

CNFE-SE: A novel hybrid approach combining complex network-based feature engineering and stacked ensemble to predict the success of Intrauterine Insemination and ranking the features

Sima Ranjbari, M.S.C.¹, Toktam Khatibi, Ph.D.^{1*}, Ahmad Vosough Taghi Dizaj, M.D.², Hesamoddin Sajadi, M.D.⁴, Mehdi Totonchi, Ph.D.^{3,4*}, Firouzeh Ghaffari⁵

1. School of Industrial and Systems Engineering, Tarbiat Modares University
2. Department of Genetics at Reproductive Biomedicine Research Center, Royan Institute for Reproductive Biomedicine, ACECR, Tehran, Iran, email: vosough@royaninstitute.org.
3. Department of Reproductive Imaging, Reproductive Biomedicine Research Center, Royan Institute for Reproductive Biomedicine, ACECR, Tehran, Iran.
4. Department of Andrology, Reproductive Biomedicine Research Center, Royan Institute for Reproductive Biomedicine, ACECR, Tehran, Iran.
5. Department of Endocrinology and Female Infertility, Reproductive Biomedicine Research Center, Royan Institute for reproductive biomedicine, ACECR, Tehran, Iran, email: ghafaryf@yahoo.com

*Corresponding Authors: Toktam Khatibi
Email: toktam.khatibi@modares.ac.ir

Abstract

Background: Intrauterine Insemination (IUI) outcome prediction is a challenging issue with which the assisted reproductive technology (ART) practitioners are dealing. Predicting the success or failure of IUI based on the couples' features can assist the physicians to make the appropriate decision for suggesting IUI to the couples or not and/or continuing the treatment or not for them. The large number of studies that have been focused on predicting the IVF and ICSI outcome by machine learning algorithms. But, to the best of our knowledge, a few studies have been focused on predicting the outcome of IUI. The main aim of this study is to develop an automatic classification and feature scoring method to predict intrauterine insemination (IUI) outcome and ranking of the most significant features, based on patients' cycle's characteristics under IUI treatment.

Methods: For this purpose, a novel hybrid approach combining complex network-based feature engineering and stacked ensemble (CNFE-SE) is proposed. Three complex networks are extracted considering the patients' data similarities. The feature engineering step is performed on the complex networks. The original feature set and/or the features engineered are fed to the proposed stacked ensemble to classify and predict IUI outcome for couples per IUI treatment cycle. Our study is a retrospective study of a 5-year couples' data undergoing IUI. Data is collected from Reproductive Biomedicine Research Center, Royan Institute for 8,360 couples who underwent 11,255 IUI cycles were included.

Results: Experimental results show that the proposed method outperforms the compared methods with AUC of 0.84 ± 0.01 , sensitivity of 0.79 ± 0.01 , specificity of 0.91 ± 0.01 , and accuracy of 0.85 ± 0.01 for the prediction of IUI outcome.

Conclusions: The most important predictors for predicting IUI outcome are semen parameters (sperm motility and concentration) as well as female BMI.

Key Words: IUI outcome prediction, complex networks, feature engineering, stacked ensemble classifier, feature selection

1 Background

Infertility is defined as the failure of the female partner to conceive after at least one year of regular unprotected sexual intercourse [1]. More than 186 million people are suffering from infertility in the worldwide that most of them are living in developing countries [2]. In most cases, the causes of infertility are not clear, which complicates the treatment procedure. These problems have been exacerbated for several reasons, such as lifestyle changes, infection, and genetic issues. In many cases, the only way to get pregnant has been through the use of assisted reproductive technology (ART), and its performance has not yet been optimized [3].

Every year, more than 1.5 million ART cycles are carried out all over the world [4]. ART consists of three basic procedures including intrauterine insemination (IUI), in-vitro fertilization (IVF), intracytoplasmic injection (ICSI) which are generally carried out in different steps of the treatment [5]. The first-line treatment, second and the third stages of ART are IUI, IVF, and ICSI, respectively [6]. In comparison with other sophisticated methods of ART, IUI has been considered as the easiest, minimally invasive and less expensive one. Most of the recent researches have shown the efficacy of IUI [6, 7].

IUI outcome prediction is a challenging issue with which the ART practitioners are dealing. Predicting the success or failure of IUI based on the couples' features can assist the physicians to

make the appropriate decision for suggesting IUI to the couples or not and/or continuing the treatment or not for them [5].

Machine Learning approaches, as the modern scientific discipline, concentrates on how to detect hidden patterns and get the information from data. Machine learning provides different methods and algorithms to predict the output from some input predictors which can be used for clinical decision making [8].

The large number of studies have been focused on predicting the IVF and ICSI outcome by machine learning algorithms as summarized in Table 1.

Table 1- summarizing the previous studies of predicting ART outcome prediction

As illustrated by Table 1, the previous studies related to outcome prediction of ART methods are listed which have analyzed data using data mining and/or statistical methods. For this purpose, classifiers such as Decision Tree (DT), Logistic Regression (LR), Naïve Bayes (NB), K-Nearest Neighbors (K-NN), Support Vector Machines (SVM), Multi-Layered Perceptron (MLP) and Radial Basis Function (RBF) as Artificial Neural Networks (ANN) and Random Forest (RF) have been used in the previous studies for predicting the clinical pregnancy after the complete cycles of different ART methods. A main drawback of the most of the considered previous studies is small volume of dataset and a few number of features. Small dataset increases the risk of overfitting the models while training on this dataset. Overfitting is occurred when a trained model has good performance for predicting the output of the training dataset but show poor performance while applying on the test dataset. Models with high overfitting property has less generalization ability. In this study, the larger volume of data records with many more features are considered for IUI outcome prediction which is a main advantage of this study compared to the considered previous studies.

