the true positives plus false positives. High accuracy means that the model is likely to be true when it predicts a positive class [8].

Recall is the fraction of accurately predicted positive cases in the dataset divided by the total number of positive instances. It evaluates the model's ability to correctly identify all positive samples. The formula for recall is true positives divided by the true positives plus the false negatives. A greater recall value means fewer false negatives [16].

The F1 score is the harmonic mean of precision and recall, resulting in a single statistic that balances both. It combines precision and recall into a single statistic that measures a classification model's overall performance. The F1 score is a value between 0 and 1, with a higher value indicating better performance. It is extremely beneficial when attempting to strike a balance between precision and recall, which is especially important in imbalanced datasets [17].

5.1. Random Forest

Random Forest is an ensemble learning technique that can be used for classification and regression. The algorithm is computationally efficient and capable of dealing with huge datasets [18]. Using the scikit-learn library to train a Random Forest classifier on the dataset and evaluate its performance using the various pre mentioned metrics. The Data Frame is separated from the characteristics (X) and the target variable (y). This is performed by appropriately assigning the Data Frame columns. The data is then separated between training and testing sets. The classifier is trained using the training set, and its performance is evaluated using the testing set.

To ensure the reliability and generalizability of the predictive models, we employed a couple of validation strategies. The dataset was divided into training and testing subsets using an 80/20 split. The training set was used to develop the models, while the testing set evaluated their performance on unseen data, ensuring that the models were not overfitted to the training data. Additionally, a 10-fold cross-validation technique was applied to further validate the models. This approach involves dividing the dataset into ten equal parts, using nine parts for training and one for testing in each iteration, and cycling through all parts. The results from these iterations were averaged to provide robust and unbiased performance metrics. This validation process ensured that the AUC scores reported for the RF, NN, and RIMARC algorithms were consistent and generalizable.

In addition to evaluation metrics for assessing the classification algorithm's performance, an essential output was the predicted success rate for each occurrence or case, allowing for more informed and correct decisions and the most beneficial benefit to patients. We generate success rates for each row using the Random Forest algorithm predictions. This method returns the probability estimates for each class label. Fig. 1 shows the success rate histogram for Random Forest and Fig. 2 shows the success rate histogram for Class 0 (Pregnant) and Class 1 (Not Pregnant) for Random Forest.

The evaluation metrics for the Random Forest predictive model demonstrate strong performance, with an Area Under the Curve (AUC) score of 0.98, indicating robust discrimination between positive and negative outcomes. The model exhibits high precision at 0.99, reflecting a minimal number of false positive predictions. Accuracy stands at 0.98, highlighting the overall correctness of predictions, while recall, or sensitivity, is at 0.96, indicating accurate identification of positive cases. The F1 Score, which balances precision and recall, is 0.98.

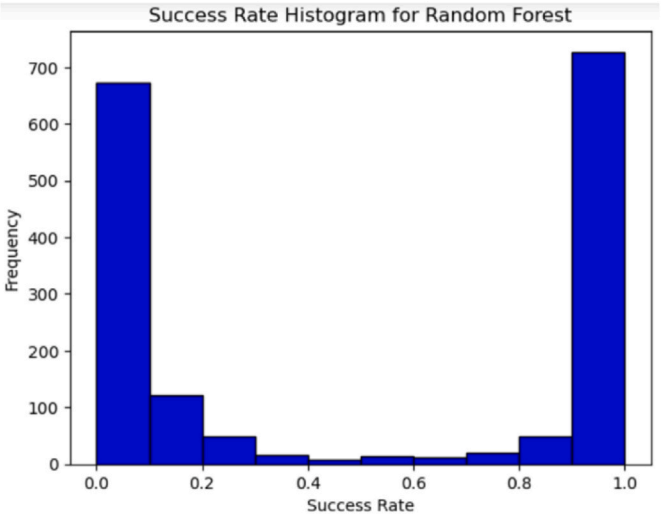


Fig. 1. The success rate histogram for Random Forest.

