Exploring Female Infertility Using Predictive Analytics

Simi M S
Computer Science
Adi Shankara Institute of Engineering and Technology
Eranakulam, India
simims619@gmail.com

Sankara Nayaki K
Dept.of Information Technology
Adi Shankara Institute of Engineering and Technology
Eranakulam, India

Murali Parameswaran
Dept. Of Computer Science
line 3: *name of organization (of Affiliation)*
Eranakulam, India

Sabine Sivadasan
Dept. of Reproductive Medicine
Sabine Hospital, Muvattupuzha

***Abstract*—With the availability of medical data for large number of patients in hospitals, early detection of diseases has been made easier in the recent past. Conditions like Infertility which are hard to detect or diagnose can be now diagnosed with greater precision with the help of predictive modeling. One of the key challenges for early detection and timely treatment is in identifying and recording key variables that contribute to specific variance of infertility. In this paper we consider 26 variables and identify relevant variables for early detection of 8 variant classes of female infertility. We compared various techniques and determined that the Random forest is the best method offerings 88% of accuracy for a reasonably large hospital dataset of size 965.**

Keywords— Data analytics, Infertility, Classification, Random Forest, J48, Medical Data, Predictive modeling, Healthcare;

I. INTRODUCTION

With the emergence of electronic medical records, there has been a steady growth of data in our medical systems [1] . This growth of data is attributed to growing number of patients as well as the amount of data stored per patient. Data analytics is the process towards developing actionable insights through problem definition and the use of statistical models and analysis against existing data [3] . Analysis of this large data can be leveraged to generate information that enables earlier and better diagnosis of certain diseases [4]. Adopting a comprehensive data analytics platform has thus become essential in the health care industry [2] . Infertility is one group of the diseases that can be treated more effectively if there is advanced warning about likelihood of its incurrence. Medical data related to infertility along with a decision support system shall enable the doctor and hospital to shift towards treatment, health management, and preventive care [5].

According to World Health Organization (WHO), 60 to 80 million people suffer from infertility [6] and 17% females in the age group 20 to 24 suffer from infertility. There are multiple reasons for female infertility to occur. In some cases, there might be certain physiological reasons for the disease. Some of the reasons for infertility can be ovulation disorders, endometriosis, tube damage, uterine disorders, and even due to lifestyle and environmental factors [7] . Sometimes, there might be no apparent reason for the disease. One of the challenges is inordinate time it takes to diagnose the actual cause for infertility. Typically, a test might last six months before a disease can be confirmed, but this delay in diagnosis might affect the probability for complete cure or speed of curing the disease. In our work, we are looking at early detection of infertility problem including unexplained infertility.

Various aspects of infertility problem in health care have been studied in the past [7], [8]. Due to the need to convert the immense medical data sets into actionable knowledge, predicative modeling is been preferred by many researchers using machine learning [9], [10], [11], [4], [12] . In medical domain, application of predicative models is challenging due to two main reasons. Practitioners do not know a priori all the various variables that must be incorporated in the model. Secondly, though almost all hospitals store data, they are seldom available in a single place. Hence the size of data set is typically smaller, affecting the accuracy of the results in real world. Moreover, the data stored is not usually conducive for direct use in any machine learning tool. The available data has to be performed before it can be applied in disease prediction or analysis.

Medical diagnosis of infertility using predictive modeling is still in its nascent state of development. Most articles focus only on predicting infertility as either likely or unlikely [13]. They do not explore the reasons or inferences available in the data. Most of this work has happened in hospitals with limited patient data sets. In our work, we have been to classify into a wider set of inferences and have been able to flag likely, unlikely as well as other probable though not imminent cases of infertility. We have been able to reasonably predict with an accuracy of over 90% for five specific reasons for infertility and with over 80% accuracy for the eight different cases of infertility. In our work, we expanded the number of variables to include twenty six variables in total, of which thirteen variables have been used for the first time by us. Another major contribution of our work is the adaptation of random forest [14] and J48 [15] for prediction.

The remainder of this article is structured as follows: Section 2 describes literature survey of related works. Section

3 describes the dataset used and elaborates on the predictive modeling used in this paper. Section 4 analyses the performance evaluation and discussion of the results. Section 5 concludes with a summary and an outlook.