On the other hand, most of the previous studies have considered the outcome prediction for IVF or ICSI. To the best of our knowledge, a few studies have been focused on predicting the outcome of IUI which have used clustering methods [9, 10] or regression analysis [11].

The previous studies which have been based on regression analysis only have considered the weights of the independent features to predict the overall pregnancy probability and they have not assessed the interconnection among the features [11-17]. Besides, the number of the patients have been studied in the previous researches has been small, these studies have suffered from the lack of statistical power [17, 18]. Also, the AUC performance of the previously proposed models for predicting IUI outcome have been low [12]. Therefore, it is required to improve the prediction performance by proposing novel classifiers and considering more data records.

Most of the considered previous studies have used single classifiers and/or RF as a simple ensemble classifier. Some previous studies have illustrated that the stacked models can improve the classification performance for other applications and other datasets [19-21]. Therefore, in this study, a novel stacked ensemble is designed and proposed for improving the performance of IUI outcome prediction.

The main aim of this study is to develop an automatic classification and feature scoring method to predict intrauterine insemination (IUI) outcome and ranking of the most significant features, based on patients' cycle's characteristics under IUI treatment. For this purpose, a novel hybrid approach combining complex network-based feature engineering and stacked ensemble (CNFE-SE) is proposed. Three complex networks are extracted considering the patients' data similarities. The feature engineering step is performed on the complex networks. The original feature set and/or the features engineered are fed to the proposed stacked ensemble to classify and predict IUI outcome

for couples per IUI treatment cycle. Our study is a retrospective study of a 5-year couples' data undergoing IUI. Data is collected from Reproductive Biomedicine Research Center, Royan Institute for 8,360 couples who underwent 11,255 IUI cycles were included.

The main novelty of this study lies in four folds including:

- Proposing a method for feature scoring and classification based on weighted complex networks and stacking ensemble classifiers
- Proposing feature engineering method based on complex networks
- Designing a novel stacked ensemble classifier for predicting IUI outcome

2 METHODS

The main steps of the proposed hybrid approach combining complex network-based feature engineering and stacked ensemble (CNFE-SE) to predict the success of Intrauterine Insemination and ranking the features are illustrated in Figure 1.

Figure 1- the main steps of the proposed method (CNFE-SE) for feature scoring and classifying the patients to predict IUI outcome

The main steps of the proposed method (CNFE-SE) as depicted in Figure 1 include the modules for data collection and preparation, feature scoring and classification and finally model evaluation and validation. The first module consists of data collection, sampling from data, preprocessing the collected data and filtering irrelevant features. In the next module, ignoring a feature, constructing three complex networks from the patients, extracting features from the constructed complex networks, training the classifiers based on the extracted features and finally scoring the ignored feature are performed. The last module evaluates and validates the models trained in the previous module. More details about the mentioned tasks are described in the following subsections.

2.1 Data collection

Dataset studied in this article is collected from Royan Institute, a public none-profitable organization, affiliated to the academic center for education, culture and research (ACECR) in Iran. It includes the features describing the patients having been treated by IUI method in the Infertility clinic at Royan Institute between January 2011 and September 2015. 11,255 IUI cycles related to 8,360 couples are considered in which the women were aged between 16 and 47 with the mean of 29. This dataset contains 1,622 positive outcomes and 9,633 negative ones. Therefore, the overall pregnancy rate is 14.41% per completed cycle and 19.4% per couple. Each couple is treated for 1.31 ± 0.59 (mean \pm SD) IUI cycles (range, 1-7 cycles).

The features describe patients based on their demographic characteristics, historical data, about the patients' diseases, the clinical diagnosis, the treatment plans and the drugs which have been prescribed, semen quality, laboratory tests and the clinical pregnancy outcome. The considered demographic features include age, BMI, education, the relationship between the couples and so on. The information about the history of the patients' subfertility consists of the duration and type of infertility, length of marriage and so on.

Among the considered features; there are 86 continuous, 152 binaries, 51 nominal and 7 binominal variables. The detailed description of the data is presented in Appendix A.

In the collected dataset, the majority of patients (almost 72%) have been treated for one cycle, 22% two cycles, 5% three cycles, and less than 1% have been treated up to seven cycles. Figure 2 depicts the distributions of positive and negative clinical pregnancy rates for patients per treatment cycle.

Figure 2 –The ratio of positive and negative clinical pregnancy per treatment cycle

As illustrated by Figure 2, 63% of the couples belonging to the positive class (positive clinical pregnancy after complete cycle) have been pregnant after the first treatment cycle. 26% of data records in the positive class have received positive outcome after the second cycle. Moreover, 74% of the couples in the negative class have been considered after the first cycle.

2.2 Data sampling

Data should be randomly partitioned into training and testing datasets with no overlapping among these two subsets. The models are trained on the training dataset and finally are evaluated by applying to the testing datasets.

K-fold cross validation (C.V.) is a common and popular sampling strategy used for this purpose. In this method, data is randomly divided into K disjoint equal-size subsets. Every time, one of these K subsets are considered as the testing dataset and all the remaining subsets consist the training one. The model is trained K times on K training datasets and applied to the corresponding testing datasets to evaluate the performances of the trained models.

