International Journal of E-Health and Medical Communications
Volume 11 • Issue 2 • April-June 2020

Predicting Seminal Quality and its Dependence on Life Style Factors Through Ensemble Learning

Satya Ranjan Dash, KIIT University, Bhubaneswar, India
Ratula Ray, KIIT, Bhubaneswar, India

ABSTRACT

The awareness related to fertility is of great importance due to the change in lifestyle habits. Semen analysis is a reliable confirmatory test to check the fertility in men. The supervised machine learning models of base classifiers include Decision Tree, Logistic Regression and Naive Bayes classifiers in which logistic regression shows a promising accuracy of 88%. Comparing with the bagging ensemble method for the weakest classifier, the results show a leap in accuracy from 78.80% to 90.02%. The authors have also attempted to design a novel voting classifier which votes over the ensemble learners and creates a more complex model to give an accuracy of 89%. Apart from this, the authors have also analyzed the receiver operating characteristic (ROC) curve for Extra Tree classifier which shows a 66% of area under the curve (AUC). The validation procedure used is a 5 fold cross-validation. The authors have further analyzed the lifestyle habits responsible for contributing to this problem based on impurity-based feature selection and have obtained 'Age' as the most crucial factor in declining seminal quality.

KEYWORDS

Decision Tree, Ensemble Learning, Feature Importance, Fertility, Logistic Regression, Naives Bayes, Seminal Analysis

1. INTRODUCTION

Quality of the sperm production in male depends upon the hormonal levels, environmental factors and genetic conditions. It has been seen that in almost 50% of the male suffering for infertility, the cause behind it cannot be determined. The lifestyle factors can directly impact the hormonal levels which in turn impacts the chances of fertility in men. Semen analysis, which is a gold standard for checking for infertility in male, depends on factors such as sperm count, motility, fructose level and pH. All these parameters are related to the lifestyle choices that a male makes during his lifetime and has a profound effect on the fertility results. The major motivation for taking up this work is the need to analyze how the male reproductive health responds to the change in lifestyle habits. Integrity of the male's sperm cells can be disrupted by smoking, alcohol consumption and stress. Oxidative damage can also affect the sperm at the DNA level. Some of these choices are modifiable which

DOI: 10.4018/IJEHMC.2020040105

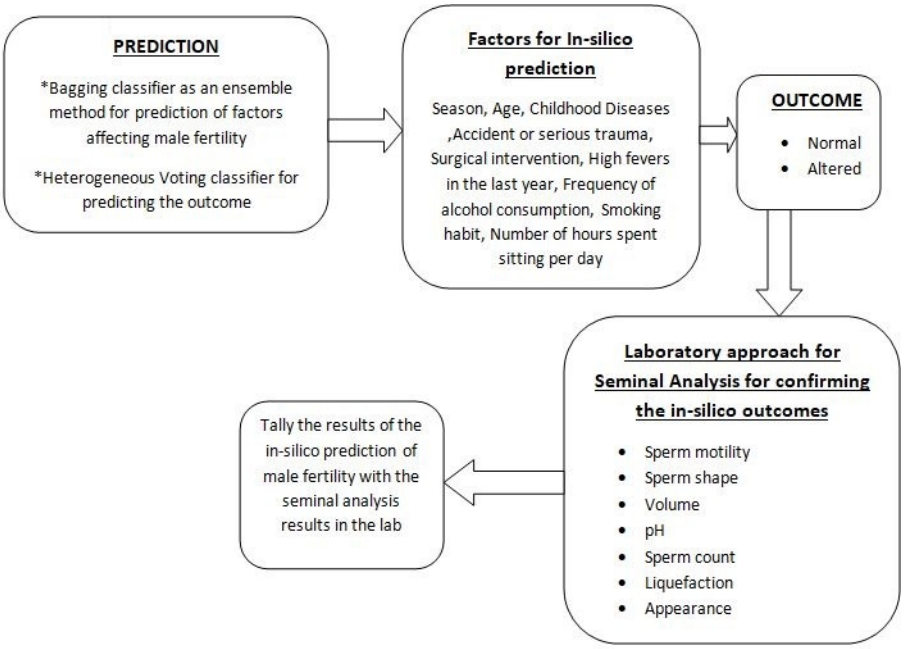
gives a greater chance of avoiding infertility among men if corrected at the right time. It is therefore important to understand how each of these features contribute to the infertility. Stress has been identified to be linked to reduce luteinizing hormone and testosterone secretion in male which can cause spermatogenesis to go down. Age is also been seen as a major factor which is again dependent on issues like job security, because of which people want to delay childbearing. Environmental stress can also contribute to the case, which is much more relevant in industrialized countries (Sharma et al., 2013; Harper et al., 2017). Therapeutic decision making for the couple suffering from infertility is very crucial and in order to achieve that, identifying the causes that lead to this problem is the most important point to look into. Identification of these causes at the right time can lead to an improvement in reproductive health and might increase chances of conceiving.

Current trend in Electronic Health Care (EHR), aimed towards creating a better and more personalized treatment regime for the patients using Deep learning techniques have also been discussed. Studies suggest the roles of certain deep learning techniques in developing effective EHR such as Convolutional Neural Networking (CNN), Multilayer perceptron (MLP), Recurrent Neural Networks (RNN), Restricted Boltzmann Machine(RBM) and Autoencoders (Shickel et al., 2018).The work of Machine learning in clinical applications is not only restricted to diagnose patient's physical health related diseases from a bunch of data but current techniques also make it possible to assess the emotional state of the patient which becomes crucial in certain stages such as in women suffering from postpartum depression after pregnancy (Moreira et al., 2019). Factors other than lifestyle habits of the patients have also been considered in case of assessing the reproductive health of male partners as discussed in the paper. Endocrine Disrupting Chemicals (EDC) such as pthalates in seminal plasma have also been identified as effective biomarkers associated with the diminished quality of the sperm health which affects its morphology and viability (Smarr et al., 2018).

The research gaps in fertility studies exist in the context that there is a knowledge gap when it comes to fertility and its relationship with lifestyle habits. Misconceptions regarding this exist which makes it difficult to implement newer techniques to organize an effective medical regime for patients with infertility. There is a requirement of extensive surveys which will also take into account epigenetic markers that are responsible for occurrences of oligozoospermia and oligoasthenoteratozoospermia. This pool of information at genetic level combined with the power of Machine Learning can solve the research gaps that exist with relation to screening of the factors responsible for infertility in male. Augmentation of patient care using the concepts of Machine learning is the technical motivation for utilizing this technology to move from theoretical to clinical reality. Figure1 gives a layout of the scenario we are aiming for in near future. A number of needs can be met using such concepts in clinical scenario such as reduction in patient readmissions, prevention of hospital acquired infections, prediction of chronic disease, reduction in mortality rate, saving time and increased accuracy in the test results.

