Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

Sima Ranjbari
Tarbiat Modares University
Toktam Khatibi (toktamk@gmail.com)
Tarbiat Modares University https://orcid.org/0000-0001-5824-9798
Ahmad Vosough Taghi Dizaj
Royan Institute
Hesamoddin Sajadi
Royan Institute
Mehdi Totonchi
Royan Institute
Firouzeh Ghaffari
Royan Institute

Research article

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

Posted Date: December 11th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-24259/v3

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Version of Record: A version of this preprint was published on January 2nd, 2021. See the published version at https://doi.org/10.1186/s12911-020-01362-0.

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

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 Ghaffari[5]

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 which the assisted reproductive technology (ART) practitioners are dealing with. 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. Many previous studies have been focused on predicting the in vitro fertilization (IVF) and intracytoplasmic sperm injection (ICSI) outcome using 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 propose an automatic classification and feature scoring method to predict intrauterine insemination (IUI) outcome and ranking the most significant features.
Methods: For this purpose, a novel 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 describing 11255 IUI treatment cycles for 8,360 couples. Our dataset includes the couples' demographic characteristics, historical data about the patients' diseases, the clinical diagnosis, the treatment plans and the prescribed drugs during the cycles, semen quality, laboratory tests and the clinical pregnancy outcome
Results: Experimental results show that the proposed method outperforms the compared methods with Area under receiver operating characteristics curve (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 body mass index (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 of the world's population specifically people living in developing countries are suffering from infertility [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) and 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 which the ART practitioners are dealing with. 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 the hidden patterns and extract 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].

To the best of our knowledge, many previous studies have been focused on predicting the IVF and ICSI outcome using machine learning methods as summarized in Table 1.

Table 1- summarizing the previous studies of predicting ART outcome

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), Random Forest (RF), and Artificial Neural Networks (ANN) such as Multi-Layered Perceptron (MLP) and Radial Basis Function (RBF) 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 the considered features. Small dataset increases the risk of overfitting the trained models. Overfitting occurs when a model has good predictive ability for training dataset but shows poor performance for test dataset. Models with high overfitting property has lower generalization ability.

In this study, a dataset including the features of 11255 IUI treatment cycles for 8360 couples is considered for IUI outcome prediction. Our dataset includes the couples' demographic characteristics, historical data about the patients' diseases, the clinical diagnosis, the treatment plans and the prescribed drugs during the cycles, semen quality, laboratory tests and the clinical pregnancy outcome. Considering the large number of couples and their corresponding IUI treatment cycles 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 ISCI. 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]. Many previous studies have suffered from the lack of statistical power due to their small dataset [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 methods 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 the features describing the couples and their corresponding IUI treatment cycles. For this purpose, a novel 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 describing 11255 IUI treatment cycles for 8,360 couples.
The main novelty of this study lies in three 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 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

Our research is approved by the Institutional Review Board of the Royan Institute Research Center and the Royan Ethics Committee consistent with Helsinki Declaration with the approval ID of IR.ACECR.ROYAN.REC.1398.213. Anonymity and confidentiality of data were respected.
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.
In this retrospective study, a completed episode is defined as a sequence of treatment cycles resulting in positive clinical pregnancy or when the treatment with IUI is stopped. The inclusion criteria for the couples to be treated under IUI cycles were male factor, ovulatory disorders such as PCOS, hypothalamic amenorrhea, diminished ovarian reserve, combined causes, and unexplained subfertility. The couples' duration of infertility was at least 1 year. Male infertility was defined as the semen quality parameters lower than the standards determined by WHO including sperm concentration lower than 15 million/ejaculate, semen volume lower than 1.5 mL, and total motility lower than 40% [22]. The male partners with donor sperms, Varicocele, and semen samples with total motile sperm count lower than 1×10^6

were excluded from being candidates for IUI treatment. Additionally, patients with anatomical and metabolic abnormalities, severe endometriosis and/or systemic diseases were excluded from our study.

11,255 IUI cycles related to 8,360 couples are considered in which the women age ranges from 16 to 47 with the average age 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 ± Standard Deviation) IUI cycles which ranges from 1 to 7 cycles.

The features describe the couples' demographic characteristics, historical data about their diseases, the clinical diagnosis, the treatment plans and the prescribed drugs to the couples, male semen quality, laboratory tests and the clinical pregnancy outcome. The considered demographic features include age, body mass index (BMI), education level, consanguinity with spouse and some other features. The information about the history of the patients' subfertility consists of the duration and type of infertility, length of marriage and so on.

The types of feature values are numerical, binary, nominal and binominal types for 86, 152, 51 and 7 features, respectively. More details about the features is shown in Appendix A.

In the collected dataset, the majority of couples (almost 72%) have been treated for one cycle, 22% of couples have underwent two cycles, 5% of couples have been treated for three cycles, and less than 1% have been treated more than three cycles. The maximum number of cycles for treating a couple is seven. 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 completing the 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 test datasets with no overlapping among these two subsets. The models are trained on the training dataset and finally are evaluated by applying them to the test 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 is considered as the test dataset and all (K-1) remaining subsets make the training one. The model is trained K times on K training datasets and applied to the corresponding test datasets to evaluate the performances of the trained models.

Before sampling from data, the features having missing value rate higher than 20% are removed from the study. Moreover, the patient records with high missing value rate (higher than 20%) are excluded from the study and then, 5-fold C.V. is used for sampling from the collected dataset, in this study.

At first, dataset is partitioned into non-overlapping subsets D_1, D_2, …, D_K based on K-fold Cross Validation strategy. Then, the models are trained on K training datasets composed of all D_1, …, D_K subsets excluding D_i for $1 \leq i \leq K$. Therefore, the i^{th} training dataset consists of all D_1, …, D_K but D_i and the it^h test dataset is D_i. The ith training dataset is balanced using over-sampling strategy.

Moreover, a strategy for classification structural risk assessment is used named as A-Test which will be described in the evaluation and validation subsection with more details. The number of instances of positive and negative outcomes in each folder of 5-fold is 324-325 and 1926-1927, respectively. therefore, the imbalance ratio of the training set in each of 5-folds is about 0.168.

2.3 Data preprocessing

Preprocessing of data is one of the most essential steps in the knowledge discovery tasks. A previous study have stated that 80% of total time in data mining projects is allocated for data preparation and preprocessing step [23].

In the first step, the initial collected dataset includes almost 86,000 data records describing the partners and about 1,000 features. The data records describing one couple per IUI treatment cycle are aggregated to make our dataset. Thus, the aggregated dataset includes 11,255 data records and 296 features describing a couple during an IUI treatment cycle.

The nominal features are converted to dummy binary variables. If a nominal features has m different levels or values, it will be converted to (m-1) dummy binary variables. Therefore, instead of considering a nominal feature in the classification and feature ranking, its corresponding dummy binary variables are considered in the mentioned tasks.

The missing values for numeric and categorical features are imputed based on the average and the most frequent values, respectively [24]. All numerical and ordinal features are normalized using min-max normalization method and the nominal features are converted into dummy binary variables.

Outlier detection is performed in this study based on isolation forest method which has been proposed by Liu et al. [25] as an appropriate outlier detection method for high dimensional data. The hyperparameters of Isolation Forest including the number of estimators, maximum number of the samples, contamination coefficient, maximum number of the features, bootstrapping or not, and the number of jobs are tuned using grid search method. For evaluating the performance of Isolation Forest, its results are compared to other outlier detection methods such as One-class SVM with kernel of Radial Basis Function (RBF), boxplot analysis and expert's opinions. Three outliers are identified by this method and excluded from the study.

2.4 Filtering irrelevant features

Since the aggregated dataset consists of many 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 [26]. 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 after removing a feature. Therefore, the higher values of MDA indicate the higher predictive ability of the corresponding features.

2.6 Constructing complex networks of patients

For modeling nonlinear data, complex networks are effective method [27]. 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 their corresponding edges e (v_i, v_j) of E.

Three complex networks are constructed from the training datasets and one data record which should be classified independent from it belongs to training or test dataset. The first one is comprised of all the training data records and one data record which should be classified as its nodes and is called CN1. The second and the third complex networks consist of one data record which should be classified and all training data records excluding the negative and positive classes and named as CN2 and CN3, respectively. If the considered data record belongs to training dataset, its class label is excluded from its corresponding complex networks.

In other words, the nodes of CN1, CN2 and CN3 are one data record which should be classified and all the training data records, positive labeled and negative labeled training data records, respectively. Therefore, for each data record, three complex networks are constructed.