Before sampling from data, patient records with high missing value rate are excluded from the study and then, 5-fold C.V. is used for sampling from the collected dataset, in this study.

2.3 Data preprocessing

Preprocessing of data is one of the most essential steps in knowledge discovery. The study of Zhang et al affirms that 80% of total time in data mining processing is allocated for preparation step [22].

In the first step, the initial collected dataset includes almost 86,000 data records (each one describes of each partner) and about 1,000 features. The data records describing one couple per IUI treatment cycle are aggregated and form the same data record.

Thus, the aggregated dataset includes 11,255 data records and about 296 features describing a couple during an IUI treatment cycle.

The missing values for numeric and categorical features are imputed based on the average and the most frequent values, respectively [23].

2.4 Filtering irrelevant features

Since the aggregated dataset consists of the large number of the features, the irrelevant features can be removed to reduce the computational time required for processing and analyzing data. Thus, the features having very low correlation with the output feature or very high correlation with other input features are excluded from this study. The linear correlation coefficient between pairs of the features F_p and F_q are calculated as Eq. (1):

$$Corr(F_p, F_q) = \sum_i \frac{(F_{i,p} - m_p)(F_{i,q} - m_q)}{\sqrt{\sum_j (F_{j,p} - m_p)^2} \sqrt{\sum_j (F_{j,q} - m_q)^2}} \quad (1)$$

Where $F_{x,p}$ ($F_{x,q}$) indicates the x^{th} row of the feature F_p (F_q) and m_p (m_q) denotes the average of the feature F_p (F_q), respectively.

If two features F_p and F_q have low (high) correlation, $Corr(F_p, F_q)$ tends to zero (-1 or +1).

2.5 Ignoring a feature

Breiman has proposed measuring the feature importance by mean decrease in accuracy (MDA) of random forest [24]. This study aims at ranking the features according to their predictive power for classifying the instances to positive or negative clinical pregnancy. For this purpose, all the steps

6-9 are performed by considering all the features excluding one feature each time and MDA for the trained proposed classifier is calculated on the validation dataset. MDA values show the amount of reducing the model accuracy while removing a feature. Therefore, higher values of MDA indicate the higher predictive ability of the features.

2.6 Constructing complex networks of patients

For modeling nonlinear data, complex networks are effective method [25]. Complex network is a weighted undirected graph $G = (V, E, W)$, where V is the set of nodes, E denotes the set of edges $e(v_i, v_j)$ between the pairs of the nodes v_i and v_j and W is the weights $w(v_i, v_j)$ assigned to the edges $e(v_i, v_j)$ of E .

Three complex networks are constructed from the training datasets. The first one is comprised of all the training data records as its nodes and is called CN1. The second and the third complex networks consist of all training data records excluding the negative and positive classes and named as CN2 and CN3, respectively.

In other words, the nodes of CN1, CN2 and CN3 are all the training data records, positive labeled and negative labeled training data records, respectively.

An edge between node v_i and v_j is drawn if the distance between the input features of the i^{th} and j^{th} training data records is smaller than a user-defined threshold. For calculating the pairwise distance between data records, Euclidean distance function is used and can be calculated as Eq. (2):

$$Distance(v_i, v_j) = \sqrt{\sum_{p=1}^m (F_{i,p} - F_{j,p})^2} \quad (2)$$

Where m is the number of the input features, $F_{i,p}$ and $F_{j,p}$ denote the p^{th} input feature values for data records corresponding to v_i and v_j .

The weight of the edge $e(v_i, v_j)$ is calculated as Eq. (3):

$$w(v_i, v_j) = \frac{distance(v_i, v_j)}{\max(distance(v_k, v_h); \forall v_k, v_h \in V)} \quad (3)$$

2.7 Feature engineering based on the complex networks

In this section, three complex networks per node are constructed including all training instances as CN1 and all training instances excluding negative (positive) instances as CN2 (CN3). A simple intuitive hypothesis is that a node has more similarity with the training instances of its own class compared to the instances of the other class. Therefore, the node centrality in different complex networks CN1, CN2 and CN3 can be compared to classify the node. Features listed in Table 2 are defined based on this idea.

Table 2- list of the features engineered from the complex networks in this study

Node degree is the number of its adjacent edges. Betweenness centrality for graph nodes have been introduced by Bavelas [26] and is calculated as Eq. (21). If a node lies in many shortest paths between pairs of nodes, its Betweenness centrality will be high. Nodes with high Betweenness centrality are the bridges for information flow.

$$Betweenness(v_i) = \sum_{j < k} \frac{\text{number of the shortest paths between } v_j \text{ and } v_k \text{ passing } v_i}{\text{number of the shortest paths between } v_j \text{ and } v_k} \quad (21)$$

Node closeness centrality measures the reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the graph.

Node Eigen vector centrality is higher when the node is pointed to by many important nodes.

Clustering coefficient of a node is calculated as Eq. (22):

$$\text{clusteringCoefficient}(v_i) = \frac{\text{number of triangles connected to } v_i}{\text{number of triples centered around } v_i} \quad (22)$$

Since, the number of the instances are very high, the complex networks are partitioned into smaller communities to reduce the computational complexity for calculating the engineered features.

One complex network extracted from only 100 data records treated by IUI method as a sample is shown in Figure 3.

Figure 3- One complex network extracted from only 100 data records treated by IUI method

Figure 4 depicts two complex networks of the same samples of positive instances drawn by different thresholds.