Our motivation for this work is to take up a pressing problem like infertility in male and apply the concepts of Machine Learning to it. We have provided an approach for designing a heterogeneous weighted voting classifier which weighs the performance of the individual base classifiers and decides the weight of the models based on statistical index. Our approach is not only based on comparing the performance of bagging classifier as an ensemble method for different base classifiers, but also on designing an effective voting classifier which takes into account the strengths and weaknesses of the base classifier systems. With other ensemble methods, there is always a chance of letting in certain predictive models with inferior performance index in the process of combination, but a weighted voting classifier will exclude them and choose the individual classification models which will provide a relatively better performance, thus proving to be a more reliable tool for classification. Our novelty lies in designing of a heterogeneous weighted voting classifier and applying it for predicting seminal quality in male. Along with it, we also look into how randomness affects the performance of the decision tree model in terms of performance and accuracy:

Figure 1. First step in the supervised machine learning



- Detailed literature survey regarding factors affecting male infertility has been performed;
- Detailed analysis regarding Bagging classifier and its extensions have been done;
- Designing of novel voting classifier as ensemble method;
- Statistical analysis and error detection of the best performing classifier performed;
- Feature importance related to fertility in male have been extracted.

2. LITERATURE REVIEW

A survey was conducted to show that the seminal quality in men have declined over the past 50 years (Carlsen et al., 1992) and over time, with rapid change in lifestyle, the problem has only gotten worse. These changes have also given rise to problems such as testicular cancer, cryptorchidism and hypospadia. A report of a survey linking effects of occupational exposure to semen quality (Jurewicz et al., 2014) was released and it revealed a relationship between occupational exposures and impaired semen parameters. Another study showed how age- dependent seminal quality reduction was a major concerning factor, with the largely affected sperm motility (Eskenazi et al., 2003). Also, ultra-structural abnormalities have been observed in spermatozoa in case of people consuming more amount of tobacco. This has also been linked to erectile dysfunction in case of male (Dai et al., 2015). Chronic alcohol consumption was also associated with arrest of maturation in the germinal sperm cells at the pachytene stage, leading to a halt in the growth of the cells, with no mature sperms (Yao et al., 2016). Fertility preservation options for patients suffering from long term diseases such as cancer is required to be made available (Le Bihan-Benjamin et al., 2018) and that is why the awareness regarding fertility is of much importance. Encouragement of fertility related discussions among adolescents and young adults (Skaczkowski et al., 2018) has taken place and documentation of the proceedings has been recognized as the key to evade problems related to infertility in future.

In Healthcare systems, cost is a major factor and it has direct relation with increase in clinical laboratory tests that are needed to be performed for diagnosis of a disease (Ichikawa

et al., 2016). Some of these tests are mandatory and some are not. Studies have shown where building a strong predictive model using machine learning methods helps to understand which of the tests are actually necessary, and thus also cuts down on medical bills. Machine learning is also used to analyze the risk factor in patients using statistical tools in case of healthcare. Infection preventionists have used machine learning methods as an effective tool to design more effective treatments in order to save time and reduce the risk of life of the patients (Beeler et al., 2018). Machine Learning has allowed artificial intelligence to find out the patterns from real-world datasets and have led to data-driven modeling capabilities (Willcock et al., 2018). This allows decision makers to understand how much level of uncertainty is acceptable and allow them to make important decisions. Using ensemble classifiers in improving models in Machine Learning have been used prior by researchers in various studies which combined the power of more than one base classifiers to form a stronger predictive model rather than just selecting one of them (Raj et al., 2018). Bagging as an ensemble classifier has been used since a long time in order to take in bootstrap samples and averaging the prediction results at the end. Also, in order to remove irrelevant features from the dataset and to lower the complexity of the classification systems, feature selection has been recognized as an important step. Novel strategies to remove the redundant features have been employed in various studies (Gao et al., 2018) in order to increase the performance of the base classifiers being used. Work in this field has also been done using neutrosophic rule-based classification system (NRCS) which generalizes the outcome of fuzzy rule-based classification and gives a better result with lesser complexity and reduction in time consumption. Literature suggests the use of hybrid classification technique called the genetic neutrosophic rule-based classification system (GNRCS) which uses the genetic algorithm in conjunction with NRCS (Basha et al., 2018). Prediction of fertility using IVF (In-Vitro Fertilization) treatment has also been conducted which utilizes the hill-climbing wrapper algorithm for most important features selection and to get rid of the redundant features (Hassan et al., 2018).

In our literature review, we did not find in-depth work done on Bagging as an ensemble method for prediction of fertility based on lifestyle habits and we wanted to exploit the bias free and variance free method in our paper and its impact on prediction, which has lesser computational expense as compared to other methods like Neural Networking being used previously. We have also made an effort to design an ensemble of ensemble classifiers to test whether it can give better performance. Also, our proposed algorithm involves understanding the dependency of lifestyle factors in case of male infertility and up to what extent each of them affects the results of seminal analysis.

3. MATERIALS AND METHOD

3.1. Dataset Details

We have chosen the 'Fertility Dataset' from UCI Machine Learning Repository to serve the purpose of our experiment. The semen samples have been given by 100 volunteers and have been analyzed depending on the criteria set by WHO in 2010. The dataset enlists the attributes and descriptions of the dataset. Season in which the analysis was performed, Age at the time of analysis, Childish diseases, Accident or serious trauma, Surgical intervention, High fevers in the last year, Frequency of alcohol consumption, Smoking habit, Number of hours spent sitting per and Diagnosis are the attributes which have been considered for predicting the class labels:

Number of Attributes: 10
Number of Instances:100

3.2. Classification Systems - Base Classifiers

3.2.1. Decision Tree Classifier

Decision Trees in Machine Learning are statistical tools used for classification and regression purposes. These trees can split the data taken in based on various attributes and reach a final decision. CART refers to Classification and Regression Trees and these are decision trees used for classification and regression modeling systems involved in predicting outcomes and is used for binary classification problems. In CART, at first, we add a root node to the tree. This node will receive all the information of the trained dataset and will ask a True or False question about each feature at the split point on the basis of which further splitting will take place. This splitting gives arise to child nodes which keeps on splitting unless we get the purest form of labels at each node (leaf node). Based on the feature variable, the split point is determined by an attribute value test. This test includes metrics such as Gini Impurity and Entropy (Shannon et al., 2001). Following two equations describe Entropy and Gini impurity (Kim et al., 2016), respectively:

$$H_e(S) = -\sum_{y \in C} p(y) \log_2 p(y) \tag{1}$$

$$H_g(S) = \sum_{y \in C} p(y)(1 - p(y)) = 1 - \sum_{y \in C} p(y)^2 \tag{2}$$

where S represents the dataset, C is a set of classes, and $p(y)$ is the proportion of the number of samples with the class label, y in C. Both Gini impurity and entropy have 0 when there is the only one class in C and reach the maximum value when all classes are equally probable. The split rule is determined by a reduction in entropy and Gini impurity after the split, which is called information gain and can be given as:

$$G(r,S) = H(S) - \sum_t p(t) H(t) \tag{3}$$

where r is a certain split rule and t represent the child nodes induced by r on the set S at the current node. $p(t)$ is the proportion of the number of samples corresponding to t. The attribute and the value for split rule are determined to obtain the maximum information gain at the current node.