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):

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

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):

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

2.7 Feature engineering based on the complex networks

In this section, three complex networks per data record are constructed including the considered data record, 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 hypothesis.

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 [28] and is calculated as Eq. (4). 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.

$$Betweenness(v_i) = \sum_{j<k} \frac{number\ of\ the\ shortest\ paths\ between\ v_j\ and\ v_k\ passing\ v_i}{number\ of\ the\ shortest\ paths\ between\ v_j\ and\ v_k} \quad (4)$$

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. (5):

$$clusteringCoefficient(v_i) = \frac{\text{number of triangles connected to } v_i}{\text{number of triples centered around } v_i} \quad (5)$$

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 by Figure 4, reducing the threshold for keeping the edges in the complex network even with a small value 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 [29]. 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 mode is written as Eq. (6):

$$f_n(x,V) = V_{n,k}\left(f_{n-1}(x,V_{n-1,1}), f_{n-1}(x,V_{n-1,2}), \ , \ f_{n-1}(x,V_{n-1,D_{n-1}})\right) \qquad (6)$$

Where n indicates the n^{th} layer of the stacked ensemble, 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 n^{th} hidden layer and finally, k is the k^{th} neuron in the n^{th} layer.

Some previous studies have illustrated that the stacked models can improve the performance of the classification [20, 21, 30]. 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 are 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.

Several different classifiers are trained and verified. The classifiers for using in the ensemble layers of our proposed stacked ensemble classifier are chosen among different trained classifiers with different values of hyperparameters based on their accuracy and diversity on the validation dataset. A previous study has proposed a method to choose classifiers for ensemble learning based on accuracy and diversity which is used in this study for the same purpose (Yang, 2011). The pairwise diversity of the classifiers are calculated using Q statistic.

Logistic regression (LR) [31], support vector machines (SVM) [32], decision tree (DT) [33], random forest (RF) [26], Adaboost [34] and LightGradient Boosting Machine (LightGBM) [35] are the base classifiers chosen based on their accuracy and diversity in both ensemble layers.

LR, SVM with linear kernel and DT are appropriate classifiers for classifying linearly separable data. SVM with non-linear kernels, 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.

The hyperparameters of the classifiers are tuned based on grid search method and the best values for hyperparameters leading to the highest accuracy for validation dataset are considered for each 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 during the base classifiers' training in both layers.

Mathematical calculation is performed in this study to show the performance improvement obtained by stacked ensemble compared to traditional one-layer ensemble and the individual classifiers.

Without loss of generality, it is assumed that each base classifier in the first layer of stacked ensemble has the error rate of $\varepsilon < 0.50$. If the aggregation of the base classifiers is performed with bagging strategy which is the simplest aggregation method and uses majority voting, the error rate of the first ensemble layer (ε_{L1}) can be calculated as Eq. (7):

$$\varepsilon_{L1} = \sum_{i=(\frac{M}{2}+1)}^{M} \binom{M}{i} \varepsilon^i (1-\varepsilon)^{M-i} \quad (7)$$

Where M is the number of the base independent classifiers in the first ensemble layer. For misclassifying a data record using bagging strategy as the aggregation method, more than half of the base classifiers should misclassify the record. If it is assumed that i is the number of the base classifiers which misclassify the data record, i should be more than M/2 for misclassifying it with the first ensemble layer. For example, if M is 25, at least 13 base classifiers should misclassify data for erroneous classifying data in ensemble of these base classifiers. Now, if ε is 0.35 for each of 25 base classifiers, ε_{L1} will be 0.04. It shows the first layer of ensemble or traditional ensemble can improve the error rate of the single independent classifiers significantly.

Now, it is assumed that we have one more ensemble layer such as a two-layer stacked ensemble. Bagging strategy uses simple majority voting for classifying data as Eq. (8):

$$classLabel_{ensemble}(r_j) = \begin{cases} Positive & if \sum_{i=1}^{M} \delta\big(classLabel_i(r_j) == Positive\big) > \frac{M}{2} \\ Negative & otherwise \end{cases} \quad (8)$$

Where r_j indicates the j^{th} data record and i denotes the i^{th} base classifier. As shown in Eq. (8), a simple decision tree or SVM with linear kernel can provide rules or find hyperplanes to classify data according to Eq. (8). Therefore, it can be shown that the performance of each base classifier in the second layer will not be worse than the simple bagging aggregation strategy used in the first ensemble layer.

This conclusion is true because each base classifier will try to find the hyperplane or rules to discriminate the training samples of two classes. But, bagging strategy uses simple majority voting. Furthermore, the input features (the first meta feature set as shown by Fig.6) for the base classifiers of the second ensemble layer are the same as the input features fed to the bagging strategy in the first ensemble layer. These input features are the output class labels generated by the base classifiers in the first layer. Therefore, the error rate of each base classifier in the second ensemble layer would be at most ε_{L1}.

The aggregation rule in the first ensemble layer is majority voting in the bagging strategy. The base classifiers try to separate the instances of different classes using linear or non-linear hyperplanes or rules. The input dataset for majority voting in the first ensemble layer is the first meta feature set. Therefore, the input of the majority voting rule and the base classifiers of the second ensemble layer is the same. The majority voting rule can be stated as Eq. (9) for the first meta feature set with M columns:

$$label_{MV}(r_j) = \begin{cases} Positive & if \ \sum_{i=1}^{M} classLabel_i(r_j) > 0 \\ Negative & otherwise \end{cases} \quad (9)$$

Where MV is the majority voting strategy. Majority voting strategy is similar to using a hyperplane considering the equal coefficients for all of its input features as the separator of two classes.

The base classifiers try to find a best hyperplane for discriminating the instances of two classes. Therefore, their fitted hyperplane will not be worse than the hyperplane used with majority voting strategy. Thus, their performance will be more than or equal to the performance of the majority voting in the first ensemble layer. According to the Eq. (7), it is shown that the performance of the majority voting will be much better than the performance of the single classifiers in the first ensemble layer. Therefore, the performance of the single classifiers in the second ensemble layer will be better than the performance of the single classifiers in the first ensemble layer.

According to Eq. (7), if the bagging strategy is used for the second ensemble layer, the error rate of the second ensemble layer in the stacked ensemble would be ε_{L2} which can be calculated as Eq. (9):

$$\varepsilon_{L2} = \sum_{j=(\frac{M_2}{2}+1)}^{M2} \binom{M2}{j} \varepsilon_{b2}^{j}(1-\varepsilon_{b2})^{M2-j} \leq \sum_{j=(\frac{M_2}{2}+1)}^{M2} \binom{M2}{j} \varepsilon_{L1}^{j}(1-\varepsilon_{L1})^{M2-j} \quad (9)$$

Where M_2 is the number of the base classifier in the second ensemble layer of the stacked ensemble and ε_{b2} is the error rates of the base classifiers in the second ensemble layer. As mentioned in the previous paragraph, the error rate of each base classifier in the second layer would be at most ε_{L1}. Therefore, ε_{b2} will be not more than ε_{L1}.

According to Eq. (7) and Eq.(9), the relationship among ε, ε_{L1} and ε_{L2} can be shown in Eq. (10):

$$\varepsilon_{L2} \ll \varepsilon_{L1} \ll \varepsilon \qquad (10)$$

A previous study have demonstrated that adding more layers to stack ensemble can improve the classification performance in terms of accuracy and AUC [1].
Based on the obtained results, it can be shown that adding more layers to stacked ensemble can improve its performance. Although, adding more layers has higher burden of time complexity and memory usage, too. There are a few studies considering the effect of the ensemble size or cardinality (the number of the base classifiers in the ensemble classifier) on the performance of the ensemble method [1, 2]. The previous studies have shown the ensemble size depends on the diversity of the base classifiers included in the ensemble and its aggregation rule [1, 2]. In addition, a previous study has examined different ensemble sizes including 10,20, 50 and 100 classifiers for bioinformatics applications [3]. They have shown that the best ensemble size has been 50 but the ensemble size of 10 is sufficient to achieve to highly reasonable performance [3].

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, the performance measures for classification problems are used in this study including Accuracy, Sensitivity, Specificity and F-Score as shown in Eq. (11) -(14):

$$Accuracy = \frac{TP + TN}{N} \quad (11)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (12)$$

$$Specificity = \frac{TN}{TN + FP} \quad (13)$$

$$F-Score = 2 \times \frac{Sensitivity \times Specificity}{Sensitivity + Specificity} \quad (14)$$

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 (ROC) 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.