Figure 4- two complex networks drawn from the positive training data samples by (a) threshold of 0.7 * average of the distance Matrix, (b) threshold of 0.5 * average of the distance matrix

As shown in Figure 4, reducing the threshold for keeping the edges in the complex network even with a small decrease lead to the network with more sparsity and more small-sized communities.

Figure 5 illustrates three complex networks from the samples of both classes, negative and/or positive classes.

Figure 5- three complex networks extracted from the samples of (a) both classes, (b) negative class, and (c) positive class with the same threshold

As shown by Figure 5, for the same thresholds, complex network considering the instances of both classes has the most density and the complex network from only positive instances has the most sparsity and consists of several small communities.

2.8 Training the stacked ensemble classifier

Stacked ensemble classifier which is a scalable meta-modeling methodology has been first introduced by Wolpert in 1994 [27]. It has been inspired by neural networks whose classifiers have been considered as the nodes. Instead of a linear model, the stacked classifier can use any base classifier. The stacking operation has been performed by either a normal stacking or a re-stacking mode. In the normal stacking mode, the base classifiers in each layer, use the output scores of the previous ones as the predictors similar to a typical feedforward neural network. The formula of normal stacking is written as Eq. (23):

$$f_n(x, V) = V_{n,k} \left(f_{n-1}(x, V_{n-1,1}), f_{n-1}(x, V_{n-1,2}), \dots, f_{n-1}(x, V_{n-1,D_{n-1}}) \right) \quad (23)$$

Where N is the size of hidden layers, n indicates the nth layer of the N, x denotes a sample of a dataset, V presents a vector holding the neurons (the base classifiers), D is the number of hidden neurons through the nth hidden layer and finally, k is the kth neuron in the nth layer.

Some previous studies have illustrated that the stacked models can improve the performance of the classification [20, 21]. Therefore, in this study, a new stacked ensemble classifier is proposed and designed based on the normal stacking mode. In the beginning, some of the basic classifiers are trained, and those outperforming the others have been selected to be considered as the base classifiers in the stacked ensemble layers. The architecture of the proposed stacked ensemble classifier is shown in Figure 6.

Figure 6- (a) input datasets, and (b) the architecture of the proposed Stacked Ensemble classifier

As illustrated in Figure 6, input dataset consists of the features in OFS, FS-Fi, EFS and/or EFS-Fi. Input dataset is fed to the base classifiers in the first layer of the proposed stacked ensemble classifier. Logistic regression (LR) [28], support vector machines (SVM) [29], decision tree (DT) [30], random forest (RF) [24], Adaboost [31] and LightGBM [32] are the base classifiers in both layers.

LR, SVM with linear kernel and DT are appropriate classifiers for classifying linearly separable data. RF, Adaboost and LightGBM are ensemble classifiers which can classify nonlinearly separable data with high performance. All the mentioned classifiers can be trained fast. Therefore, they are chosen as the base classifiers of the proposed stacked ensemble classifier.

After training the base classifiers in the first layer, their outputs are considered as Meta features according to the normal stacking mode. The Meta features are fed into the base classifiers of the second layer for training them. Finally, the outputs of the base classifiers in the second layer are aggregated by weighted voting aggregation rule.

The weight of each base classifier is obtained by measuring its accuracy for classifying the validation dataset. The validation dataset is about 20% of the original training dataset which is excluded while training the base classifiers in both layers.

2.9 Scoring the ignored feature

As mentioned in section 1.5, MDA score is calculated for each feature and is considered as the feature importance score.

2.10 Evaluating and validating the trained models

To evaluate the performances of the trained models, performance measures for classification problems are used in this study including Accuracy, Sensitivity, Specificity and F-Score as shown in Eq. (24)-(27):

$$Accuracy = \frac{TP + TN}{N} \quad (24)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (25)$$

$$Specificity = \frac{TN}{TN + FP} \quad (26)$$

$$F - Score = 2 \times \frac{Sensitivity \times Specificity}{Sensitivity + Specificity} \quad (27)$$

Where TP and FP (TN and FN) indicate the number of instances in the positive (negative) classes which are classified correctly and incorrectly, respectively.

Moreover, the area under the curve (AUC) of the receiver operating curve is considered.

In order to validate the results, the experiments are repeated 50 times, and each time the data is selected based on 5-fold C.V.

3 Experimental results

In this section, the performance of the proposed model (CNFE-SE) is compared with other state-of-the-art classifiers. The features are ranked based on MDA while ignoring them during the training of CNFE-SE. Finally, the partial dependencies between high-ranked features are discussed.

3.1 Comparing the performance of CNFE-SE with other state-of-the-art classifiers

Table 3 lists the performance measures for comparing CNFE-SE with other state of the art classifiers.

Table 3- comparing the performance of CNFE-SE with other state of the art classifiers

The models are executed and trained on different random training samples up to 50 times and the mean \pm standard deviation values are depicted in Table 3. The CNFE-SE outperforms the compared models by AUC of 0.84 ± 0.01 , sensitivity of 0.79 ± 0.01 , specificity of 0.91 ± 0.01 , and accuracy of 0.85 ± 0.01 .

Table 4 shows the confusion matrix of CNFE-SE for total dataset.

Table 4- the confusion matrix of CNFE-SE for total dataset

Figure 7 depicts ROC curve for CNFE-SE trained with all features.