3.2.2. Naive Bayes Classifier

Unlike Decision Tree Classifiers, Naive Bayes utilizes probabilistic methods instead of prediction under uncertainty to predict the output result of the class. Given a training dataset, $D = \{X_1, X_2, \ldots, X_n\}$, each data record is represented as, $X_i = \{x_1, x_2, \ldots, x_n\}$. D contains the following attributes $\{A_1, A_2, \ldots, A_n\}$ and each attribute A_i contains the following attribute values $\{A_{i1}, A_{i2}, \ldots, A_{ih}\}$. The attribute values can be discrete or continuous. D also contains a set of classes $C = \{C_1, C_2, \ldots, C_m\}$. Each training instance, $X \in D$, has a particular class label C_i. For a test instance, X, the classifier will predict that X belongs to the class with the highest posterior probability, conditioned on X. That is, the NB classifier predicts that the instance X belongs to the class C_i, if and only if $P(C_i|X) > P(C_j|X)$ for $1 \leqslant j \leqslant m, j \neq i$. The class C_i for which $P(C_i|X)$ is maximized is called the Maximum Posteriori Hypothesis (Farid, D. M., *et al.,* 2014):

$$P(Ci|X) = [P(X|Ci)P(Ci)]/P(X)$$

This algorithm assumes that the features we use are independent of each other. For a Gaussian Naive Bayes model, in addition to the probability calculation for input values for each class, we also calculate the mean and standard deviation to summarize our results.

3.2.3. Logistic Regression

It quantifies the relationship between a dependent categorical outcome and one or more independent predictor variables. The logistic regression gives predicted probabilities for each category (Stylianou et al., 2015). Logistic Regression predicts the output using a function called the logistic function. The logistic function is a, which takes any input and gives the output as a value between zero and one. The equation can be given as:

$$\frac{1}{\left(1+e^{-value}\right)}$$

where, e is the base of the natural logarithms function and 'value' is the actual numerical value that you want to transform. Logistic regression measures the output between categorically dependent variable and one or more independent variables.

3.3. Classification Systems - Ensemble Classifiers

Their basic idea of the ensemble classifiers (Saleh et al., 2017) is that the analysis of different aspects of the problem, through the use of a diverse set of classifiers, can improve the performance of any single individual classification system. In medical studies especially, ensemble classifiers are of great importance because it improves the function of a base classifier, and prevents problem such as over-fitting. Ensemble based classifiers perform better than single classifier by either combining powers of multiple algorithms or introducing diversification to the same classifier by varying input (Bijalwan et al., 2016).

3.3.1. Bagging

Bagging (Kuncheva et al., 2002) stands for Bootstrap aggregating. In particular, it mainly works by reducing the variance component of the loss function (usually, the misclassification probability) of a given base classifier. It utilizes the averaging method to build independent classifiers and average their predictions. This is an effective ensemble method because it greatly reduces variance compared to the single classifiers. Figure2. shows an illustration of Bagging Classifier. Here, n is the no. of instances in the training sample and n' is the number of instances that we put in each bag. The different models trained from each bag is tested upon by similar features (X) to get a mean model which give us the final output(Y).

We can train M different trees to train different subsets of the data to create an ensemble classifier. The equation can be given as the following.

The averaging method of ensemble method can be extended to two more classification systems which uses weak learners like Decision Trees to increase their accuracy results. These are: Random Forest Classifier and Extra Tree Classifier.

3.3.2. Random Forest Classifier

The random forest (RF) model (Patel et al., 2016) is based on the grouping of trees for classification and regression (CART). The method is based on the tree-type classifier. Each tree classifier is named a component predictor. A large number tree makes RF from sub-dataset. The distinctive training set is completed by using bagging. Random Forest Classifier develops lots of decision trees based on random selection of data and random selection of independent features to create a forest. Figure 3 below shows the illustration of a Random forest classifier.

Figure 2. Second step in the supervised machine learning to show us the final predicted outcome

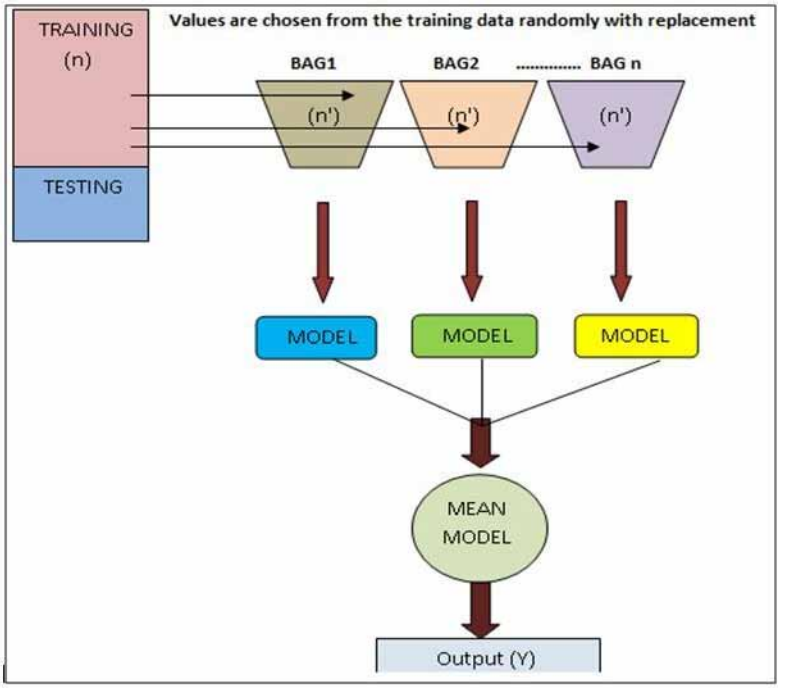
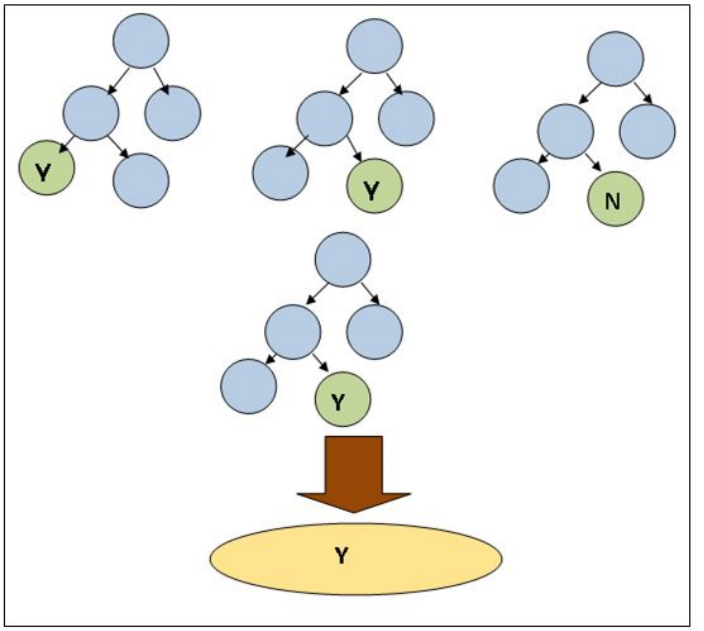


Figure 3. Demonstrates the bagging technique



3.3.3. Extra Tree Classifier

Also referred to as Extremely Randomized Tree Classifier (Geurts et al., 2006), it is a collection of a number of already produced randomized trees which creates an ensemble which gives rise to more randomization. Two main distinctions that it might have with other tree-based classification system are that: firstly, it chooses its cut points entirely at a random state and secondly, it uses the entire training dataset to grow the tree, unlike in case of Bagging. This is used to reduce the variance of the classification model more and produce a much better accuracy.