A novel method named as A-Test has been proposed in a previous study to calculate the structural risk of a classifier model as its instability with the new test data [36]. A-test calculates the misclassification error percentage $\Gamma_{\zeta,K}$ for different K values using the balanced K-fold validation. In this study, the values of $\Gamma_{\zeta,K}$ will be reported for different classifiers and different feature sets. $\Gamma_{\zeta,K}$ is calculated as Eq. (15):

$$\Gamma_{\zeta.K} = \frac{100}{N}\left(\sum_{i=1}^{N}\delta((predictedLabel == Negative).(realLabel == Positive)) + \sum_{i=1}^{N}\delta((predictedLabel == Positive).(realLabel == Negative))\right)$$
$$K = 2.\cdots.K_{max} \quad (15)$$

Where K_{max} cannot be more than the size of the minority class. For estimating the structural risk of a classifier method, the average of the values of $\Gamma_{\zeta,K}$ is considered as Eq. (16):

$$\hat{\Gamma_{\zeta}} = \frac{\sum_{K=2}^{K_{max}} \Gamma_{\zeta.K}}{K_{max} - 1} \quad (16)$$

Where $\Gamma\hat{}\zeta$ ranges from 0 to 100% which higher values show higher risk of classification and lower values show the higher capacity and generalization ability of the model. Therefore, the lower values of $\Gamma\hat{}\zeta$ are more desired.

3 Experimental results

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

Ranking the significance of features

Error! Reference source not found.Figure 7 represents top-20 important features with highest MDA score for IUI outcome prediction based on 50 repetitions of CNFE-SE 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 are high-ranked predictors of 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 7Error! Reference source not found., have the highest repetitions as the first informative features. Generally, the variables related to the men's semen analysis parameters are high-ranked features in this study.

Figure 7-Overview of top features ranked based on CNFE-SE

The Pearson correlation coefficients are calculated among the top-20 important features, and Figure 8 depicts the heat map of the correlation coefficients.

Figure 8- the pairwise correlation analysis of 20 most important features

As shown by Figure 8, the male semen parameters are positively correlated to each other, the more sperm concentration, the more total sperm count, and the more total motile sperm count. Also, couples' duration of infertility and duration of marriage are positively correlated.
Figure 9 shows the exact values of MDA score for top-20 features in this study.

Figure 9- MDA values of top-20 features in this study

In addition, Table 3 lists MDA values of top-20 features.

Table 3- MDA values of top-20 features

3.1 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 changes in the clinical pregnancy probabilities in terms of 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 more than0.52. This finding underlines the value of the machine learning approach by determining the complicated association between individual 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 obtained by CNFE-SE (y-axis) as a function of a feature (x-axis) in IUI.

According to the results of the partial dependency plots as shown by Figure 8, 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 [37].

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

Table 4 lists the performance measures for comparing CNFE-SE with other state of the art classifiers.

Table 4- comparing the performance of CNFE-SE with other state of the art classifiers

Two different feature sets are considered as the input variables fed to the classifiers including all 296 features and only the most important features (top-20 features shown in Fig.6). Moreover, CNFE-SE is trained and evaluated twice (one time without doing feature engineering (FE) and another time with performing feature engineering).

The models are executed and trained on different random training samples up to 50 times and the mean ± standard deviation values are depicted in Table 4. 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 when trains on all 296 features. Moreover, CNFE-SE has the superior performance when only 20-top features are fed to it as input variables with AUC of 0.87±0.01, sensitivity of 0.82±0.01, specificity of 0.92±0.01 and accuracy of 0.87±0.01. Our obtained results show that feature engineering and considering only 20-top features improve the performance of CNFE-SE.
Table 5 shows the confusion matrix of CNFE-SE for total dataset.

Table 5- the confusion matrix of CNFE-SE for total dataset

Figure 11 depicts ROC curve for CNFE-SE trained with all features.

Figure 11- ROC curve for CNFE-SE trained with all features

As shown by Figure 11, AUC of CNFE-SE trained on all features is 0.84 ± 0.01. As illustrated by Table 4, the compared single 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 using single classifier is 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 [35].

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 12 indicates the precision-recall curves for CNFE-SE trained using top-20 features.

Figure 12 -Precision-recall curves for CNFE-SE trained using top-20 features

As shown in Figure 12, CNFE-SE predicts both classes with highly reasonable performance.
Moreover, the results of A-test method for structural risk calculation for different combinations of feature sets and classifiers are shown in Table 6.

Table 6- Results of the A-Test: The values of $\Gamma\hat{}\zeta$ and the minimum value of Γ_ζ

Lower values of Γ^ζ and Γζ shows lower risk of the classifier for classifying previously unseen records and the higher capacity and generalization ability of the model. Therefore, the feature set and classifier achieving the lower values of Γ^ζ and Γζ is more desired. As shown by Table 6, CNFE-SE trained using top-20 features has the superior performance based on A-Test results.

4 DISCUSSION

In the current study, among the various features that significantly affect the IUI outcome, the most potential predictors are female BMI and semen quality parameters. Semen data such as sperm count and motility are illustrated as the most prognostic factors in pregnancies, conceived by IUI and their association with IUI outcome have demonstrated in some previous studies [38]. Moreover, some previous studies have confirmed that semen descriptors, after the swim-up procedure have been more important than the ones before sperm washing process [39, 40]. 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 [41, 42]. 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 [43]. Thus, if their corresponding values are more than 20% and less than 15%, respectively, the IUI success rate is higher.

Furthermore, the results of this study indicates that the IUI success rate is almost low when the female BMI is abnormal (BMI is lower than 20 or larger than 30). If female BMI is about 25 as the normal BMI value, the probability of pregnancy increases. This finding is mentioned in the previous studies, too [44].

Previous studies have shown that pregnancy rate could be reduced by increase in the female age [42, 45]. 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 [46].

As shown in 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 [47].

The total dose of gonadotropins is taken into account in this study as an important feature. Moreover, 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 demonstrated as the influential attributes in some previous studies [48-50].

Eventually, the CNFE-SE is trained using the 20 most important features and it yields surprisingly good performances (AUC =0.87, 95% CI 0.86-0.88). 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. For example, they have considered the first cycle information [16, 51]. Our reanalysis of the primary cycle data revealed that the AUC performances of Light-GBM and CNFE-SE are 0.62 ± 0.01 and 0.84 ± 0.01, respectively, which does not change significantly when all the cycles are taken into account. Moreover, as shown in the materials and methods section, increasing the number of cycles augment the clinical pregnancy rate which are in line with the importance of this feature in subsequent IUI outcome [52, 53]. 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 second, third and more cycles, which approximately 74% of the data belongs to the first cycle of IUI treatment.

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 their high missing value rate. 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 low quality 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 13.

Figure 13- 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 experimental results in this study show that 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 advanced balanced sampling strategies.

Determining the optimal ensemble size is a challenging issue, yet. It is suggested that the impact of the ensemble size on the overall performance of stacked ensemble is studied in the future studies on different tasks and different datasets.

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

Ethics approval and consent to participate:

This study is approved by the institutional review board of the ROYAN Institute (IR.ACECR.ROYAN.REC.1398.213). The informed consent requirement for this study was waived because this was a retrospective study with little patients' sensitive or personal information, and all data were anonymized.
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.

Consent for publication:

Not applicable.

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. But, we are not allowed to share the original dataset because of the privacy and security issues.

Competing interests:

The authors declare that there are no conflicts of interest.

Funding:

This study was not funded by any organization.

Authors' Contributions

Conceptualization: SR, TK and MT
Data curation: SR, TK and MT
Formal analysis: SR, TK and MT
Funding acquisition: there is no funding.
Investigation: SR, TK and MT
Methodology: SR and TK
Project administration: TK
Software: SR and TK
Supervision: TK and MT
Validation: SR, TK, AVTD, HS, MT, FG
Visualization: SR, TK and MT
Writing – original draft: SR and TK
Writing – review & editing: SR, TK, AVTD, HS, MT, FG
All authors have read and approved the manuscript.

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.