Figure 7- ROC curve for CNFE-SE trained with all features

As shown by Figure 7, AUC of CNFE-SE trained on all features is 0.84 ± 0.01 . As illustrated by Table 3, the compared classifiers show almost weak performances. The main reason is that the patients treated with IUI do not have complicated conditions and the leading cause of their infertility is idiopathic. Therefore, the data of the two classes have high similarity with each other, and their differentiation by single classifier was not an easy task. However, among these models, Light-GBM as one of state-of-the-art machine learning algorithms has the second best performance because it is a gradient boosting framework that uses tree-based learning algorithms and not only covers multi hyper-parameters but also has more focus on the accuracy of the results [32].

When the classes are imbalanced, Precision-Recall curve is a useful instrument for the presentation of prediction success. A great area under this curve shows both high precision, which is related to low false-positive rate, and high recall, refers to low false-negative rate. Figure 8 indicates the precision-recall curves for CNFE-SE.

Figure 8 -Precision-recall curves for CNFE-SE

As shown in Figure 8, CNFE-SE predicts positive class better than negative one.

Ranking the significance of features

Figure 9 represents top-20 important features with highest mean decrease of accuracy score influencing the IUI outcome in 50 repetitions of CNFE-SE execution and training on different training samples. Post wash total motile sperm counts, female BMI, sperm motility grades a + b, total sperm motility, and sperm motility grade c have been ranked as the important features for predicting IUI outcome. Additionally, post-wash total motile sperm counts, female BMI, and total sperm counts are the features illustrated with dark blue colors in Figure 9, have the highest repetitions as the first informative features. Generally, the variables relating to the men's semen analysis parameters are high-ranked features in this study.

Figure 9-Overview of top features ranked based on CNFE-SE

3.2 Partial dependency between the features

Figure 10 depicts the partial dependency plots for the most important features. Partial dependency plots show whether a feature has a positive or negative effect on the response variable when the other ones are controlled. However, in order to interpret the graphs, we should note that shifts in clinical pregnancy probabilities over the value of the features, even the most significant ones, are roughly small (the y-axis range is 0.44-0.52). Therefore, it is noteworthy that none of the features could individually and significantly alter the pregnancy rates above the 0.52 threshold. This finding underlines the value of the machine learning approach by determining the complicated association between individually delicate predictors to make an effective classification model.

Figure 10- Partial dependency plots of nine features among the important features which the blue curves indicate locally weighted smoothing. It shows pregnancy variation obtaining from CNFE-SE (y-axis) as a function of a feature (x-axis) in IUI.

According to the results of the partial dependency plots, the clinical pregnancy rate has raised with increased number of post-wash total motile sperm counts and after processing sperm concentration. Also, when their values respectively vary upper than 100 million and 30 million spermatozoa per ml, the rate of pregnancy reaches its highest rate. In addition, the likelihood of IUI success increases through growing the number of total sperm counts which is mentioned in the previous studies, too [33].

4 DISCUSSION

In the current study, among the various features that significantly affect the IUI outcome, the most potential predictive ones are female BMI and semen quality parameters. Semen data, i.e. sperm count and motility are illustrated as the most prognostic factors in pregnancies, conceived by IUI and key features of some previous studies [34]. Moreover, some previous studies confirm that semen descriptors, after the swim-up procedure are more important than the ones before sperm washing process [35, 36]. Similarly, the percentage of motile sperm and its progression in the ejaculate have been known as significant predictors in IUI outcome prediction in the literature [37]. Sperm motility grades a + b (progressive motility) and grade d (immotile sperms) are also determined in this study as potential predictive factors for a successful IUI. Thus, when each of one is more than 20% and less than 15%, the IUI success rate is higher.

Furthermore, the outcome of this study indicates that the IUI success rate is almost low when the female BMI is abnormal (BMI is less than 20 or larger than 30) and as much as it changes to a normal range of about 25, the probability of pregnancy increases. This finding is mentioned in the previous studies, too [38].

Meanwhile, reducing pregnancy rate can be mitigated by increase in the female age [39]. The present study identifies that the women older than 38, have a lower chance of successful IUI. However, Edrem et al. have not found the female age to be a prognostic factor in the prediction of IUI outcome [40].

As shown in the figure 8, the duration of infertility inversely affects the fertility rate, and the decline in fecundity is acclaimed by some previous works, as well. Also, the previous studies have shown that when the couples' duration of infertility is less than six years, the pregnancy success rate is higher [41].

The total dose of gonadotropins is taken into account in this study as an important feature in which its significance has been considered recently, too [11]. This study identifies that the total dose of gonadotropin is positively correlated with the pregnancy rate. Moreover, other factors contributing to failure or success of IUI outcome according to this study's findings include semen volume, male age, sperm normal and amorphous morphology, duration of the marriage, and endometrial thickness which some of them have been considered as the influential attributes in some previous similar studies [42-44].

Eventually, the CNFE-SE is rebuilt with these 20 most important features and it yields surprisingly good performances (AUC =0.82, 95% CI 0.80-0.84). It shows that the model carried out by these features, demonstrates a highly reasonable performance.

Some studies consider different patients' cycles as independent of each other, which may lead to a biased result, so they considered the first cycle information within a couple [16, 45]. Our reanalysis of the primary cycle data revealed that the AUC performances of Light-GBM and CNFE-SE are 0.62 ± 0.01 , 0.84 ± 0.01 which did not change significantly when all the cycles were taken into account. Moreover, as shown in the materials and methods section, increasing number of cycles augment the clinical pregnancy rate which are in line with the importance of this feature in subsequent IUI outcome [46, 47]. On the contrary, the variable cycle number has not identified as an important feature according to CNFE-SE feature scores. This finding may be due to the high number of data in the first cycle compared to the rest of cycles, which accounts for approximately 74% of the data.