3.3.4. Voting Classifier

The idea of designing a Voting Classifier involves taking in entirely different classifiers and using either majority vote (hard voting) or averaged probabilities (soft voting) in order to build a much better ensemble method to predict the class on out-of-sample data.

In this paper, we have used an experimental soft-voting technique and have applied on bagged base classifiers. The base classifiers that we used in this paper are of different types and thus, we have chosen to vote on the ensemble of heterogeneous base classifiers. The idea behind this was to build a more complex model and see if it can give a better result than the individual ensemble classifiers(bagged) we had already built by balancing out their weaknesses. In case of Soft voting, we use predicted probabilities 'p' for each classifier and assign weights 'w' in order to get a final average value where we combine all the classifiers. Soft voting returns the class label as *argmax* of predicted probabilities. In the equation, *'i'* is the class label and 'j' is the individual classifier that we are using. The equation for a Soft-voting classifier can be given as:

$$\hat{y} = \arg\max_{i} \sum_{j=1}^{m} w_j p_{ij}$$

These average values for each class label are compared in order to understand which is the best predicted class label and the decision taken has contributions from all the individual classifiers.

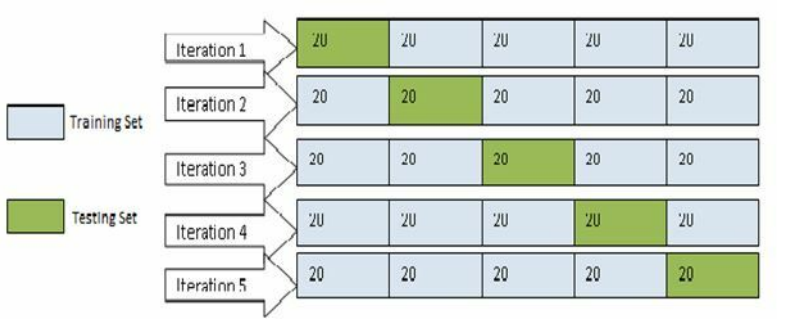
3.4. Validation Procedure

In our paper, we have used k-fold cross validation as a standard method for training our classification models. Train- Test split method can also be used to split between our training and testing data set, but it runs into the problem of under-fitting as a part of the training set has already been removed. In k-fold cross validation system, we divide the entire dataset into k parts with equal amounts of data in each part. Now, we consider (k-1) to be the training model and the rest is the testing model. This is repeated for k times with (k-1) as training set and the rest as validation(testing) set. The error estimation is averaged over the no. of trials conducted and the total effectiveness of the model is calculated accordingly. Figure4. shows us the 5-fold cross validation.

3.5. Feature Importance

Understanding the feature importance is of utmost priority in order to build a better classification model in future. All the features of a dataset are not equally important and understanding this will enable us to choose the best ones. It will reduce complexity in the dataset and provide a much better chance to build an efficient classifier. We have used an impurity-based feature selection method in this paper where the weight of each decision tree in the forest is calculated based on summing of the dependency of each feature at each split point. For an ensemble classifier, these dependencies are averaged. In this paper, we are aiming to isolate the best classifier and analyze the features to understand the importance of each feature.

Figure 4. Depicts a random forest classifier



3.6. Procedure

We have used Scikit Learn (a python library) to apply the concepts of Machine Learning and Ensemble Methods to our dataset. We have used default parameters of the classifiers in the prediction. Figure5 shows the flowchart o the entire working model that we have proposed to build in this paper. The entire procedure can be summarized as follows:

- We at first apply 3 base classifiers: Decision Tree Classifier (CART), Gaussian Naive Bayes Classifier and Logistic Regression to our Fertility Dataset and analyze the performance of each. We have used a 5-fold cross validation procedure;
- Then for each model, we use Bagging Classifier as the ensemble method;
- We then analyze the performance of the Bagged Classifiers to understand which is the weakest of the 3 models;
- We then take this weak model and go on to improve its accuracy by applying extensions of the Bagging Classifier, i.e. Random Forest and Extra Tree classifier;
- We then compare the performance of all these ensemble methods and find out which is showing the best result. The detailed analysis of the best classifier is then done;
- Another novel method has been proposed in this paper, i.e.to create a much more complex ensemble learner over individual ensemble learners (bagged classifiers) by introducing a Soft Voting technique;
- We also use Feature Importance tool to understand the importance of each feature in the best isolated ensemble classifier so that feature selection can be done.

4. RESULTS

4.1. Performance Analysis

Using various metrics, we can compare the performance of the base learners. Table 1 summarizes the average outcome of each of these classifiers used.

From Table 1, it is clear that Decision Tree (CART) is the weakest learner of all the 3 base classifiers. In an attempt to increase their performance, we have used Bagging ensemble method for each of the classifier and analyzed their performance. Table 2. summarizes the outcome of 'Bagging' on base classifiers.

In this paper, we have also shown a novel method for creating a complex ensemble classifier over already existing ensemble learners (Bagged classifiers). We have used the concept of Voting Classifier and have applied Soft voting technique to create a stronger ensemble model. Soft Voting utilizes the average weighted probabilities of each class label from all the classifiers, and estimate

Figure 5. Depicts the k fold cross validation of 100 instances

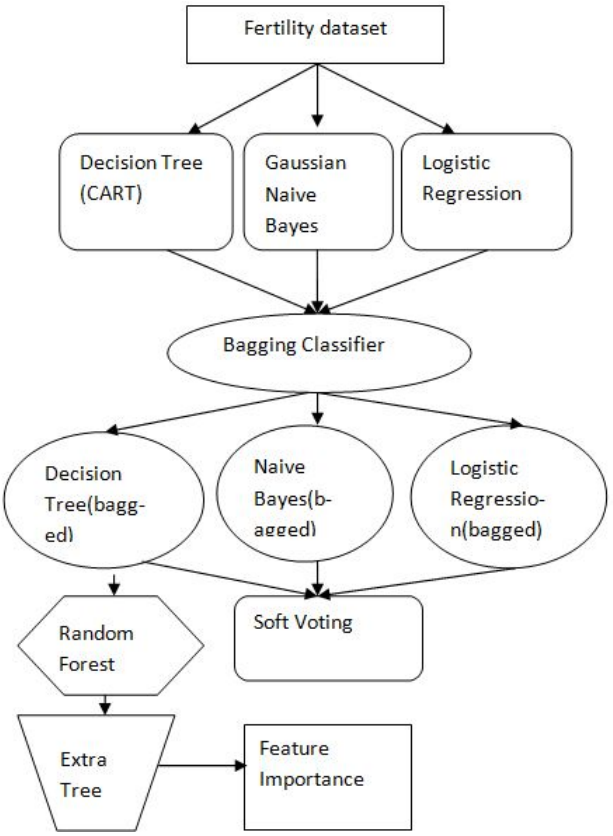


Table 1. Overview of the performance of the base classifiers

Base Classifiers	Precision	Recall	F1 Score	Accuracy%
Decision Tree (CART)	0.79	0.78	0.79	78%
Logistic Regression	0.77	0.88	0.82	88%
Gaussian Naive Bayes	0.77	0.83	0.80	83%