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. Borght, 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. Tocci, A. and C. Lucchini, *WHO reference values for human semen.* Human reproduction update, 2010. **16**(5): p. 559-559.
23. Zhang, S., C. Zhang, and Q. Yang, *Data preparation for data mining.* Applied artificial intelligence, 2003. **17**(5-6): p. 375-381.
24. Han, J., J. Pei, and M. Kamber, *Data mining: concepts and techniques*. 2011: Elsevier.
25. Liu, F.T., K.M. Ting, and Z.H. Zhou, *Isolation Forest*, in *2008 Eighth IEEE International Conference on Data Mining*. 2008, IEEE. p. 413-422.
26. Breiman, L., *Random Forests.* Machine Learning, 2001. **45**: p. 5-32.
27. 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.
28. Bavelas, A., *A mathematical model for group structure, human organization.* Appl. Anthropol., 1948. **7**(3): p. 16-30.
29. Wolpert, D.H., *Stacked generalization.* Neural networks, 1992. **5**(2): p. 241-259.
30. Güneş, F., R. Wolfinger, and P.Y. Tan. *Stacked ensemble models for improved prediction accuracy*. in *Static Anal. Symp.* 2017.

31. Sperandei, S., *Understanding logistic regression analysis.* Biochem med, 2014. **24**(1): p. 12-18.
32. Cortes, C. and V. Vapnik, *Support-vector network.* Machine Learning, 1995. **20**: p. 1-25.
33. Quinlan, J.R., *Induction of Decision Trees.* Machine Learning, 1986. **1**: p. 81-106.
34. Zhu, J., et al., *Multi-class AdaBoost.* Statistics and its interfere, 2009. **2**: p. 349-360.
35. Ke, G., et al. *Lightgbm: A highly efficient gradient boosting decision tree*. in *Advances in Neural Information Processing Systems*. 2017.
36. Gharehbaghi, A. and M. Linden, *A Deep Machine Learning Method for Classifying Cyclic Time Series of Biological Signals Using Time-Growing Neural Network.* IEEE Transactions on Neural Networks and Learning Systems, 2018. **29**(9): p. 4102-4115.
37. 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.
38. 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.
39. 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.
40. Ombelet, W., et al., *Semen quality and intrauterine insemination.* Reproductive BioMedicine Online, 2003. **7**(4): p. 485-492.
41. 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.
42. Duran, H.E., et al., *Sperm DNA quality predicts intrauterine insemination outcome: a prospective cohort study.* Human Reproduction, 2002. **17**(12): p. 3122-8.
43. Muriel, L., et al., *Value of the sperm chromatin dispersion test in predicting pregnancy outcome in intrauterine insemination: a blind prospective study.* Human Reproduction, 2006. **21**(3): p. 738-744.
44. 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.

45. 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.
46. 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.
47. 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.
48. 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.
49. 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.
50. 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.
51. Blank, C., et al., *Prediction of implantation after blastocyst transfer in in vitro fertilization: a machine-learning perspective.* 2019. **111**(2): p. 318-326.
52. Nuojua-Huttunen, S., et al., *Intrauterine insemination treatment in subfertility: an analysis of factors affecting outcome.* Human Reproduction, 1999. **14**(3): p. 698-703.
53. 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 * average of the distance Matrix, (b) threshold of 0.5 * 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-Overview of top features ranked based on CNFE-SE

Figure 8- 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.