II. BACKGROUND

Patients records have been used in medical data analysis to uncover hidden knowledge, such as predicting the reasons and diagnosis of diseases [5] . Emir, et al [9] developed a predictive model from a six week observation data to predict the response to pregabalin for the treatment of neuropathic pain. The training dataset had information about 9,187 patients and testing contains 6,114 patient information. They used 8 predictors to analyze result. The prediction suggested that adhering to a pregabalin medication regimen is essential for an ideal end-of-treatment result. Stephanie Revels, et al [10] applied Auto Regressive Integrated Moving Average (ARIMA) [23] time series analysis to model the obesity data published by the Center for Disease Control and Prevention to forecast the future cost associated with obesity related healthcare. They concluded that percentage of population defined as overweight will drop slowly in the next 20 years and thus the cost will also decline. Ali Dag, et al [11] developed a data-driven approach for forecasting survival results at numerous time-points. The developed model is based on decision trees, artificial neural networks, support vector machines, and logistic regression. Their method successfully predicted short, mid & long-term heart transplantation outcomes with an accuracy of 62.4%, 67.6%, and 83.8%, respectively with logistic regression.

Idowu, et al [13] evaluated the use of predictive modelling for dealing with infertility. With a small dataset of 39 patients with 14 attributes, they empirically show that J48, with an accuracy of 87.18% is more effective than Radom forest(53.8% accuracy) for predicting infertility using Weka tool. The dataset used by them is small, especially as a part of it shall be used for training and testing. For a 90:10 ratio of training and testing, the testing phase shall be done with merely 4 inputs which is rather small to make meaningful judgments. Vijayalakshmi et al [16] collected a dataset of 575 patents with 9 attributes by creating a questionnaire regarding various factors like fibroid, endometriosis, cysts, polycystic ovary syndrome (PCOS). They have showed a result of 96.35% accuracy with j48 and 85.9 % accuracy with random forest. However, instead of using clinical data while choosing the parameters, the authors relied on questionnaires. Various studies [17] are available that list the important feature set for identifying infertility. We have chosen 26 attribute values based on our clinical study in hospital.

III. METHODOLOGY

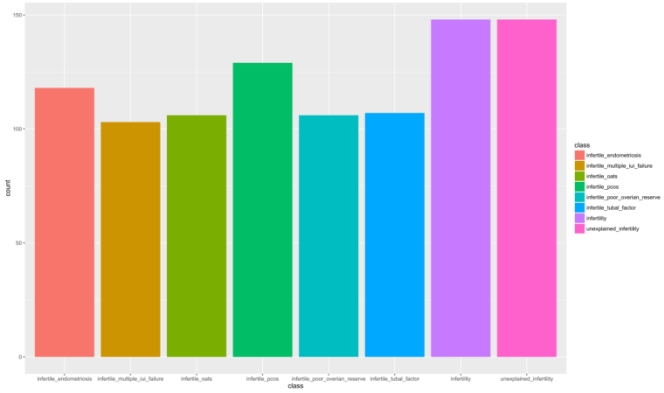
While infertility has numerous definitions by different research bodies some of them are, infertility is a disease of the reproductive system defined by the failure to achieve a clinical pregnancy after 12 months or more of regular unprotected sexual intercourse [18]. Infertility is the inability of a sexually active, non-contracepting couple to achieve pregnancy in one year. The male partner can be evaluated for infertility or subfertility using a variety of clinical interventions, and also from a laboratory evaluation of semen [19] . Fathalla reports that more than 75 million people suffer from infertility worldwide [20] .

A. Medical Dataset

From data analytics point of view the important challenge is finding a right dataset, especially in the case of infertility. We obtained data from a fertility center. This has 965 instances and 26 attributes. It contains records of patients diagnosed between 2014 January to 2016 October. Our analysis of women focused on age less than 50 years (in practical terms, premenopausal subjects). We performed feature selection with Mean decrease Accuracy (MDA), a variable importance in R3 and also by adopting the suggestions from clinical doctors. Variables with a large mean decrease in accuracy are more important for classification of the data [21] . Figure.1 describes the entire dataset.