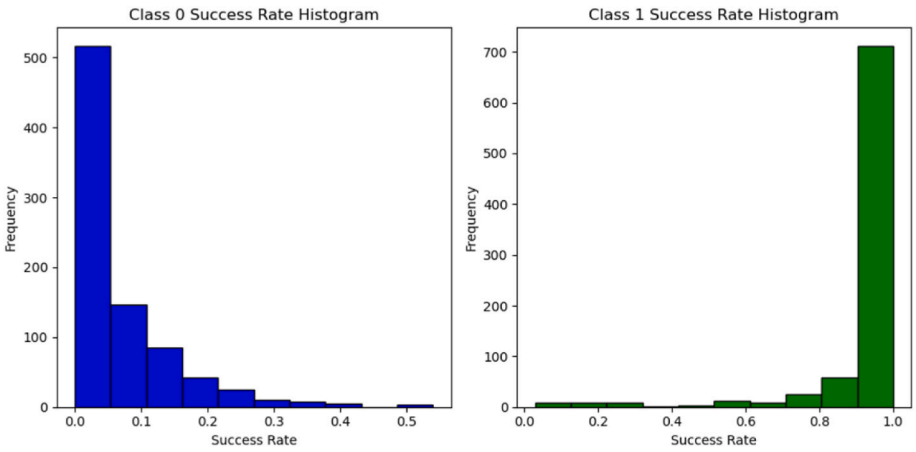


Fig. 2. The success rate histogram for Class 0 (Pregnant) and Class 1 (Not Pregnant) for Random Forest.

5.2. Neural Networks

Neural networks are a sophisticated and versatile classification system that has been successfully applied to a wide variety of classification issues across multiple fields. Its capacity to handle complicated relationships, flexibility in collecting both linear and nonlinear patterns, and ability to learn from vast volumes of data are all advantages for classification jobs. They do, however, necessitate careful tuning of hyperparameters, correct data preparation, and enough training data to avoid overfitting.

Using the scikit-learn library to train a Neural Networks classifier on the dataset and evaluate its performance using the various pre mentioned metrics. You split the data into training and testing sets. After that, normalizing features using standardization is a crucial step. Next, define a simple neural network architecture using the Sequential API from TensorFlow Library. The network consists of two hidden layers with 64 units each, using the ReLU activation function, and an output layer with a sigmoid activation function for binary classification then occurs. A compilation of the model with the Adam optimizer and binary cross-entropy loss. The model is then trained on the normalized training data. After training, predictions were conducted on the normalized test data. The predicted probabilities are then converted to binary predictions using a threshold of 0.5. Then, calculate evaluation metrics (accuracy, recall, F1 score, precision, and AUC). Fig. 3 shows the success rate histogram for Neural Network and Fig. 4 shows the success rate histogram for Class 0 (Pregnant) and Class 1 (Not Pregnant) for Neural Network.

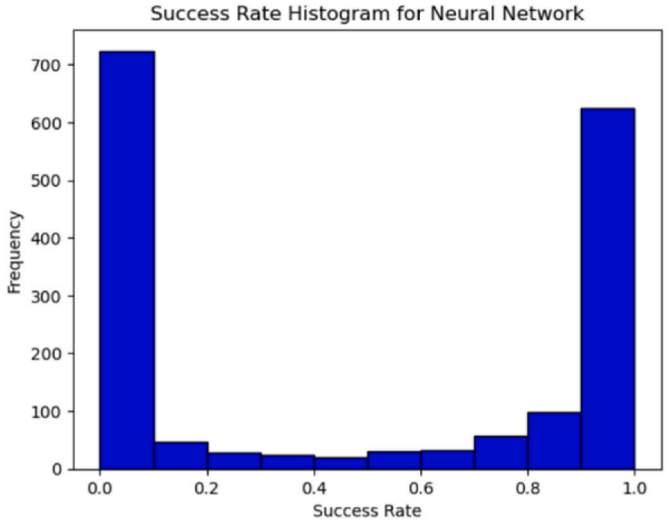


Fig. 3. The success rate histogram for Neural Network.