Table 2. Overview of the performance of the bagged classifiers

Classifiers	Precision	Recall	F1 Score	Accuracy%
Decision Tree (Bagged)	0.87	0.89	0.87	88%
Logistic Regression (Bagged)	0.77	0.88	0.82	88%
Gaussian Naive Bayes (Bagged)	0.84	0.87	0.85	84%

which class label shows the highest result. Our aim was to compare the performance of a new complex ensemble (Voting Classifier) over individual Bagged base learners. Bagging is a powerful ensemble technique which averages the predictions of individual base learners and reduce variance of the classifiers. In our model, we have attempted to use these bagged classifiers to construct one single model with improved performance. From Table 3, we can see that the Voting Classifier has indeed shown an increased performance over the individual Bagged classifiers. The Voting classifier which we designed combines the performances of all the ensemble classifiers using average predicted probabilities and vote for the best class labels to provide better accuracy. The Voting Classifier is the complex ensemble of the Bagged classifiers.

One of our main aims of this paper was also to analyze the weak base classifier and through extensions of Bagging ensemble method, build a better classification system. Out of all the classifiers that we have used, we can see that Decision Tree is the weakest learner and we are aiming to build a better version of this. In Table 4, we discuss versions of Bagging Ensemble classifier on the Decision Tree to analyze their performances.

From the data in Table 4, we can say that Extra tree classifier can be considered as a good averaging ensemble technique to be used with this dataset. This classifier takes in extremely randomized forests of decision trees and averages out the result to give an accuracy with lower variance. Comparing it with the base learner Decision Tree that we have used before, we get a 12% increase in accuracy, which is huge. So, we can say that the bagging ensemble classification is indeed an effective technique to improve the performance of the weak learning classifiers like Decision Trees and can be effectively used in medical datasets.

4.2. Feature Importance Analysis

Evaluating the importance of each feature is essential in order to build a much better classifier and enhance its performance. In our paper, we have discussed the most important lifestyle features that contribute to fertility in male and have offered an Impurity based feature selection to isolate the most important features. Understanding this will give us a much clearer idea on what factors mainly this problem depends upon and how the course of treatment should proceed. In medical diagnosis, it is of utmost importance to understand which areas to focus on and how each parameter affects the health of the individual. Here, we will discuss the feature importance in the Fertility dataset

Table 3. Voting classifiers

Classifiers	Accuracy
Decision Tree (Bagged)	88%
Gaussian Naive Bayes (Bagged)	84%
Logistic Regression (Bagged)	88%
Voting Classifier (Soft Voting)	89%

Table 4. Shows the performances of extensions of bagging ensemble method on decision tree

Classifiers	Accuracy%
Decision Tree	78.80%
Bagged Decision Tree	88.12%
Random Forest Classifier	89.07%
Extra Tree Classifier	90.02%

used in Extremely Randomized Forests of decision trees using Filter method, which suggests a statistical method like Gini impurity to assign the score to each feature and get a ranking based on their dependency on class labels.

From the data in the graph in Figure6, we can tabulate the rankings and understand which the most important features are, and which are the redundant ones. Extra Tree Classifier consists of a number of Decision Trees and at each node of the tree, splitting occurs based on a particular feature. The optimal conditions to be selected for splitting at a node is measured by Gini impurity or Information Gain/Entropy in case of classification problems. In case of Decision trees, the main idea of calculating feature importance is to see how much each of the features affect the weighted impurity in the decision tree. In case of a forest of trees, the mean of decreased impurity due to each feature contribution is found out and the ranking is given accordingly in Table 5.

4.3. Result Analysis

The basic difference between a Bagged Decision Tree and Random Forest is that in case of Bagging, we are considering all the features for splitting at the node, but in case of Random Forest, we are considering a subset of features being chosen randomly from the entire set and the best split feature from the subset is used for splitting each node in the Decision Tree. Extremely randomized forests on the other hand tests splits randomly over a subset of features (unlike in case of Random forest where all the splits for a subset of features were considered). This avoids any chance of biasness. As we are introducing more randomization in the Decision Trees (like in case of Extra Tree Classifiers), we are minimizing the chance of variance and thus obtain a fairly reliable model for out of sample predictions. Figures 7, 8, 9 and 10 depict the different graphical representation of the classifiers depending on the accuracy score. From the results, we can conclude that Extra Tree Classifier shows the best accuracy among all other bagging classifiers and the Voting classifier which we designed also gives us a pretty good score.

Figure 6. The workflow of the proposed model

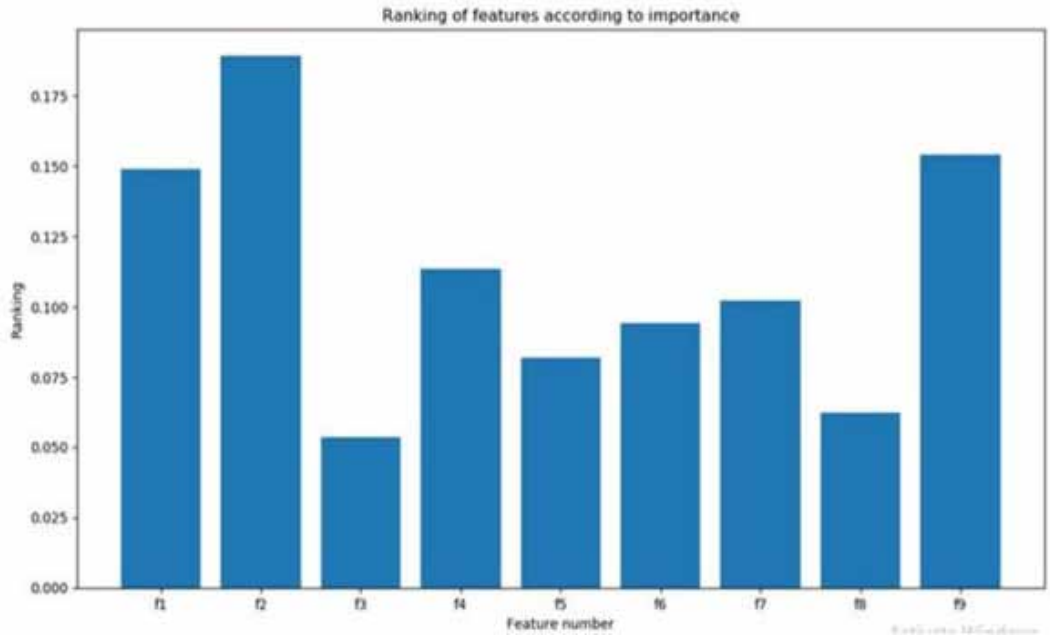
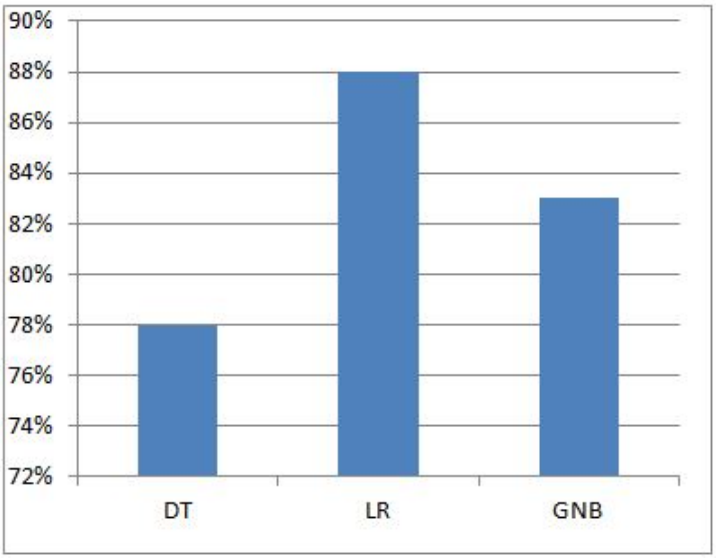