Figures

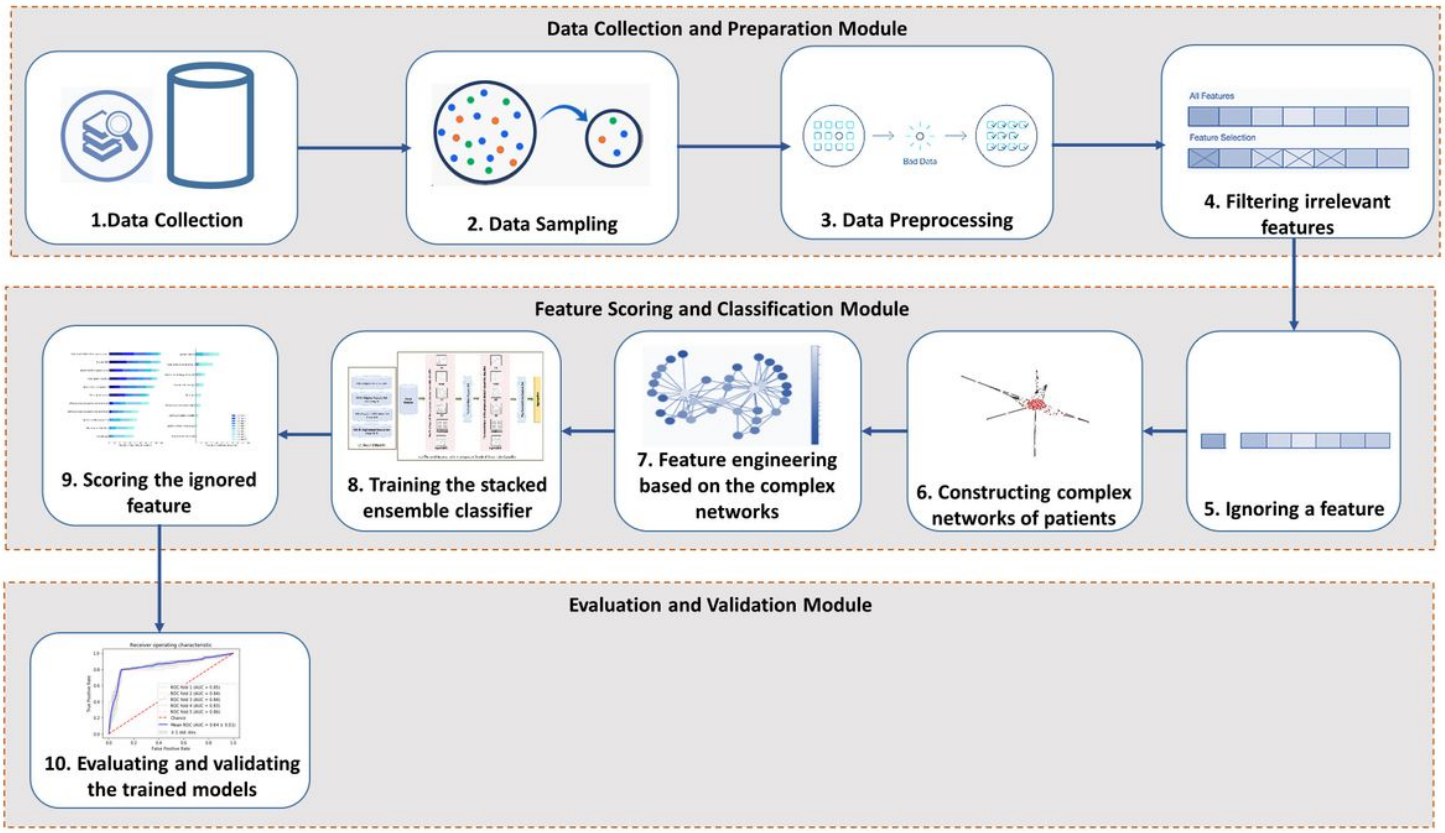

Figure 1

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

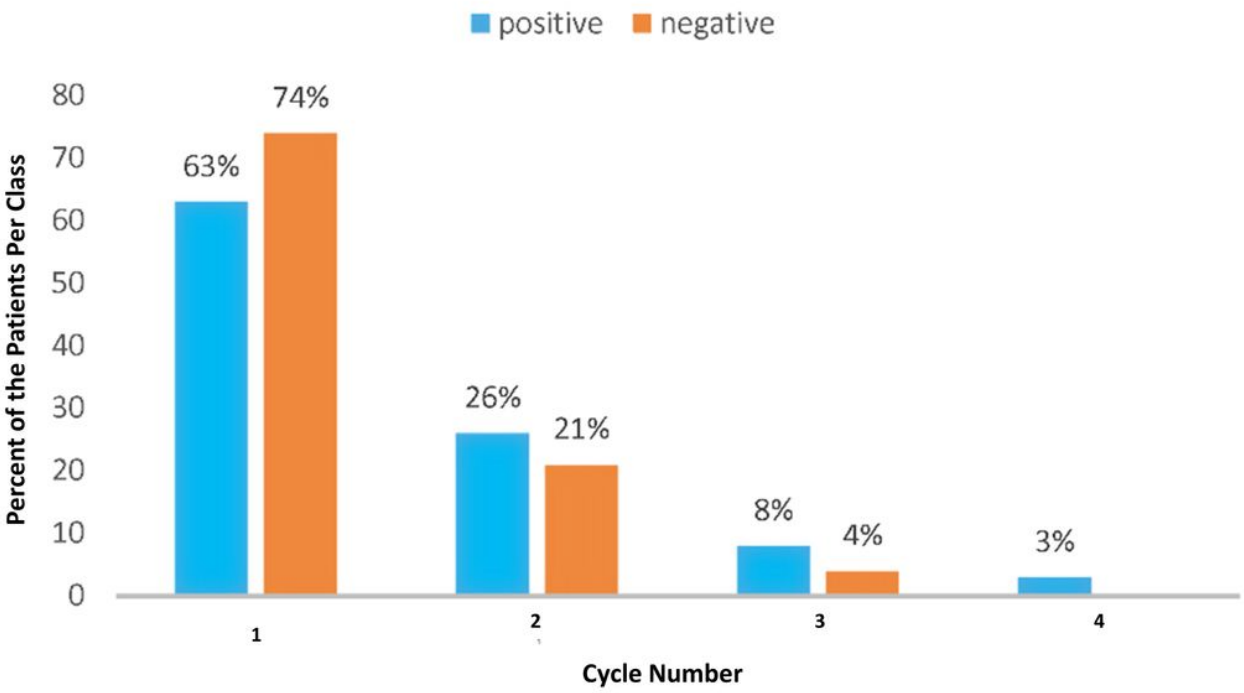

Figure 2

The ratio of positive and negative clinical pregnancy per treatment cycle

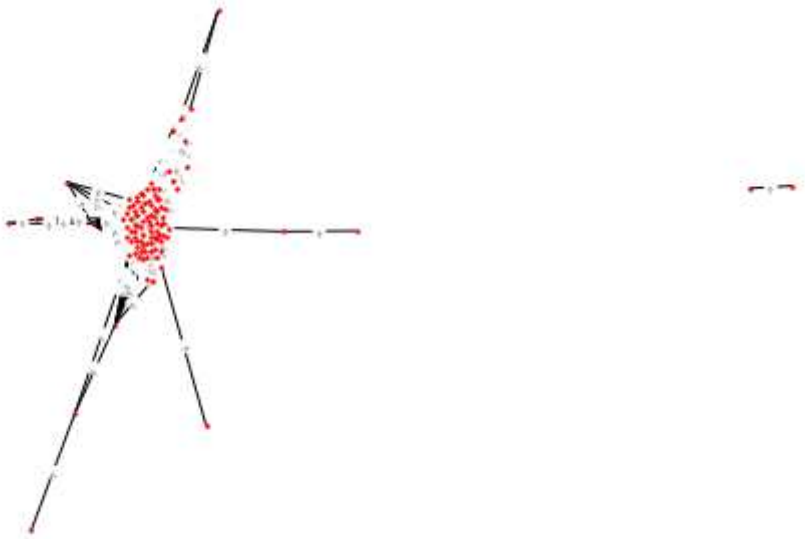

Figure 3

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

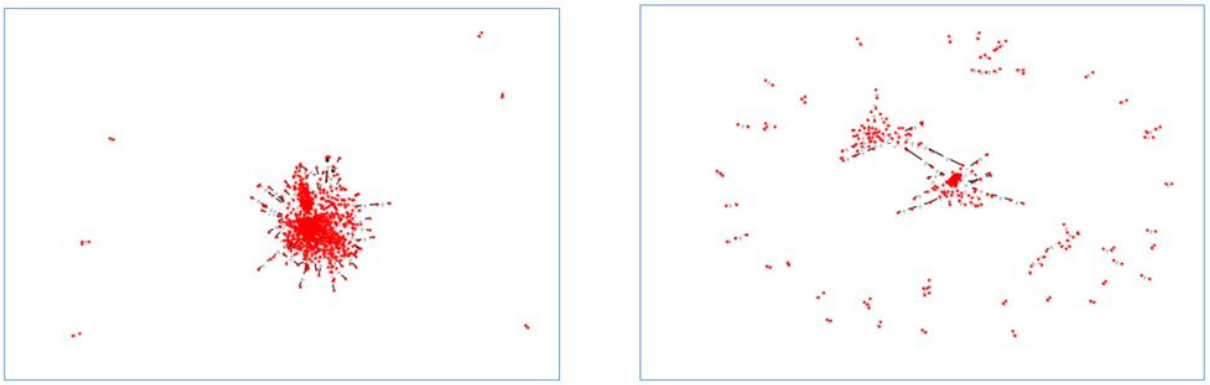