Finally, calculate the predicted success rate for each occurrence or case. The performance metrics for the Neural Networks predictive model are as follows: an AUC score of 0.95, indicating good discrimination between positive and negative outcomes; precision of 0.95, implying a relatively low number of false positives; accuracy at 0.95, signifying overall correctness of predictions; recall, or sensitivity, is at 0.96, indicating the model's ability to accurately identify positive cases; and an F1 Score of 0.95, effectively balancing precision and recall. The success rates for individual rows vary, with some instances showing high probabilities, such as 0.99, indicating confident success, while others display lower probabilities, suggesting less certainty in the corresponding success.

5.3. RIMARC algorithm

RIMARC is a supervised algorithm that learns a scoring function to rank instances. It does not make any assumptions about the data and has no parameters to tune for optimizing the performance. The RIMARC algorithm aims to maximize the AUC value, since the area under the ROC curve (AUC) has become a widely accepted performance evaluation metric in evaluating the quality of ranking. It learns a ranking function which is a linear combination of nonlinear score functions constructed separately for each feature. Each of these nonlinear score functions aims to maximize the AUC by considering only the corresponding feature in ranking. It has been shown that, for a single categorical feature, it is possible to derive a scoring function that achieves the maximum possible AUC [3]. The top 20 rules learned from RIMARC considering all features are presented in Table 2. Performance results of RIMARC are shown in Figs. 5 and 6.

The Results after applying the RIMARC algorithm to our dataset show that the AUC = 0.92, and the execution time is about 956 ms. An AUC value of 0.92 suggests that the model has a relatively high ability to discriminate between positive and negative instances. It implies that the model has a good balance between sensitivity and specificity, and it can make reasonably accurate predictions across different thresholds.

6. Discussion

Based on the results, RF algorithm achieved the highest AUC score of 0.97, followed by the NN with a score of 0.95, and the RIMARC algorithm with a score of 0.92. AUC is a widely used metric for evaluating the performance of binary classification models. It quantifies a model's ability to distinguish between positive and negative samples by measuring the area under the Receiver Operating Characteristic (ROC) curve. A higher AUC score indicates superior discriminatory power and

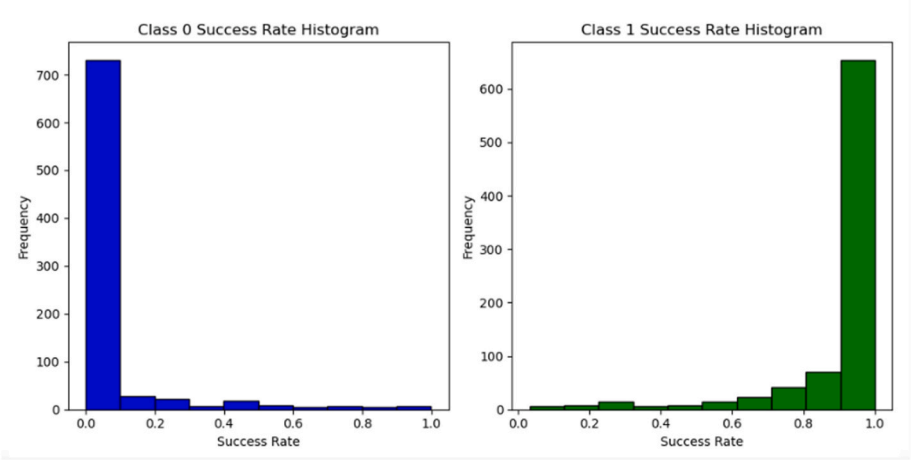


Fig. 4. The success rate histogram for Class 0 (Pregnant) and Class 1 (Not Pregnant) for Neural Network.

Table 2
Top 20 Cases Rules learned from RIMARC.