Finally, our study has some restrictions. Some of the female hormonal tests including FSH, TSH, LH, and AMH have not been measured in all the patients before beginning IUI cycle, and therefore they are eliminated from the analysis due to a large number of their missing data. At the Royan center, the patients who are entering the IUI treatment cycles are those who do not have complicated conditions, and the women's hormonal tests are usually normal. Moreover, the male BMI is excluded because of its high rate of missing values. The features describing the geographic information of couples' habitats are removed from the study due to their inappropriate data entry.

Currently machine learning algorithms has been increasingly employed in different medical fields [8]. Therefore, through using machine learning methods, we are able to predict the success or failure of the IUI cycle treatment outcome for each couple, based on their demographic

characteristics and cycle information. In other words, our proposed CNFE-SE model shows superior performance among the compared state of the art classifiers. A decision support system (DSS) can be designed and implemented based on CNFE-SE. This DSS can help the physicians to choose other treatment plans for the couples and reduce patients' costs if their IUI cycle success rate is low. The schematic of this medical assistance system is shown in Figure 11.

Figure 11- Schematic of our proposed medical decision support system for IUI outcome prediction

The proposed DSS is trained on the training dataset by CNFE-SE after preprocessing the collected dataset. After completing the training of CNFE-SE, every time a new data record is registered in the DSS, it can be classified by CNFE-SE into positive or negative outcome. The predicted outcome for the new data record can assist the physicians to decide to treat the couple with IUI method or not.

Conclusion:

In conclusion, the use of machine learning methods to predict the success or failure rate of the IUI could effectively improve the evaluation performances in comparison with other classical prediction models such as regression analysis. Furthermore, our proposed CNFE-SE model outperforms the compared methods with highly reasonable accuracy. CNFE-SE can be used as clinical decision-making assistance for the physicians to choose a beneficial treatment plan with regards to their patients' therapy options, which would reduce the patients' costs as well.

The most important features for predicting IUI outcome are semen parameters (sperm motility and concentration) as well as female BMI.

Some features which have been identified as good discriminative features for IUI outcome prediction in the previous studies are excluded from this study because of their high missing value rate. For example, some of the female hormonal tests including FSH, TSH, LH, and AMH are not routinely measured in all the patients before IUI and they are excluded from the study. It is proposed to augment dataset with data records without missing value in the mentioned features and consider the excluded features to CNFE-SE, and then try to rank the augmented feature set and evaluate the performance of the classifier.

On the other hand, some data records have noisy information which can reduce the performance of the classifiers. As future work, it is suggested that improving the robustness of CNFE-SE against the noisy data by including vote-boosting and other previously proposed methods for increasing the noise robustness of the classifiers. Moreover, the data is highly imbalanced which can have negative effect on the classifiers' performance. As another research opportunity, it is suggested that reducing the influence of data distribution per class by incorporating the balanced sampling strategies combined with k-fold cross validation.

Abbreviations:

ACECR: academic center for education, culture and research

AMH: Anti-müllerian hormone

ANN: Artificial neural networks

ART: assisted reproductive technology

AUC: Area under curve

BMI: Body mass index

CN1: Complex network which is comprised of all the training data records as its nodes

CN2: Complex network which includes all training data excluding negative class

CN3: Complex network which includes all training data excluding positive class
 CNFE-SE: Complex network-based feature engineering and stacked ensemble
 C.V.: Cross validation
 DSS: Decision support system
 DT: Decision tree
 FN: False negative
 FP: False Positive
 FSH: Follicle-stimulating hormone
 HCA: Hierarchical clustering analysis
 ICSI: Intracytoplasmic injection
 IUI: Intrauterine Insemination
 IVF: In-vitro fertilization (IVF)
 K-NN: K-nearest neighbors
 LH: Luteinizing Hormone
 LR: Logistic regression
 MDA: Mean decrease of accuracy
 MLP: Multi-layer perceptron
 N: Negative
 NB: Naïve Bayes
 P: positive
 PCA: Principal component analysis
 RF: Random forest
 RBF: Radial basis function
 SVM: Support vector machines
 SD: Standard deviation
 TN: True negative
 TP: True positive
 TSH: Thyroid-stimulating hormone

Declarations:

Funding:

This study was not funded by any organization.

Competing interests:

The authors declare that there are no conflicts of interest.

Ethics approval and consent to participate:

The full name of the ethics committee who approved this study is IR.ACECR.ROYAN.REC which ROYAN Institute belongs to. The committee's reference number is IR.ACECR.ROYAN.REC.1398.213.

Acknowledgments:

The authors acknowledge the Royan institute staffs, especially the informatics department for their valuable contributions. There is no conflict of interest in this study.

Authors' Contribution

Conceptualization: S Ranjbari, T Khatibi and M Totonchi

Data curation: S Ranjbari, T Khatibi and M Totonchi

Formal analysis: S Ranjbari, T Khatibi and M Totonchi

Funding acquisition: there is no funding.