Figure 4

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

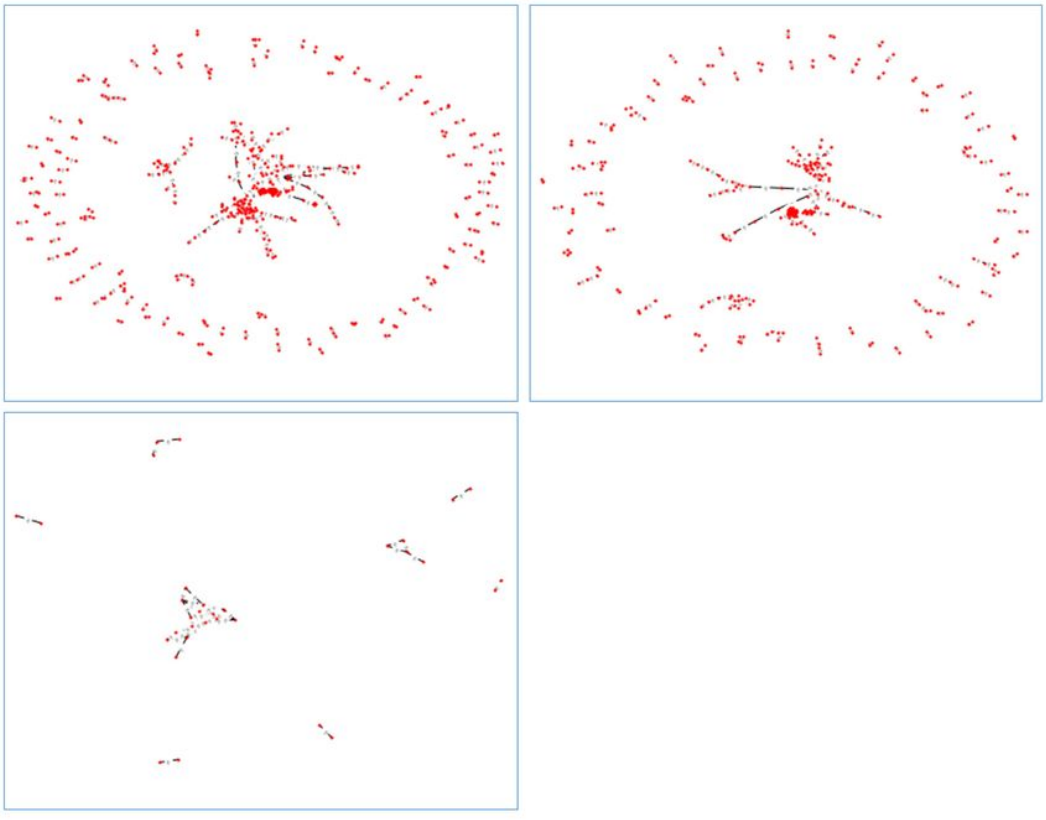

Figure 5

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

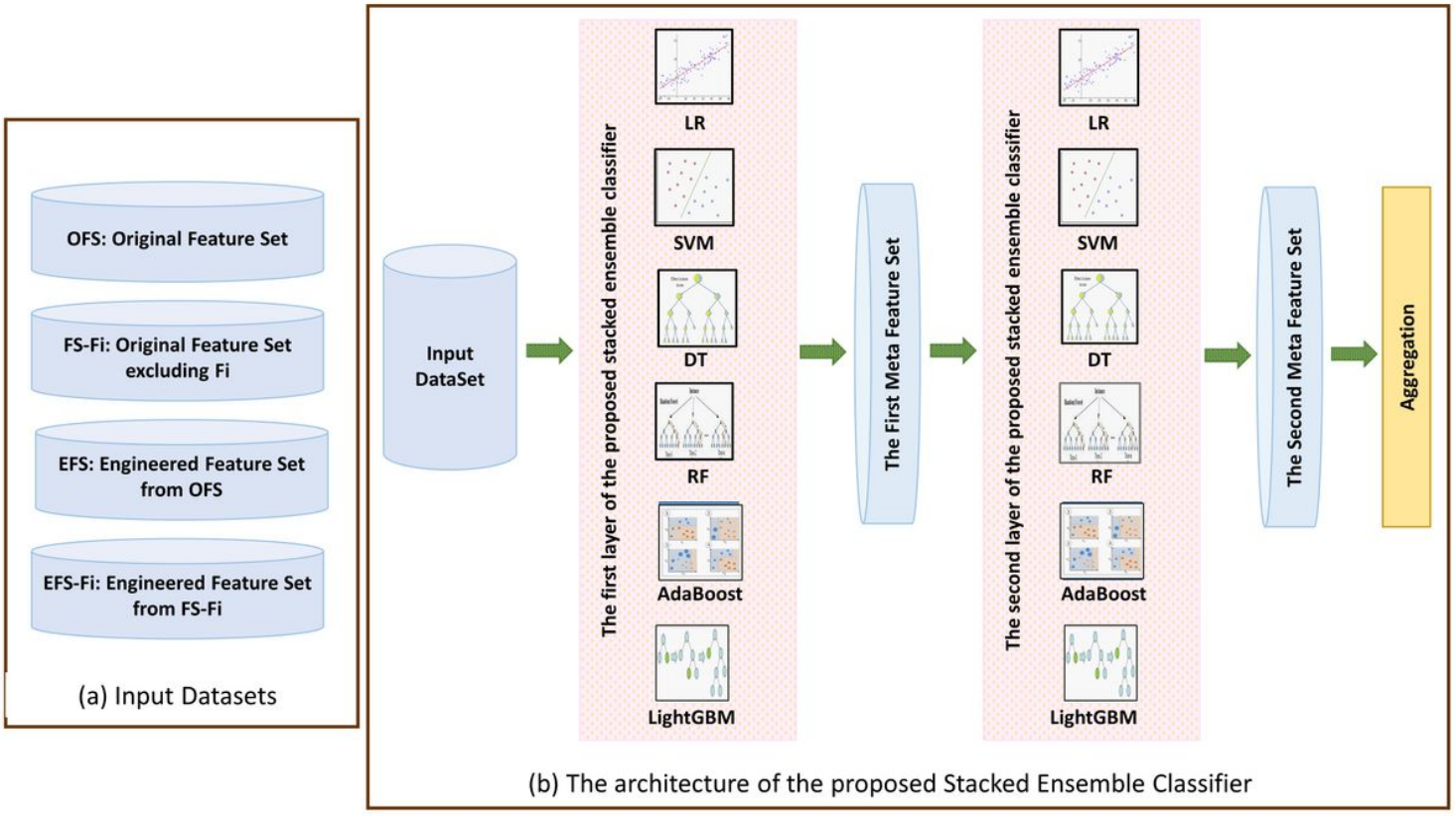

Figure 6

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

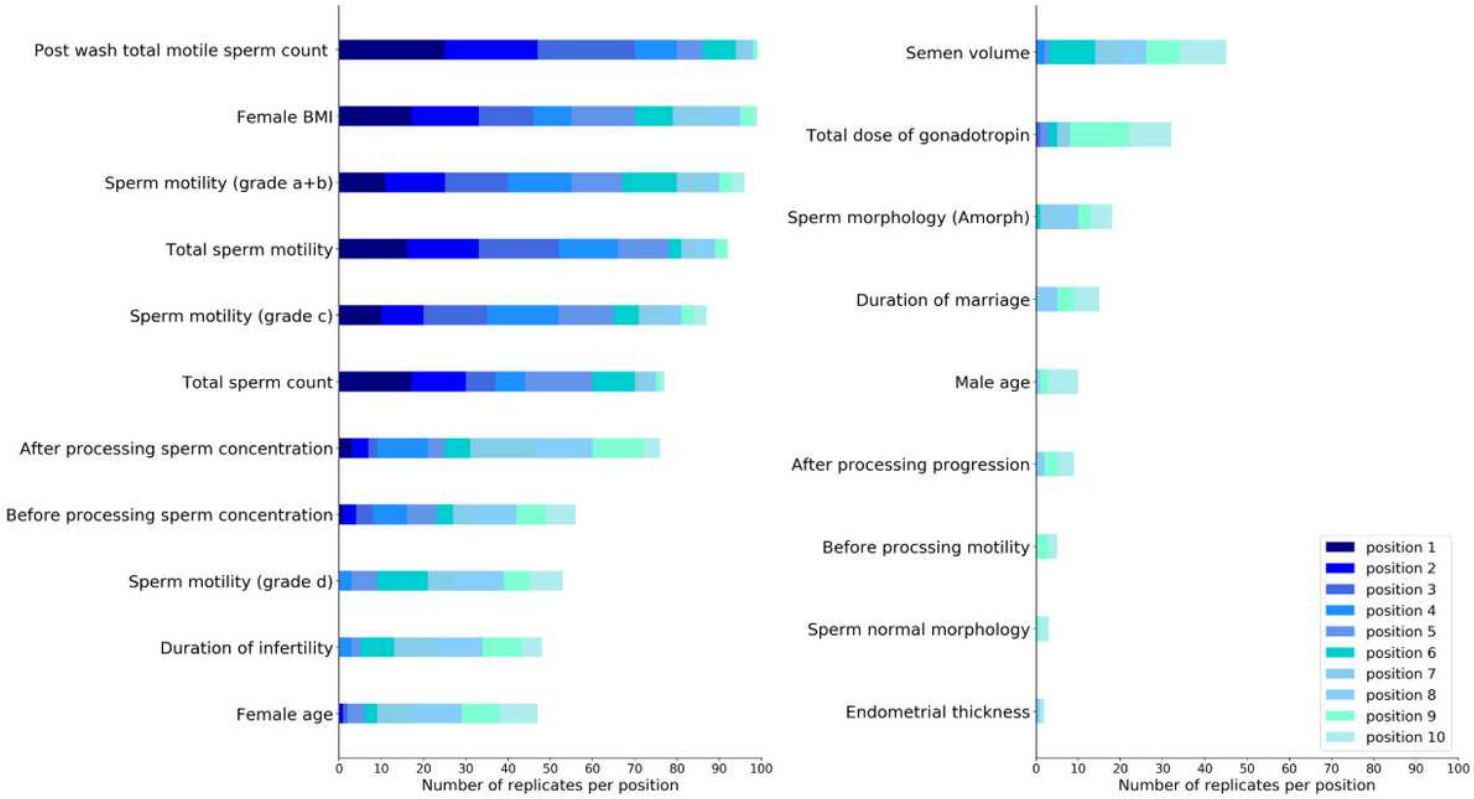

Figure 7

Overview of top features ranked based on CNFE-SE