Table 5. Ranking of the features according to the importance

Feature Number	Features	Ranking
f2	Age	0.189
f9	Number of hours spent sitting per day	0.154
f1	Season	0.148
f4	Accident or serious trauma	0.113
f7	Frequency of alcohol consumption	0.102
f6	High fevers in the last year	0.094
f5	Surgical intervention	0.082
f8	Smoking habit	0.062
f3	Childish diseases	0.053

Figure 7. Plot showing an improved accuracy on using extremely randomized forest as ensemble classifier over the decision tree classifier



We have also found out the Confusion Matrix and have done a Receiver Operating Characteristic (ROC) curve analysis for Extra Tree classifier which gives us the maximum accuracy among all the ensemble classifiers used. Our predictions show a 66% of Area Under Curve (AUC) which can be further improved by using GirdSearchCV for optimizing the sensitivity of the classifier. The threshold value can be adjusted to create a balance between sensitivity and specificity in future results. Table 6. shows the Confusion Matrix for Extra Tree Classifier and Figure 11 shows the plot for ROC curve for the same classifier. Other performance analysis parameters for the Extra Tree classifier from the confusion matrix have been depicted in Table 7.

Tables 8 and 9 also discuss about the time required for building different Ensemble models used for prediction of male fertility. Computational time is another important factor which is to be considered. We should aim at building a classifier which gives better result and does not consume much computational time, which is a drawback for powerful classifiers like as neural networks. Our

Figure 8. Plot showing the ranking of the features in order from most important to least important

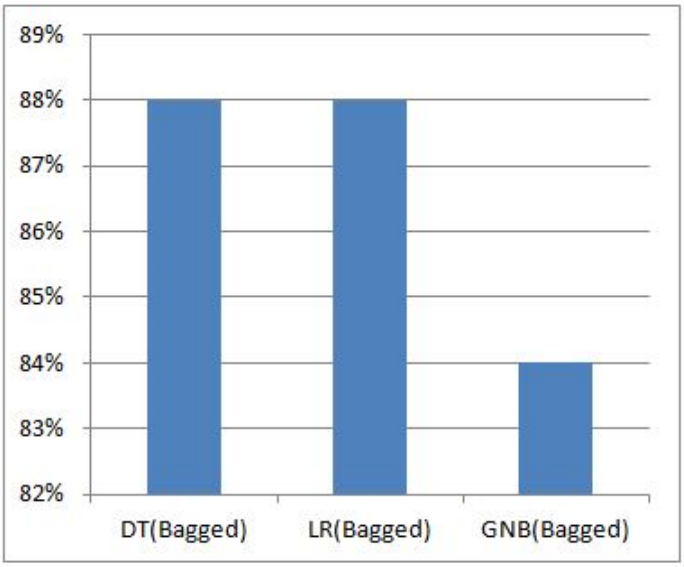
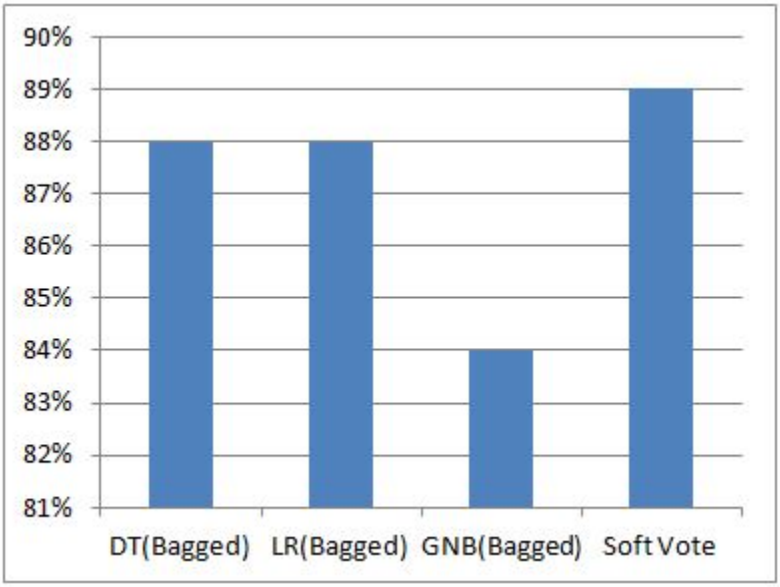


Figure 9. Graphical representation of the classifiers depending on accuracy score



results show that Voting classifier, which combines the performance of the different base classifiers takes maximum time to predict than other ensemble methods.

Understanding the best possible features for deciding the diagnosis for infertility in male is of utmost importance. Since the most improved results in accuracy have been found out to be in case of Extremely Randomized Tree classifier, the feature ranking in case of this ensemble model can give us important information like which lifestyle factors affect most in case of the diagnosis and which are the least important features. From the data that we have collected (refer to Table 5), we

Figure 10. Graphical representation of the classifiers depending on accuracy score