B. Variable Description

The variables of females we are selected for our study are Age of Patient, Marital life, Weight, Systolic blood pressure, Diastolic blood pressure, Previous history of pregnancy, Hemoglobin, Glucose Challenge Test, Thyroid stimulating hormone, Prolactin, Anti-Mullerian hormone, Endometrium thickness, Right ovary size, Left ovary size, Type of uterus, White blood cells, Luteinizing hormone, Follicle-stimulating hormone, Neutrocytes, Lymphocytes, Bilirubin, Eosinophils, Progesterone , Sodium, Potassium, and Calcium.



C. Variable selection

Developed interest and practical use of data in different scientific areas is in the peak today. The variable importance plot is a critical yield of the data analytics. We would not use all the variables for all analytics. After developing model with all variables feature selection is performed. In this work we used the Mean decrease Accuracy (MDA) variable importance as measured by an ensemble to find the variable importance [21] also doctors suggestions are incorporated for variable selection and selected as Biomarkers. For every variable it discloses how critical that variable is in classifying the data. The plot demonstrates every variable on the x-axis, and their significance on the y-axis. They are ordered top-to-bottom as

most- to least-important. In this manner, the most important factors are at the top and their significance is given by the position on the x-axis. We would to utilize the most essential factors, as decided from the variable importance plot. Ordinarily, we ought to search for a substantial break between factors to choose what number of vital factors to pick. This is an important tool for decreasing the number of factors for data analytics. . For finding the optimal number of variable we performed tuning with cart package in R [22].

IV. Results and Discussion

Two classification methods, Random Forest (RF) [14] and J48 [15] methods were applied to the entire dataset in order to determine the accuracy of predictor. The variable importance feature of RF (Mean Decrease Accuracy) was used initially to determine the important variables [21] . In order to find the number of important variables tuning with cart package were performed [22]. We obtained the result as in the so the number of important variables as 12. The key variables we observed in our study which contributing infertility are age, hemoglobin, anti-mullerian hormone, right ovary, left ovary, luteinizing hormone, follicle-stimulating hormone, Sodium, Pottassium, Calcium, progesterone, lymphocytes.

A. Classification

Firstly J48 model trained with all variables. Classification accuracy was obtained via use training set and 10 fold method for J48 algorithm. In use training set method all the data have been used for training, in 10 fold cross validation splitting the original dataset into 10 equal parts (folds) takes out one fold aside, and performs training over the rest 9 folds and measures the performance repeats the process 10 times by taking different fold each time. First we take all attributes for training and testing. Using training set method, it was measured that there were 933 (96.68%) correct classified instances and 32 (3.32%) incorrect classified instances, showing an accuracy of 96.6%. 10 fold method has been used for creating test and training data, shows that there were 835 (86.52%) correct classified instances and 130 (13.47%) incorrect classified instances, showing an accuracy of 86.5%. Second part implementations of J48 using Biomarkers were performed. Using training set method 931 instances out of 965 were correctly classified (96.47%) and 34 instances are incorrectly classified (3.52%) with an accuracy of 96.4%. With 10 fold cross validation method 846 instances out of 965 were correctly classified (87.66%) and 119 instances are incorrectly classified (12.33%) given an accuracy of 87.7%. The Table 1 shows the results of Infertility Predictor (I.P) and the comparison with the previous studies.

TABLE I. Analogy of J48

Statistics	P1 [20]	P2 [17]	Infertility Predictor (I.P)	I.P with Biomarker Variables
Correctly Classified Instances	96.35 %	87.18%	96.68 %	96.48%
Incorrectly Classified Instances	3.65%	12.82%	3.32 %	3.52%
Kappa statistic	0.90		0.96	0.96
Mean absolute error	0.05		0.01	0.01
Relative absolute error	12.10 %		5.84%	6.19%
Root relative squared error	42.63 %		24.16%	24.88%
Instances/ variables	575/9	39/14	965/26	965/12

The detailed accuracy by class of 10 fold cross validation have shown in Table 2 and Table 3. Table 2 shows detailed class accuracy with True Positive rate, False Positive rate, Precision, Failure Measure, Matthews Correlation Coefficient, Receiver Operating Characteristic Area, Receiver Operating Characteristic Area of 10 fold method with all variables. Table 3 has similar measures applied for 10 fold with Biomarkers variables as predictors. All the values are in the range of 0 through 1, I indicate maximum and 0 indicates minimum.