Feature	Score	Number of cases
FSH = "<0.5"	0.67	4881
LH = "<0.5"	0.67	4881
Ovarian surgery = "<0.5"	0.67	4880
HSG Tubes Hydrosalpinx RT = "<0.5"	0.67	4877
Abdominal surgery = "<0.5"	0.67	4876
Uterine Surgery = "<0.5"	0.67	4874
HSG Tubes Hydrosalpinx LT = "<0.5"	0.67	4873
PGD-FISH = "<0.5"	0.67	4870
Uterus status = "<0.5"	0.67	4868
Hypo Male = "<0.5"	0.67	4863
Hypo Female = "<0.5"	0.67	4858
Hystroscopy Cavity status = "<0.5"	0.67	4857
Endometrioma surgery = "<0.5"	0.67	4855
HSG Cavity status = "<0.5"	0.67	4854
PGD-PCR = "<0.5"	0.67	4853
Laproscopy Ovary LT status = "<0.5"	0.67	4853
Tubal Surgery = "<0.5"	0.67	4851
TESE = "<0.5"	0.67	4850
Laproscopy Ovary RT status = "<0.5"	0.67	4847
Laproscopy Tube LT status = "<0.5"	0.67	4840

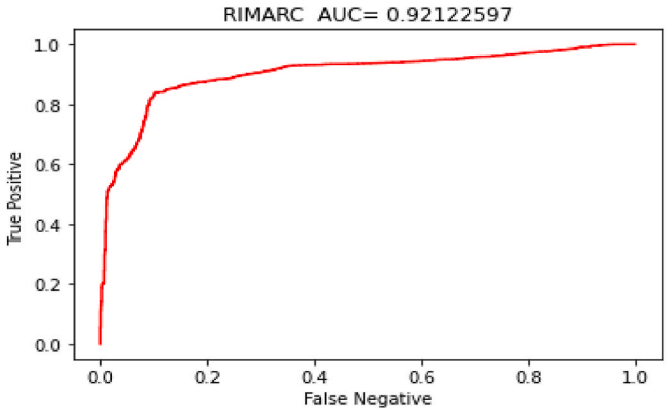


Fig. 5. Rimarc AUC.

overall model performance. Therefore, judging by the AUC scores, it appears that RF algorithm outperformed the other two algorithms in terms of the evaluated metric. However, it's essential to consider additional evaluation metrics and assess the overall performance and suitability of each algorithm. This becomes particularly relevant in the case of RIMARC because we only had access to the executable code of the

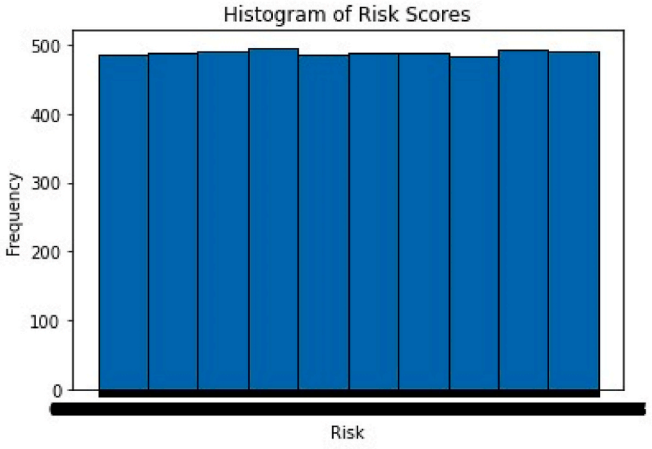


Fig. 6. Histogram of risk score for RIMARC.

algorithm.

Predicting the success rate of ICSI treatment using machine learning holds significant importance in the field of assisted reproductive technology. The results we obtained, with the RF algorithm achieving the highest AUC score, highlight the potential of machine learning in optimizing the decision-making process for both healthcare providers and patients. By accurately predicting the likelihood of success for ICSI treatment, patients can make more informed choices about their reproductive options, reducing emotional and financial burdens associated with multiple unsuccessful attempts. Healthcare providers can use these predictive models to tailor treatment plans, optimizing resource allocation and potentially improving overall success rates. Additionally, the insights gained from such models can contribute to a deeper understanding of the factors influencing ICSI success, driving advancements in fertility treatments and ultimately improving the chances of couples achieving their desired family goals.