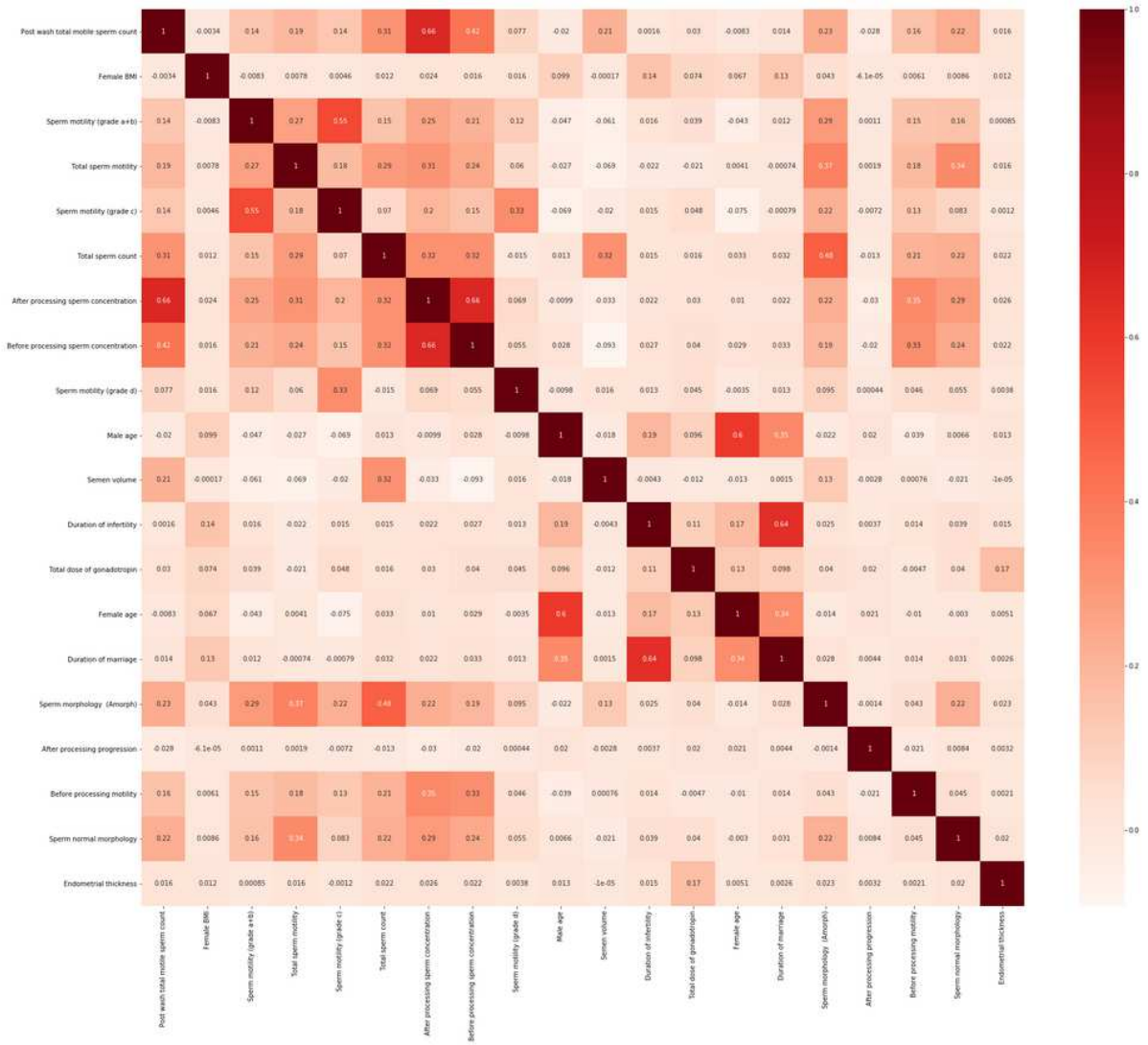

Figure 8

the pairwise correlation analysis of 20 most important features

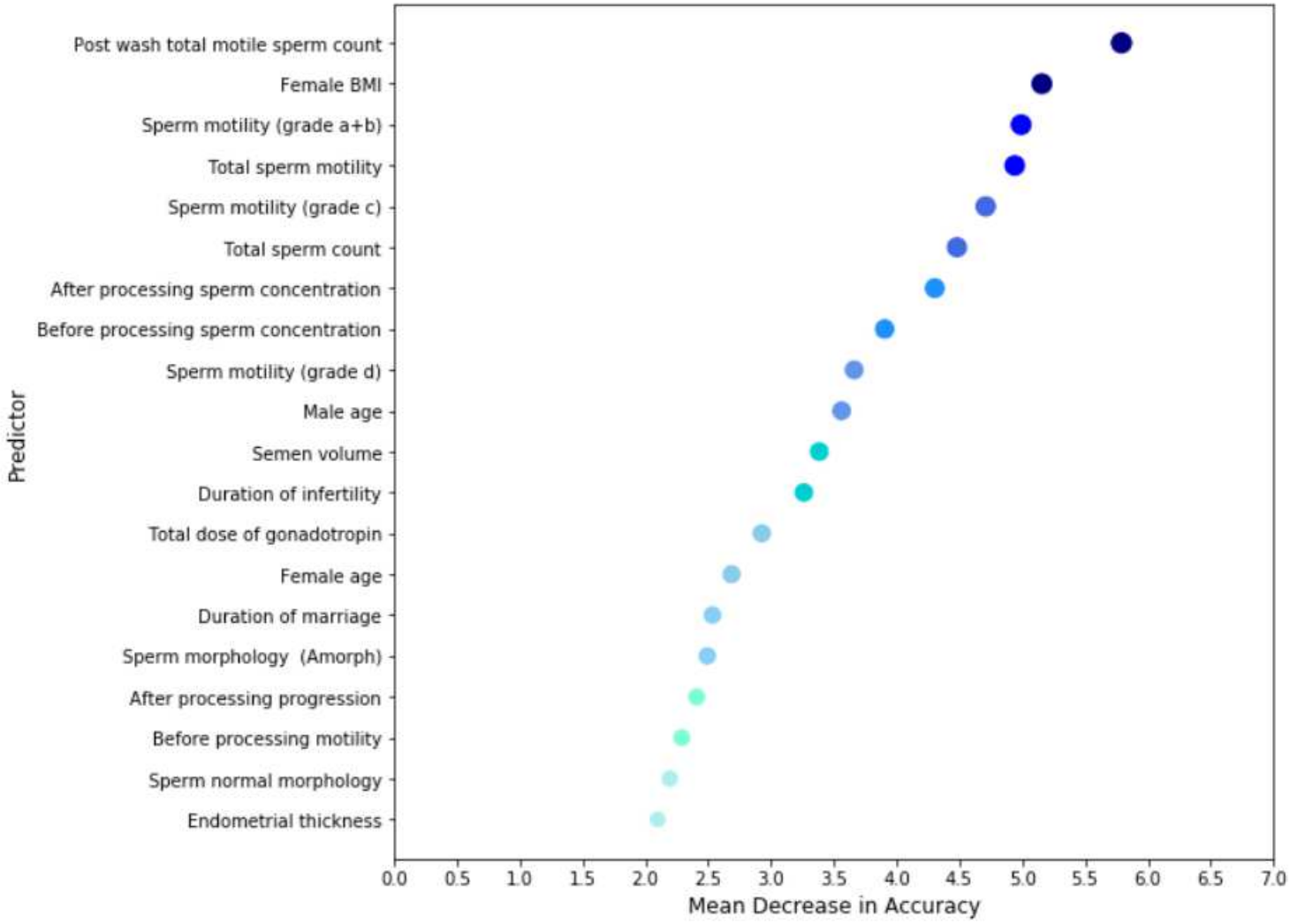

Figure 9

MDA values of top-20 features in this study

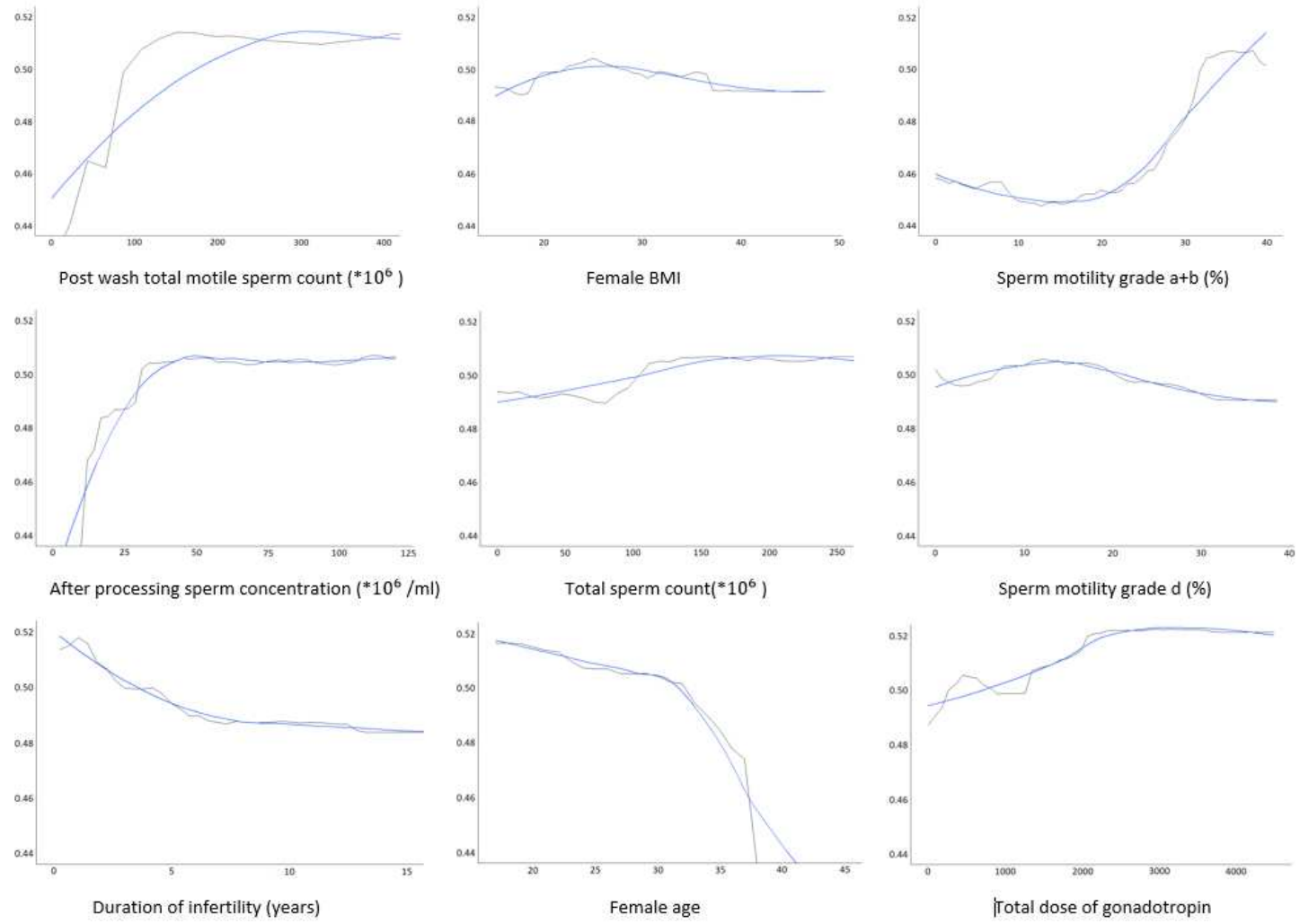

Figure 10

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

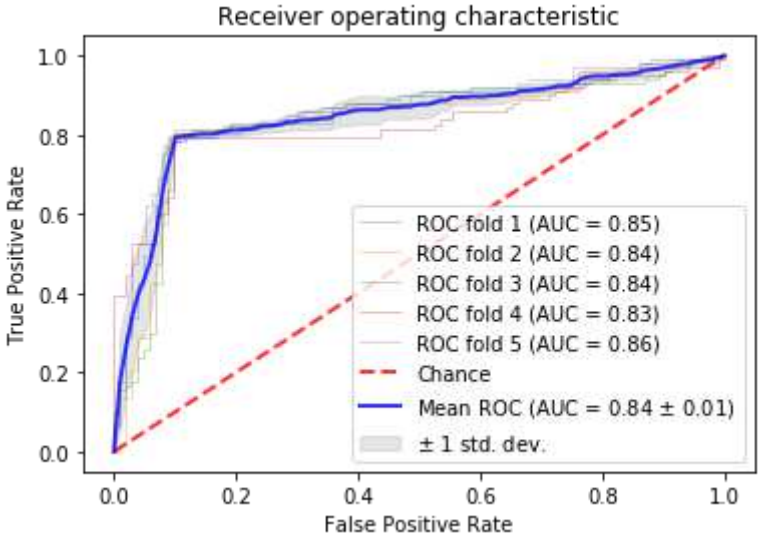