Investigation: S Ranjbari, T Khatibi and M Totonchi

Methodology: S Ranjbari and T Khatibi

Project administration: T Khatibi and M Totonchi

Software: S Ranjbari and T Khatibi

Supervision: T Khatibi and M Totonchi

Validation: S Ranjbari, T Khatibi, AV Taghi Dizaj, H Sajadi, M Totonchi, F Ghaffari

Visualization: S Ranjbari, T Khatibi and M Totonchi

Writing – original draft: S Ranjbari, T Khatibi and M Totonchi

Writing – review & editing: S Ranjbari, T Khatibi, AV Taghi Dizaj, H Sajadi, M Totonchi, F Ghaffari

Availability of data and materials

Our study is a retrospective study of a 5-year couples' data undergoing IUI. Data is collected from Reproductive Biomedicine Research Center, Royan Institute for 8,360 couples who underwent 11,255 IUI cycles were included.

REFERENCES

1. Medicine, P.C.o.t.A.S.f.R., *Definitions of infertility and recurrent pregnancy loss: a committee opinion*. *Fertil Steril*, 2013. **99**(1): p. 63.
2. Borgh, M. and C. Wyns, *Fertility and infertility: Definition and epidemiology*. *Clinical Biochemistry*, 2018. **62**: p. 2-10.
3. Milewska, A.J., et al., *Prediction of infertility treatment outcomes using classification trees*. *Studies in Logic, Grammar Rhetoric*, 2016. **47**(1): p. 7-19.
4. Blank, C., et al., *Prediction of implantation after blastocyst transfer in in vitro fertilization: a machine-learning perspective*. *Fertil Steril*, 2019. **111**(2): p. 318-326.
5. Patil, A.S., *A Review of Soft Computing Used in Assisted Reproductive Techniques (ART)*. *International Journal of Engineering Trends and Applications (IJETA)*, 2015. **2**(3): p. 88-93.
6. Bahadur, G., et al., *First line fertility treatment strategies regarding IUI and IVF require clinical evidence*. *Human Reproduction*, 2016. **31**(6): p. 1141-1146.
7. Ombelet, W., P. Puttemans, and E. Bosmans, *Intrauterine insemination: a first-step procedure in the algorithm of male subfertility treatment*. *Human Reproduction*, 1995. **10**(suppl_1): p. 90-102.
8. Deo, R.C., *Machine learning in medicine*. *Circulation*, 2015. **132**(20): p. 1920-1930.

9. Milewska, A.J., et al., *Analyzing Outcomes of Intrauterine Insemination Treatment by Application of Cluster Analysis or Kohonen Neural Networks*. *Studies in Logic, Grammar Rhetoric*, 2013. **35**(1): p. 7-25.
10. Kooptiwoot, S. and M.A. Salam, *IUI mining: human expert guidance of information theoretic network approach*. *Soft Computing*, 2006. **10**(4): p. 369-373.
11. Ghaffari, F., et al., *Evaluating the effective factors in pregnancy after intrauterine insemination: a retrospective study*. *International journal of fertility and sterility*, 2015. **9**(3): p. 300.
12. Steures, P., et al., *Prediction of an ongoing pregnancy after intrauterine insemination*. *Fertil Steril*, 2004. **82**(1): p. 45-51.
13. Goldman, R.H., et al., *Patient-specific predictions of outcome after gonadotropin ovulation induction/intrauterine insemination*. *Fertil Steril*, 2014. **101**(6): p. 1649-1655. e2.
14. Marshburn, P.B., et al., *Spermatozoal characteristics from fresh and frozen donor semen and their correlation with fertility outcome after intrauterine insemination*. *Fertil Steril*, 1992. **58**(1): p. 179-186.
15. Moro, F., et al., *Anti-Müllerian hormone concentrations and antral follicle counts for the prediction of pregnancy outcomes after intrauterine insemination*. *International Journal of Gynecology and Obstetrics*, 2016. **133**(1): p. 64-68.
16. Lemmens, L., et al., *Predictive value of sperm morphology and progressively motile sperm count for pregnancy outcomes in intrauterine insemination*. *Fertil Steril*, 2016. **105**(6): p. 1462-1468.
17. Arslan, M., et al., *Predictive value of the hemizona assay for pregnancy outcome in patients undergoing controlled ovarian hyperstimulation with intrauterine insemination*. *Fertil Steril*, 2006. **85**(6): p. 1697-1707.
18. Florio, P., et al., *Evaluation of endometrial activin A secretion for prediction of pregnancy after intrauterine insemination*. *Fertil Steril*, 2010. **93**(7): p. 2316-2320.
19. Shah, S. and A. Kusiak, *Cancer gene search with data-mining and genetic algorithms*. *Computers in Biology and Medicine*, 2007. **37**(2): p. 251-261.
20. Kaya, A., *Cascaded classifiers and stacking methods for classification of pulmonary nodule characteristics*. *Computer Methods and Programs in Biomedicine*, 2018. **166**: p. 77-89.
21. Wang, S.Q., J. Yang, and K.C. Chou, *Using stacked generalization to predict membrane protein types based on pseudo-amino acid composition*. *Journal of theoretical biology*, 2006. **242**(4): p. 941-946.
22. Zhang, S., C. Zhang, and Q. Yang, *Data preparation for data mining*. *Applied artificial intelligence*, 2003. **17**(5-6): p. 375-381.