TABLE II. Detailed Accuracy by Class (10 fold using all variables)

Class	TP Rate	FP Rate	Precision	MCC	ROC Area	PRC Area
infertile_ endometriosis	0.89	0.02	0.89	0.87	0.95	0.88
infertile_ multiple_iui_ failure	0.84	0.02	0.85	0.83	0.94	0.78
infertile_ oats	0.92	0.02	0.84	0.86	0.96	0.83
infertile_ pcos	0.92	0.03	0.84	0.86	0.96	0.82
infertile_poor_ overian_reserve	0.83	0.01	0.88	0.84	0.92	0.81
infertile_tubal _factor	0.94	0.01	0.89	0.91	0.98	0.90
infertility	0.89	0.02	0.91	0.88	0.93	0.82
Unexplained _infertility	0.73	0.03	0.82	0.73	0.85	0.69
Weighted Avg.	0.87	0.02	0.87	0.85	0.93	0.81

TABLE III. J48-DETAILED ACCURACY BY CLASS (10 FOLD USING BIOMARKER VARIABLES)

Class	TP Rate	FP Rate	Precision	MCC	ROC Area	PRC Area
infertile_ endometriosis	0.89	0.01	0.91	0.89	0.95	0.88
infertile_ multiple_iui_ failure	0.86	0.03	0.80	0.81	0.94	0.79
infertile_ oats	0.93	0.02	0.84	0.87	0.97	0.84
infertile_ pcos	0.91	0.03	0.84	0.85	0.96	0.86
infertile_poor_ overian_reserve	0.93	0.007	0.94	0.923	0.97	0.94
infertile_tubal _factor	0.95	0.009	0.93	0.93	0.98	0.89
infertility	0.89	0.01	0.93	0.89	0.94	0.88
Unexplained _infertility	0.70	0.03	0.83	0.72	0.85	0.70
Weighted Avg.	0.88	0.02	0.88	0.86	0.94	0.84

Summary accuracy of J48 has been provided in the following figure.2, x label is detailed accuracy and y label indicates the percentage of accuracy. The red dotted line point accuracy when the whole training set is used for training purpose, blue line points the 10 fold cross validation prediction accuracy.

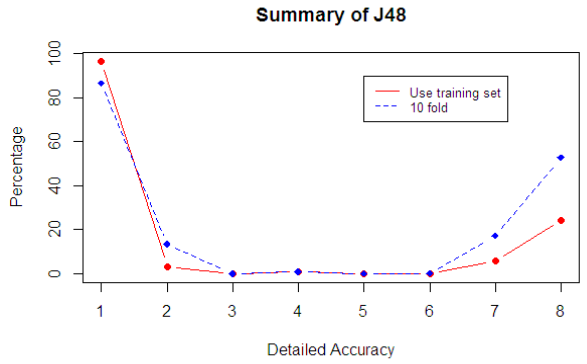
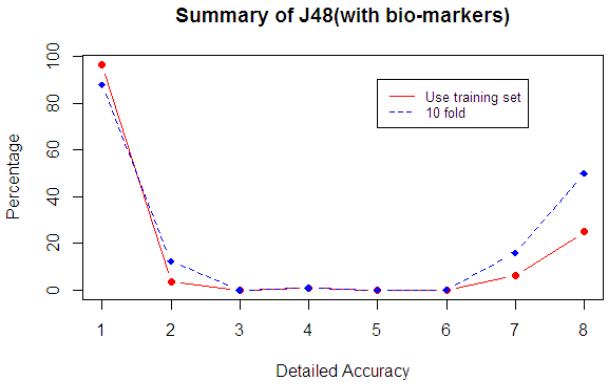


Fig. 2. (a) J48 Accuracy with all variables;



(b) Accuracy with Biomarkers .

1.pctCorrect,2. pctIncorrect, .pctUnclassified, 4. Kappa, 5.meanAbsoluteError, 6. rootMeanSquaredError,7. relativeAbsoluteError, 8. rootRelativeSquaredError

In Random Forest (RF) we developed a percentage split and also 10 fold models. While developing RF model, it take more building time than j48. In percentage split method we split the training set as 70-30 %(70 % for training and 30% for testing). We obtained 250 correctly classified instances out of 285(30% of 965) and 35 misclassification with an accuracy of 87.7%. By applying 10 fold repeated cross validation method with a repetition of 3, 253 correct classifications obtained from 285 instances thus the accuracy is 88.7% and kappa statistics is 0.871. Tune length parameter (mtry) used for the model is 4 and the RF gives the optimal mtry as 5 as in figure 3. Table.4 has shown the RF results and comparison with the previous studies.