The method employed in our analysis demonstrates considerable promise, practicality, and generalizability. Firstly, the use of machine learning algorithms, such as RF and NN, to predict the success rate of ICSI treatment showcases the potential of data-driven approaches in the field of assisted reproductive technology. These algorithms have shown robust performance, with the RF algorithm achieving a high AUC score, indicating their efficacy in making accurate predictions. Secondly, the practicality of our approach is evident in its applicability to real-world clinical settings. Healthcare providers and patients can readily benefit from these predictive models by incorporating them into the decision-

making process for fertility treatments. Patients can make more informed choices, while healthcare providers can customize treatment plans based on individualized risk assessments, optimizing resource allocation, and potentially improving success rates. Lastly, the generalizability of our method is significant. While we focused on ICSI treatment in this analysis, the underlying machine learning framework can be adapted and extended to other datasets and other areas of healthcare.

The principles of data-driven decision-making and predictive modeling can be applied to various medical disciplines, allowing for the development of tailored solutions that improve patient outcomes and resource utilization. Therefore, our method not only holds promise within the specific context of ICSI treatment but also offers a template for broader applications in healthcare and predictive analytics.

6.1. Practical implementation

To enhance the relevance of this study for clinicians, we include a section illustrating how the predictive models can be implemented into daily clinical practice. For example, consider a patient who presents with a specific set of attributes, such as age, hormonal levels, sperm quality, and previous IVF outcomes. Based on these features, the predictive models can estimate the likelihood of a successful ICSI treatment, helping clinicians make informed decisions on whether to proceed with the treatment or explore alternative options. For instance, if the model predicts a high probability of success based on the patient's profile, clinicians may decide to move forward with ICSI, while if the model predicts a lower chance of success, they might consider additional interventions or counseling. These predictive tools can be used to facilitate discussions between healthcare providers and patients, optimizing treatment planning and improving overall patient outcomes. By incorporating such real-world examples, we aim to demonstrate the practical utility of machine learning models in a clinical setting, helping clinicians understand how these tools can complement traditional decision-making processes in reproductive medicine.

7. Conclusion

The Study demonstrates the potential of machine learning techniques in enhancing clinical decision-making in the field of infertility treatment. These algorithms provide valuable insights into predicting fertility outcomes, including clinical pregnancy rates, implantation success, and ovarian response. By considering a range of clinical and biological factors, machine learning models offer personalized and precise predictions that can support clinicians in making informed decisions. However, further research and validation are needed to optimize these models and integrate them effectively into clinical practice. With continued advancements in machine learning and reproductive medicine, the utilization of these predictive models has the potential to improve pregnancy success rates and reduce the cost of assisted reproductive technologies.

CRediT authorship contribution statement

Abrar Mohammad: Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **Haneen Awad:** Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Huthaifa I. Ashqar:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Disclosure statement

No conflict of interest and no disclosure.

Ethics statement

In this study, the data were anonymized and collected retrospectively from the medical records of patients who received treatment at the Razan infertility center. As the data were used for research purposes only and patient identities were not disclosed, informed consent was not sought. However, we would like to clarify that this study was conducted in compliance with ethical standards.

Funding statement

No funding received for this project.

Declaration of competing interest

There is no conflict of interest to declare.