23. Han, J., J. Pei, and M. Kamber, *Data mining: concepts and techniques*. 2011: Elsevier.
24. Breiman, L., *Random Forests*. Machine Learning, 2001. **45**: p. 5-32.
25. Diykh, M., Y. Li, and S. Abdulla, *EEG Sleep Stages Identification Based on Weighted Undirected Complex Networks*. Computer Methods and Programs in Biomedicine, 2020. **184**: p. 105116.
26. Bavelas, A., *A mathematical model for group structure, human organization*. Appl. Anthropol., 1948. **7**(3): p. 16-30.
27. Wolpert, D.H., *Stacked generalization*. Neural networks, 1992. **5**(2): p. 241-259.
28. Sperandei, S., *Understanding logistic regression analysis*. Biochem med, 2014. **24**(1): p. 12-18.
29. Cortes, C. and V. Vapnik, *Support-vector network*. Machine Learning, 1995. **20**: p. 1-25.
30. Quinlan, J.R., *Induction of Decision Trees*. Machine Learning, 1986. **1**: p. 81-106.
31. Zhu, J., et al., *Multi-class AdaBoost*. Statistics and its interfere, 2009. **2**: p. 349-360.
32. Ke, G., et al. *Lightgbm: A highly efficient gradient boosting decision tree*. in *Advances in Neural Information Processing Systems*. 2017.
33. Campana, A., et al., *Intrauterine insemination: evaluation of the results according to the woman's age, sperm quality, total sperm count per insemination and life table analysis*. Human Reproduction, 1996. **11**(4): p. 732-736.
34. Kuriya, A., C. Agbo, and M.H. Dahan, *Do pregnancy rates differ with intra-uterine insemination when different combinations of semen analysis parameters are abnormal?* Journal of the Turkish German Gynecological Association, 2018. **19**(2): p. 57.
35. Zhang, E., et al., *Effect of sperm count on success of intrauterine insemination in couples diagnosed with male factor infertility*. Materia socio-medica, 2014. **26**(5): p. 321.
36. Ombelet, W., et al., *Semen quality and intrauterine insemination*. Reproductive BioMedicine Online, 2003. **7**(4): p. 485-492.
37. Dickey, R.P., et al., *Comparison of the sperm quality necessary for successful intrauterine insemination with World Health Organization threshold values for normal sperm*. Fertil Steril, 1999. **71**(4): p. 684-689.
38. Thijssen, A., et al., *Predictive factors influencing pregnancy rates after intrauterine insemination with frozen donor semen: a prospective cohort study*. Reproductive biomedicine online, 2017. **34**(6): p. 590-597.

39. Merviel, P., et al., *Predictive factors for pregnancy after intrauterine insemination (IUI): An analysis of 1038 cycles and a review of the literature*. Fertility and Sterility, 2010. **93**(1): p. 79-88.
40. Erdem, A., et al., *Factors affecting live birth rate in intrauterine insemination cycles with recombinant gonadotrophin stimulation*. Reproductive biomedicine online, 2008. **17**(2): p. 199-206.
41. Kamath, M.S., et al., *Predictive factors for pregnancy after intrauterine insemination: A prospective study of factors affecting outcome*. Human reproductive sciences, 2010. **3**(3): p. 129.
42. Licht, R.S., L. Handel, and M. Sigman, *Site of semen collection and its effect on semen analysis parameters*. Fertil Steril, 2008. **89**(2): p. 395-397.
43. Francavilla, F., et al., *Effect of sperm morphology and motile sperm count on outcome of intrauterine insemination in oligozoospermia and/or asthenozoospermia*. Fertil Steril, 1990. **53**(5): p. 892-897.
44. Luco, S.M., et al., *The evaluation of pre and post processing semen analysis parameters at the time of intrauterine insemination in couples diagnosed with male factor infertility and pregnancy rates based on stimulation agent. A retrospective cohort study*. European Journal of Obstetrics Gynecology: Reproductive Biology Endocrinology, 2014. **179**: p. 159-162.
45. Blank, C., et al., *Prediction of implantation after blastocyst transfer in in vitro fertilization: a machine-learning perspective*. 2019. **111**(2): p. 318-326.
46. Nuojua-Huttunen, S., et al., *Intrauterine insemination treatment in subfertility: an analysis of factors affecting outcome*. Human Reproduction, 1999. **14**(3): p. 698-703.
47. Liu, W., et al., *Comparing the pregnancy rates of one versus two intrauterine inseminations (IUIs) in male factor and idiopathic infertility*. Journal of assisted reproduction genetics, 2006. **23**(2): p. 75-79.

Figure legend:

Figure 1- the main steps of the proposed method (CNFE-SE) for feature scoring and classifying the patients to predict IUI outcome

Figure 2 –The ratio of positive and negative clinical pregnancy per treatment cycle

Figure 3- One complex network extracted from only 100 data records treated by IUI method

Figure 4- two complex networks drawn from the positive training data samples by (a) threshold of $0.7 * \text{average of the distance Matrix}$, (b) threshold of $0.5 * \text{average of the distance matrix}$

Figure 5- three complex networks extracted from the samples of (a) both classes, (b) negative class, and (c) positive class with the same threshold

Figure 6- (a) input datasets, and (b) the architecture of the proposed Stacked Ensemble classifier

Figure 7- ROC curve for CNFE-SE trained with all features

Figure 8 -Precision-recall curves for CNFE-SE

Figure 9-Overview of top features ranked based on CNFE-SE

Figure 10- Partial dependency plots of nine features among the important features which the blue curves indicate locally weighted smoothing. It shows pregnancy variation obtaining from CNFE-SE (y-axis) as a function of a feature (x-axis) in IUI.

Figure 11- Schematic of our proposed medical decision support system for IUI outcome prediction