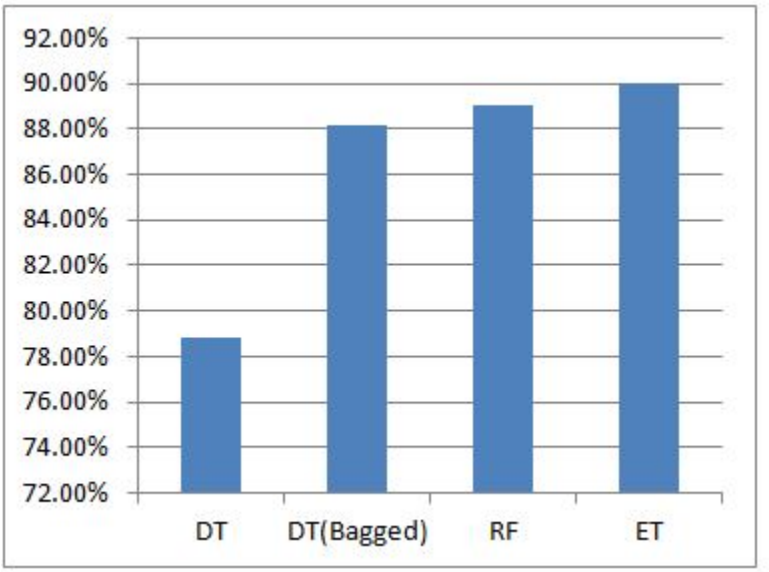
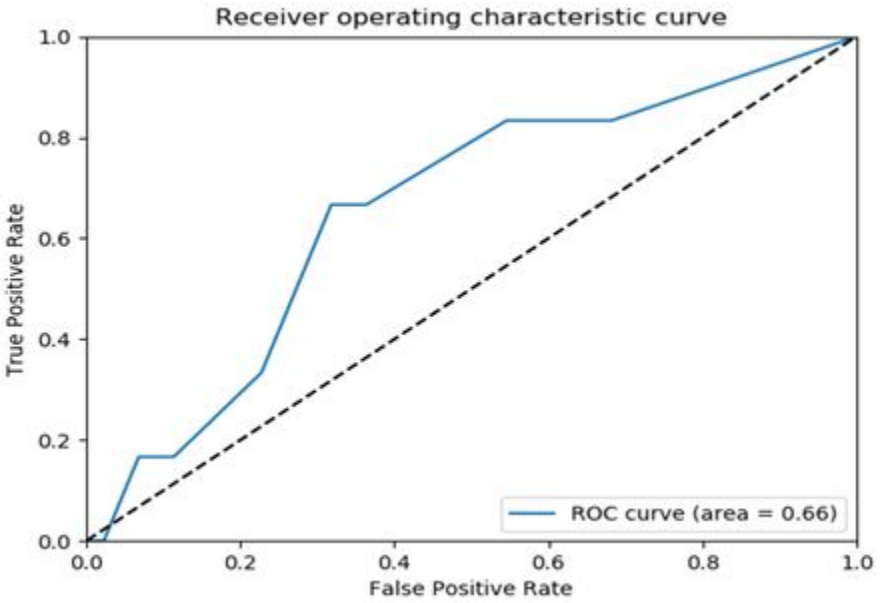


Figure 11. Plot for ROC curve



can understand that age of the person, no. of hours spent sitting per day, and season in which the analysis was performed are the most important features which influences the decision taken by the Extra Tree Classifier. Age is one of the most increasing problems and studies have demonstrated that with increasing age, the sperm quality in male declines. Sperm production in the testes also decline with age because of the reduced number of Sertoli and Leydig cells. Thickening of basal membrane

of Seminiferous tubule and reduced vascularization of the testes contribute to the reduction of these cells (Zitzmann et al., 2013).

5. CONCLUSION

The main aim of our paper was to build a strong classifier system so that we can understand which lifestyle factors affect infertility in males the most. Ensemble classifiers offers a greater chance of accuracy than base classifiers as it combines the power of more than one base classification system together. To ensure than that properties of the classifier matches that of the needs of the application domain, ensemble classifiers become essential for pooling the decision of different base classifiers so that the final decision is being reached at. Ensemble classifiers can offer an additional degree of freedom with respect to bias and variance and helps to reach at a much feasible conclusion for real world problems which are much more complex.

In this paper, we have taken up Bagging classifier and its extensions like Random Forest classifier and Extremely Randomized Forests classifier (Extra Tree) as ensemble methods to understand how effectively it can show an improvement over base classification system like Decision Trees. When we used bagging on base classifiers like Decision Trees, Gaussian Naive Bayes classifier and Logistic Regression model, we find that the most improvement is seen in case of Decision Trees where we get a leap from78% to 88%. While diving deeper into the extensions of bagging, we see an increment in accuracy from 88% in bagged Decision Tree to 90% in case of Extra Tree classifier. Thus, the true strength of Bagging is seen to be in case of unstable base classifiers such as Decision Trees which is extremely sensitive to changes. In this regard, Extremely Randomized classifier shows the best performance according to the results of our experiment. The ROC curve representation of Extra Tree classifier shows 66% of AUC. On the other hand, linear models such as Logistic Regression do not show much improvement with Bagging as they are stable (for this reason, both Logistic Regression model and bagged Logistic Regression model gives same accuracy of 88%). Voting Classifier shows an accuracy of 89% which is more than the performance of the bagged ensemble models.

The future of Machine Learning in clinical experimentations hold promising avenues to be explored. In cases of fertility prediction using Machine Learning, studies are already underway to explore the success rates of IVF in cases of female based upon the dependent variables such as patient's medical history, history of miscarriages, etc. Deep learning techniques such as artificial neural networking are used for such cases and the predictions are being made. Bagging classifier as ensemble method can mildly reduce performance for classifier such as k- nearest neighbors and it might not perform well with non-linear relationship between dependent and independent variables. The deep learning methods can help to eliminate such problems. Success rate of Intra Cytoplasmic Sperm Injection (ICSI) can also be predicted using this method by analyzing the sperm concentration, sperm vitality, etc. Implementation of Machine learning algorithms along with Internet of Things (IoT) are coming up with wearable techs which will allow to predict the time when the woman is most likely to conceive. In cases of male fertility prediction, scopes also lie in areas concerned with image processing associated with studying of the sperm morphology through feature extraction. For grasping a better understanding of the lifestyle factors that can directly correlate with the problem of infertility in males, the collection of more data in future is necessary in order to build a stronger and more reliable classifier system.

REFERENCES

Basha, S. H., Tharwat, A., Ahmed, K., & Hassanien, A. E. (2018). A Predictive Model for Seminal Quality Using Neutrosophic Rule-Based Classification System. In *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics* (pp. 495-504). Cham: Springer.

Beeler, C., Dbeibo, L., Kelley, K., Thatcher, L., Webb, D., Bah, A., & Azar, J. et al. (2018). Assessing patient risk of central line-associated bacteremia via machine learning. *American Journal of Infection Control*, *46*(9), 986–991. doi:10.1016/j.ajic.2018.02.021 PMID:29661634

Bijalwan, A., Chand, N., Pilli, E. S., & Krishna, C. R. (2016). Botnet analysis using ensemble classifier. *Perspectives on Science*, *8*, 502–504. doi:10.1016/j.pisc.2016.05.008

Carlsen, E., Giwercman, A., Keiding, N., & Skakkebaek, N. E. (1992). Evidence for decreasing quality of semen during past 50 years. *BMJ*, *305*(6854), 609–613. doi:10.1136/bmj.305.6854.609 PMID:1393072

Dai, J. B., Wang, Z. X., & Qiao, Z. D. (2015). The hazardous effects of tobacco smoking on male fertility. *Asian Journal of Andrology*, *17*(6), 954–960. doi:10.4103/1008-682X.150847 PMID:25851659

Eskenazi, B., Wyrobek, A. J., Sloter, E., Kidd, S. A., Moore, L., Young, S., & Moore, D. (2003). The association of age and semen quality in healthy men. *Human Reproduction (Oxford, England)*, *18*(2), 447–454. doi:10.1093/humrep/deg107 PMID:12571189