TABLE IV. ANALOGY OF RANDOM FOREST

Statistics	P1 [20]	P2 [17]	Infertility Predictor (I.P)	I.P with Biomarker Variables
Correctly Classified Instances	85.91%	53.84%	87.72%	88.77%
Incorrectly Classified Instances	3.65%	12.82%	12.28%	11.23%
Kappa statistic	0.91	-	0.86	0.87
Number of Instances	575	39	965	965
Number of variables	9	14	26	12

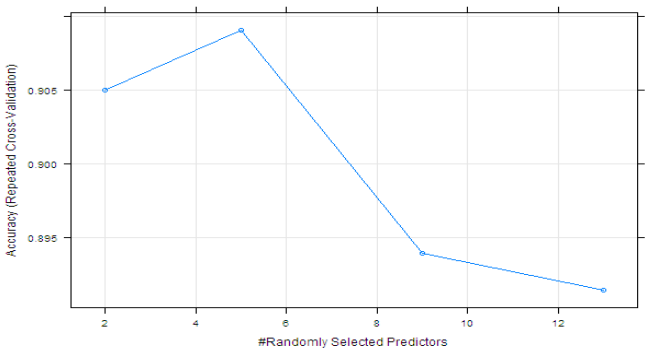


Fig.3. Resampling results across tuning parameters

RF detailed accuracy of class shows in Table.5 with Sensitivity, Specificity, Positively Predicted Value, Negatively Predicted Value, Prevalence, Detection Rate and Balanced Accuracy applied for 10 fold percentage split with biomarkers. We find that the biomarkers improved the accuracy of predictive model when compared with the other model using all variables as predictors (Table.6, Table.7).

TABLE V. .RANDOM FOREST-DETAILED ACCURACY BY CLASS (10 FOLD WITH BIOMARKER PREDICTORS)

Class	Sensitivity	Specificity	Prevalence	Detection Rate	Balanced Accuracy
infertile_ endometriosis	0.97	0.98	0.11	0.11	0.97
infertile_ multiple_ iui_failure	1.0	0.98	0.08	0.08	0.99
infertile_ oats	0.78	0.99	0.13	0.10	0.89
infertile_ pcos	0.94	0.98	0.13	0.12	0.96
infertile_ poor_ overian_ reserve	0.94	0.99	0.11	0.10	0.96
infertile_ tubal_factor	0.94	1.0	0.12	0.11	0.97
infertility	0.80	0.98	0.18	0.14	0.89
Unexplained_infertility	0.83	0.96	0.15	0.12	0.89

TABLE VI. RANDOM FOREST-DETAILED ACCURACY BY CLASS (PERCENTAGE SPLIT WITH ALL VARIABLES AS PREDICTORS)

Class	Sensitivity	Specificity	Prevalence	Detection Rate	Balanced Accuracy
infertile_ endometriosis	0.80	0.99	0.12	0.09	0.89
infertile_ multiple_ iui_failure	0.87	1.0	0.11	0.09	0.93
infertile_ oats	0.97	0.94	0.11	0.11	0.95
infertile_ pcos	0.92	0.99	0.13	0.12	0.96
infertile_poor_ overianreserve	0.97	0.99	0.11	0.11	0.98
infertile_ tubal_factor	1.0	0.99	0.11	0.11	0.99
infertility	0.82	0.98	0.15	0.13	0.89
Unexplained_infertility	0.75	0.97	0.15	0.12	0.86

The Random Forest predictor accuracy using biomarkers with 10 fold cross validation method were shown in figure 4, as in the figure the false classification using Random Forest is minimal in the result and obtained high accuracy of prediction. For 10 fold validation method the entire data is split to 70:30 ratios. First portion was used for training and the remaining for testing. The x label is prediction about classes and y label is count. As the cross validation repeated for three times this given the best predictive accuracy of 88.7%, with 253 correct classifications and 32 misclassifications out of 285 instances.