Figure 11

ROC curve for CNFE-SE trained with all features

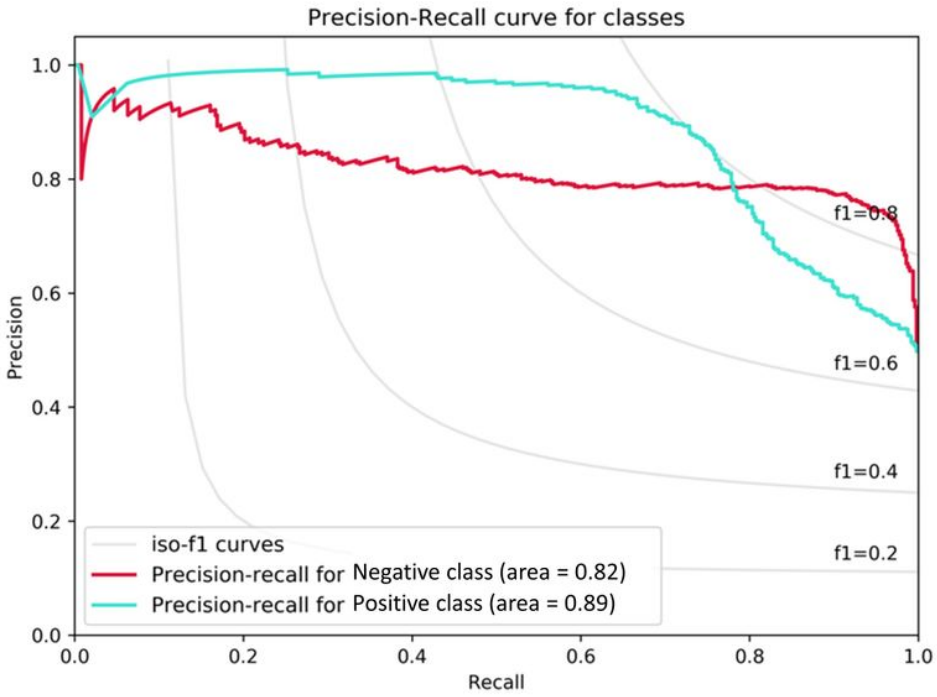

Figure 12

Precision-recall curves for CNFE-SE trained using top-20 features

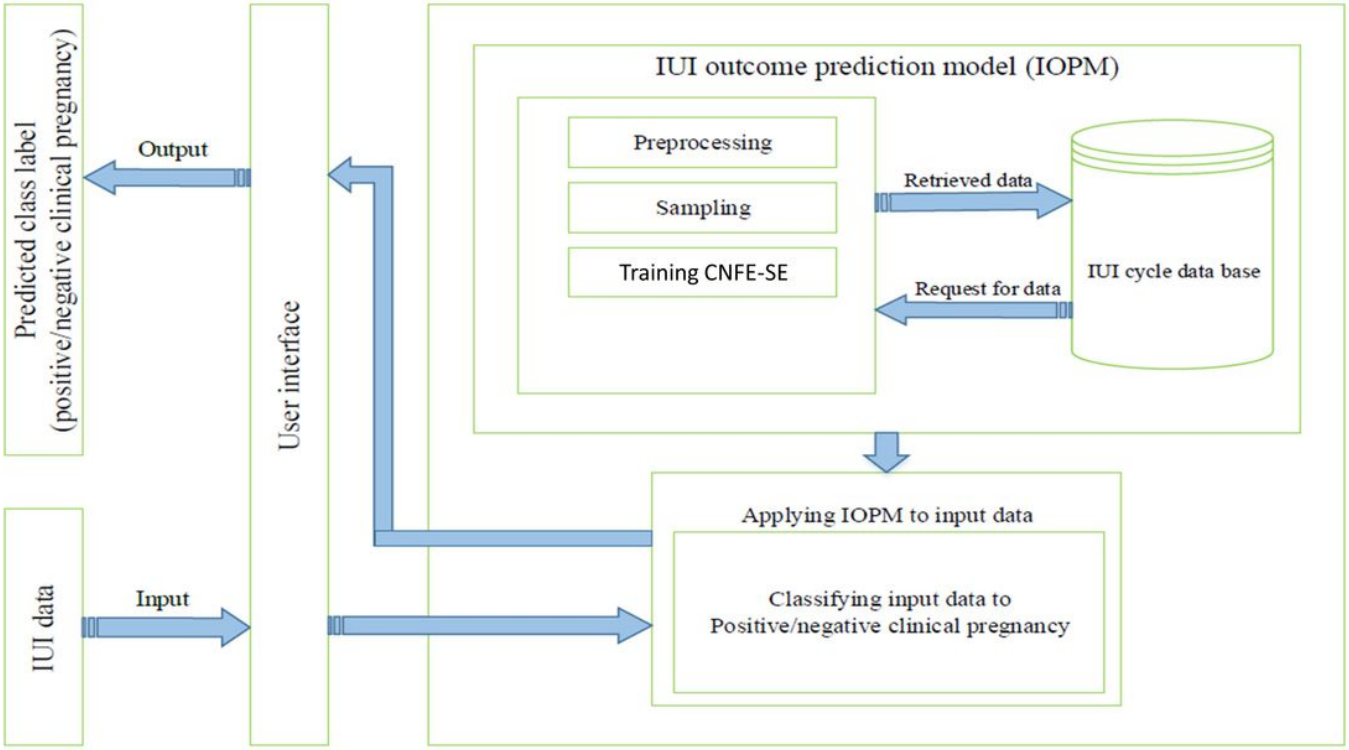

Figure 13

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

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- EditedAppendix.docx