Farid, D. M., Zhang, L., Rahman, C. M., Hossain, M. A., & Strachan, R. (2014). Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. *Expert Systems with Applications*, *41*(4), 1937–1946. doi:10.1016/j.eswa.2013.08.089

Gao, W., Hu, L., & Zhang, P. (2018). Class-specific mutual information variation for feature selection. *Pattern Recognition*, *79*, 328–339. doi:10.1016/j.patcog.2018.02.020

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3–42. doi:10.1007/s10994-006-6226-1

Harper, J., Boivin, J., O'Neill, H. C., Brian, K., Dhingra, J., Dugdale, G., & Hamzic, L. et al. (2017). The need to improve fertility awareness. *Reproductive Biomedicine & Society Online*, *4*, 18–20. doi:10.1016/j.rbms.2017.03.002 PMID:29774262

Hassan, M. R., Al-Insaif, S., Hossain, M. I., & Kamruzzaman, J. (2018). A machine learning approach for prediction of pregnancy outcome following IVF treatment. *Neural Computing and Applications*, 1-15.

Ichikawa, D., Saito, T., Ujita, W., & Oyama, H. (2016). How can machine-learning methods assist in virtual screening for hyperuricemia? A healthcare machine-learning approach. *Journal of Biomedical Informatics*, *64*, 20–24. doi:10.1016/j.jbi.2016.09.012 PMID:27658886

Jurewicz, J., Radwan, M., Sobala, W., Radwan, P., Bochenek, M., & Hanke, W. (2014). Effects of occupational exposure-is there a link between exposure based on an occupational questionnaire and semen quality? *Systems Biology in Reproductive Medicine*, *60*(4), 227–233. doi:10.3109/19396368.2014.907837 PMID:24702586

Kim, K. (2016). A hybrid classification algorithm by subspace partitioning through semi-supervised decision tree. *Pattern Recognition*, *60*, 157–163. doi:10.1016/j.patcog.2016.04.016

Kuncheva, L. I., Skurichina, M., & Duin, R. P. (2002). An experimental study on diversity for bagging and boosting with linear classifiers. *Information Fusion*, *3*(4), 245–258. doi:10.1016/S1566-2535(02)00093-3

Le Bihan-Benjamin, C., Hoog-Labouret, N., Lefeuvre, D., Carré-Pigeon, F., & Bousquet, P. J. (2018). Fertility preservation and cancer: How many persons are concerned? *European Journal of Obstetrics, Gynecology, and Reproductive Biology*, *225*, 232–235. doi:10.1016/j.ejogrb.2018.05.006 PMID:29754073

Moreira, M. W., Rodrigues, J. J., Kumar, N., Saleem, K., & Illin, I. V. (2019). Postpartum depression prediction through pregnancy data analysis for emotion-aware smart systems. *Information Fusion*, *47*, 23–31. doi:10.1016/j.inffus.2018.07.001

Patel, R. K., & Giri, V. K. (2016). Feature selection and classification of mechanical fault of an induction motor using random forest classifier. *Perspectives on Science*, *8*, 334–337. doi:10.1016/j.pisc.2016.04.068

Raj, S. S., & Nandhini, M. (2018). (in press). Ensemble human movement sequence prediction model with Apriori based Probability Tree Classifier (APTC) and Bagged J48 on Machine learning. *Journal of King Saud University-Computer and Information Sciences.*

Saleh, E., Błaszczyński, J., Moreno, A., Valls, A., Romero-Aroca, P., de la Riva-Fernández, S., & Słowiński, R. (2018). Learning ensemble classifiers for diabetic retinopathy assessment. *Artificial Intelligence in Medicine*, *85*, 50–63. doi:10.1016/j.artmed.2017.09.006 PMID:28993124

Shannon, C. E. (2001). A mathematical theory of communication. *Mobile Computing and Communications Review*, *5*(1), 3–55. doi:10.1145/584091.584093

Sharma, R., Biedenharn, K. R., Fedor, J. M., & Agarwal, A. (2013). Lifestyle factors and reproductive health: Taking control of your fertility. *Reproductive Biology and Endocrinology*, *11*(1), 1–15. doi:10.1186/1477-7827-11-66 PMID:23870423

Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE Journal of Biomedical and Health Informatics*, *22*(5), 1589–1604. doi:10.1109/JBHI.2017.2767063 PMID:29989977

Skaczkowski, G., White, V., Thompson, K., Bibby, H., Coory, M., Pinkerton, R., & Osborn, M. et al. (2018). Factors influencing the documentation of fertility-related discussions for adolescents and young adults with cancer. *European Journal of Oncology Nursing*, *34*, 42–48. doi:10.1016/j.ejon.2018.02.007 PMID:29784137

Smarr, M. M., Kannan, K., Sun, L., Honda, M., Wang, W., Karthikraj, R., & Louis, G. M. B. et al. (2018). Preconception seminal plasma concentrations of endocrine disrupting chemicals in relation to semen quality parameters among male partners planning for pregnancy. *Environmental Research*, *167*, 78–86. doi:10.1016/j.envres.2018.07.004 PMID:30014899

Stylianou, N., Akbarov, A., Kontopantelis, E., Buchan, I., & Dunn, K. W. (2015). Mortality risk prediction in burn injury: Comparison of logistic regression with machine learning approaches. *Burns*, *41*(5), 925–934. doi:10.1016/j.burns.2015.03.016 PMID:25931158

Willcock, S., Martínez-López, J., Hooftman, D. A., Bagstad, K. J., Balbi, S., Marzo, A., ... Villa, F. (2018). Machine learning for ecosystem services. *Ecosystem Services*, *33*(B), 165-174.

Wittkowski, K. (1986). Classification and regression trees-L. Breiman, JH Friedman, RA Olshen and CJ Stone. *Metrika*, *33*, 128–128.

Yao, D. F., & Mills, J. N. (2016). Male infertility: lifestyle factors and holistic, complementary, and alternative therapies. *Asian journal of andrology*, *18*(3), 410–418.

Zitzmann, M. (2013). Effects of age on male fertility. *Best Practice & Research. Clinical Endocrinology & Metabolism*, *27*(4), 617–628. doi:10.1016/j.beem.2013.07.004 PMID:24054934

Satya Ranjan Dash is an Associate Professor in School of Computer Applications, KIIT University, Bhubaneswar, India. He received his MCA degree from Jorhat Engineering College, Dibrugarh University, Assam and M.Tech. degree in Computer Science from Utkal University, Odisha. He received his Ph.D. in Computer Science from Utkal University, Bhubaneswar, Odisha in 2015. His research interest includes machine learning, bioinformatics and cloud computing.

Ratula Ray is a student of 4th year, currently pursuing the Dual Degree program in Biotechnology (B.Tech+M.Tech) from KIIT School of Biotechnology, Odhisa. She passed her 12th from St. Mary's Higher Secondary School, New Cooch Behar, West Bengal. Her research interests lie in computational biology, especially in machine learning and understanding how it can be applied to solve medical conditions and its use in clinical applications.