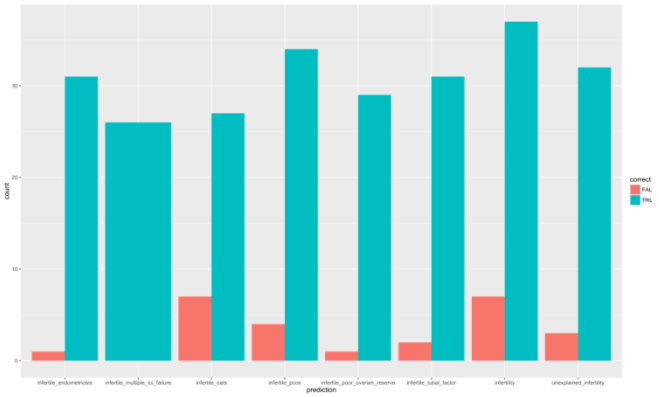


Fig .4. Random Forest Accuracy with Biomarker Predictors

The figure shows that the class infertile_multiple_iui_failure is 100% correctly classified. And for other classes five out of seven has misclassification below five. The unexplained_infertility class only has a misclassification of about 4. This shows that the importance of biomarkers. Table.6 described detailed accuracy with Sensitivity, Specificity, Positively Predicted Value, Negatively Predicted Value, Prevalence, Detection Rate and Balanced Accuracy applied for RF that used all 26 variables as the predictors.

Apart from this, we compared our best model with linear discriminant analysis (LDA) [25] , Classification tree and Bagging Model. Among all models two has the best competing performance, and among that Random Forest is the best method. Table.7 described the results. Two best competing accuracy plot are described in figure.5.

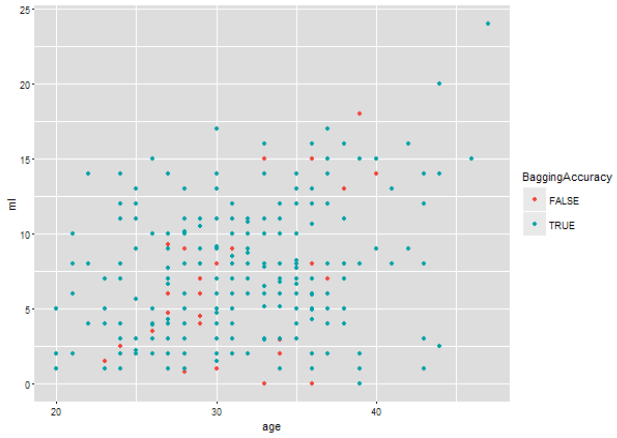
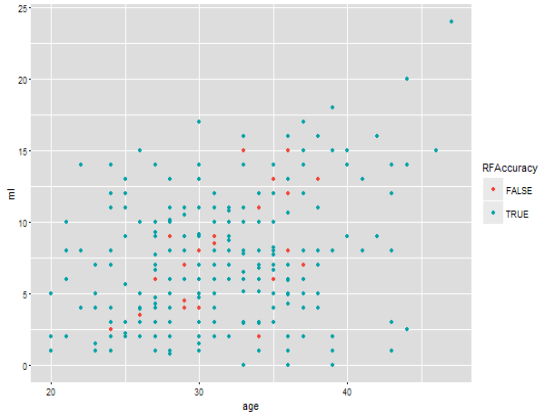


Fig.5. (a) Bagging Accuracy



(b) Random Forest Accuracy

TABLE VII. ANALOGY OF RF ALGORITHMS (PERCENTAGE SPLIT WITH ALL VARIABLES AS PREDICTORS)

Statistics	LDA-predictions	Classification tree-predictions	Random_forest predictions	Bagging predictions
Accuracy	69.47%	76.84%	87.72%	84.56%
Kappa	0.65	0.73	0.86	0.82
Error	30.52 %	23.15%	12.28%	15.43%

The simulation of the prediction models was done using R [3] . RWeka is an R interface to Weka [24]. From the simulation results, it can be inferred that Random Forest with biomarkers as predictors is the most effective and suitable prediction algorithm for infertility. It can be used to support the clinical doctors while decision making process and for the early detection of infertility.

V. CONCLUSION

Medical Data Analytics can possibly change the way the medical researchers utilize complex technologies to gain insight from their medical data repositories and make them to take decisions wisely. This research shows that the best prediction can be done with Random Forest algorithm. Another interesting observation was that the key variable selection improved the performance of predictive model. This will help for the timely detection and treatment of infertility problem. Since no other person on this, whatever results we are observed will definitely have much room for improvement.