References

[1] Raef B, Ferdousi R. A review of machine learning approaches in assisted reproductive technologies. Acta Inf Med 2019;27(3):205–11. https://doi.org/10.5455/aim.2019.27.205-211. Avicena Publishing.
[2] Durairaj M. Data mining application on IVF data for the selection of influential parameters on fertility. In: *International Journal of Engineering and advanced technology (IJEAT)* (issue 2); 2013. https://www.researchgate.net/publication/262932487.
[3] Güvenir HA, Misirli G, Dilbaz S, Ozdegirmenci O, Demir B, Dilbaz B. Estimating the chance of success in IVF treatment using a ranking algorithm. Med Biol Eng Comput 2015;53(9):911–20. https://doi.org/10.1007/s11517-015-1299-2.
[4] Gunderson SJ, Puga Molina LC, Spies N, Balestrini PA, Buffone MG, Jungheim ES, Riley J, Santi CM. Machine-learning algorithm incorporating capacitated sperm intracellular pH predicts conventional in vitro fertilization success in normospermic patients. Fertil Steril 2021;115(4):930–9. https://doi.org/10.1016/j.fertnstert.2020.10.038.
[5] Radwan A, Amarneh M, Alawneh H, Ashqar HI, AlSobeh A, Magableh AAAR. Predictive analytics in mental health leveraging LLM embeddings and machine learning models for social media analysis. Int J Web Serv Res 2024;21(1):1–22.
[6] Hu X. DB-HReduction: a data preprocessing algorithm for data mining applications. Appl Math Lett 2003;16(6):889–95.
[7] Wang R, Pan W, Jin L, Li Y, Geng Y, Gao C, Chen G, Wang H, Ma D, Liao S. Artificial intelligence in reproductive medicine. Reproduction 2019;158(4):R139–54.
[8] Kafaee Ghaeini M, Amin-Naseri MR, Aghahoseini M. Prediction of clinical pregnancy occurrence after ICSI using decision tree and support vector machine methods. Journal of Obstetrics, Gynecology and Cancer Research 2018;3(4):12–8. https://doi.org/10.30699/jogcr.3.4.12.
[9] Mehrjerd A, Rezaei H, Eslami S, Ratna MB, Khadem Ghaebi N. Internal validation and comparison of predictive models to determine success rate of infertility treatments: a retrospective study of 2485 cycles. Sci Rep 2022;12(1). https://doi.org/10.1038/s41598-022-10902-9.
[10] Ottosen LDM, Kesmodel U, Hindkjær J, Ingerslev HJ. Pregnancy prediction models and eSET criteria for IVF patients - do we need more information? J Assist Reprod Genet 2007;24(1):29–36. https://doi.org/10.1007/s10815-006-9082-9.
[11] Hafiz P, Nematollahi M, Boostani R, Jahromi BN. Predicting implantation outcome of in vitro fertilization and intracytoplasmic sperm injection using data mining techniques. International Journal of Fertility & Sterility 2017;11(3):184.
[12] Meijerink AM, Cissen M, Mochtar MH, Fleischer K, Thoonen I, De Melker AA, Meissner A, Repping S, Braat DDM, VanWely M, Ramos L. Prediction model for live birth in ICSI using testicular extracted sperm. Hum Reprod 2016;31(9):1942–51. https://doi.org/10.1093/humrep/dew146.
[13] De Santiago I, Polanski L. Data-driven medicine in the diagnosis and treatment of infertility. J Clin Med 2022;11(21):6426.
[14] Dehghan S, Rabiei R, Choobineh H, Maghooli K, Nazari M, Vahidi-Asl M. Comparative study of machine learning approaches integrated with genetic algorithm for IVF success prediction. PLoS One 2024;19(10):e0310829.
[15] Sadegh-Zadeh S-A, Khanjani S, Javanmardi S, Bayat B, Naderi Z, Hajiyavand AM. Catalyzing IVF outcome prediction: exploring advanced machine learning paradigms for enhanced success rate prognostication. Frontiers in Artificial Intelligence 2024;7:1392611.
[16] Chen Z, Zhang D, Zhen J, Sun Z, Yu Q, Yin Y. Predicting cumulative live birth rate for patients undergoing in vitro fertilization (IVF)/intracytoplasmic sperm injection (ICSI) for tubal and male infertility: a machine learning approach using

XGBoost. In: Chinese medical journal, vol. 135. Lippincott Williams and Wilkins; 2022. p. 997–9. https://doi.org/10.1097/CM9.0000000000001874. Issue 8.

[17] Ver Milyea M, Hall JMM, Diakiw SM, Johnston A, Nguyen T, Perugini D, Miller A, Picou A, Murphy AP, Perugini M. Development of an artificial intelligence-based assessment model for prediction of embryo viability using static images captured by optical light microscopy during IVF. Hum Reprod 2021;35(4):770–84. https://doi.org/10.1093/HUMREP/DEAA013.

[18] Doan T, Kalita J. Predicting run time of classification algorithms using meta-learning. International Journal of Machine Learning and Cybernetics 2016;8(6): 1929–43. https://doi.org/10.1007/S13042-016-0571-6/FULLTEXT.HTML.