REFERENCES

[1] IHTT: Transforming Health Care through Big Data Strategies for leveraging big data in the health care industry. 2013.

[2] Raghupathi W: Data Mining in Health Care. Healthcare Informatics: Improving Efficiency and Productivity. Edited by: Kudyba S. 2010, Taylor & Francis, 211-223

[3] Cooper A,. What is analytics? Definitions and essential characteristics. JISC CETIS Analytics Series, 1(5),2012.

[4] Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inf Sci Syst. 2014; 2:3.

[5] Miner L, Bolding P, Hilbe J, Goldstein M, Hill T, Practical Predictive Analytics and Decisioning-Systems for Medicine. Academic Press,2014.

[6] Calverton, Maryland, USA: ORC Macro and the World Health Organization; 2004. World Health Organization. Infecundity, Infertility, and Childlessness in Developing Countries. DHS Comparative Reports No 9.

[7] Akwasi A. Amoako and Adam H. Balen, Female Infertility: Diagnosis and Management, Endocrinology and Diabetes, pp 123-131, 2015 – Springer.

[8] Practice Committee of American Society for Reproductive Medicine, "Diagnostic evaluation of the infertile female: a committee opinion," Fertility and Sterility, vol. 98, no. 2, pp. 302–307, 2012.

[9] Emir B, Johnson K, Kuhn M, Parsons B, Predictive Modeling of Response to Pregabalin for the Treatment of Neuropathic Pain Using 6-Week Observational Data: A Spectrum of Modern Analytics Applications, Clinical Therapeutics, Volume 39, Issue 1 , Pages 98–106, January 2017.

[10] Stephanie Revels, Sathish A.P Kumar and Ofir Ben-Assuli, Predicting Obesity Rate and Obesity-Related Healthcare Costs using Data Analytis , Health Policy and Technology, 2017, http://dx.doi.org/10.1016/j.hlpt.2017.02.00

[11] Ali Dag, Asil Oztekin, Ahmet Yucel, Serkan Bulur, Fadel M. Megahed, Predicting heart transplantation outcomes through data analytics, Decision Support Systems (2016), doi: 10.1016/j.dss.2016.10.005

[12] Data acience-: Dhar, V. (2013), Data Science and Prediction, Communications of the ACM, 56, 64–73

[13] PA Idowu, JA Balogun, OB Alaba , Data Mining Approach for Predicting the Likelihood of Infertility in Nigerian Women, Handbook of Research on Healthcare Administration and Management, chapter6,pp 76-1042016

[14] Ho, Tin Kam (1995). Random Decision Forests . Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.

[15] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

[16] Vijayalakshmi N, UmaMaheswari M, data mining to elicit predominant factors causing infertility in women, IJCSMC, Vol. 5, Issue. 8, August 2016, pp.5 – 9.

[17] Gesink Law D, Maclehose R, Longnecker M. Obesity and time to pregnancy, Hum Reprod , 2007, vol. 22 ,pp. 414-420

[18] Zegers-Hochschild F, Adamson GD, de Mouzon J, Ishihara O, Mansour R, et al. (2009) The International Committee for Monitoring Assisted Reproductive Technology (ICMART) and the World Health Organization (WHO) revised glossary on ART terminology, 2009. Hum Reprod 24: 2683–2687 doi:10.1093/humrep/dep343.

[19] World Health Organization Laboratory Manual for the Examination of Human Semen and Sperm–Cervical Mucus Interaction 2010 5th edn Cambridge, UK Cambridge University Press.

[20] Fathalla MF (1992) Reproductive health: a global overview. Early Human Development29: 35–42

[21] Louppe G, Louis W , Antonio s and Pierre G, Understanding variable importances in forests of randomized trees, In Advances in Neural Information Processing Systems, pages 431–439, 2013.

[22] Kuhn M., Contributions from Jed Wing SW, Andre Williams, Chris Keefer and Allan Engelhardt. caret: Classification and Regression Training. R package version 5.15-023.

[23] Saboia. 1977. "Auto-Regressive Integrated Moving Average (ARIMA) Models for Birth Forecasting,". Journal of the American Statistical Association, 72: 264–270.

[24] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Peter, R., & Witten, I. H. (2009). The weka data mining software: An update. SIGKDD Explorations, 11(1).

[25] Izenman, A.J.: Linear Discriminant Analysis. Springer (2008)