Predictive Analytics on Female Infertility using Ensemble Methods

Prof. Simi M S[1]
Dr. Manish T I[2]
[1] Assistant Professor, ASIET, India.
[2] Professor, ASIET, India

ABSTRACT

With the accessibility of healthcare data for a significant proportion of patients in hospitals, using predictive analytics to detect diseases earlier has become more feasible. Identifying and recording key variables that contribute to a specific medical condition is one of the most difficult challenges for early detection and timely treatment of diseases. Conditions such as infertility that are difficult to detect or diagnose can now be diagnosed with greater accuracy with the help of predictive modelling. Infertility detection, particularly in females, has recently gained attention. In this work, the researchers proposed an intelligent prediction for female infertility (PreFI). The researchers use 26 variables for the early diagnosis and determine a subset of these 26 variables as biomarkers. These biomarkers contribute significantly to a better prediction of our problem. The researchers designed PreFI using ensemble methods with biomarkers and improved the performance of the predictive system with 98.4% of accuracy.

Key terms-: Health Care; Predictive Analytics; Classification; Data Analytics; Medical data; Random Forest; J48; LDA;

INTRODUCTION

The amount of data in our medical systems has steadily increased with the advent of electronic medical records and increased computing power (IHTT, 2013). The number of patients and the amount of data stored per patient have both increased, resulting in an increase in data. As a result, in the health-care industry, implementing a solid data analytics platform has become critical (Raghupathi, 2010). The process of generating actionable insights by defining problems and applying statistical models and analysis to existing data is referred to as data analytics (Cooper, 2012). The analysis of this large dataset can be used to generate data that will help doctors diagnose diseases earlier and more accurately (Raghupathi, 2014).

Electronic Health Records (EHR) have been incorporated to provide more coordinated and patient-centered care. The use of an Electronic Health Records (EHR) in the ICU significantly reduces central line-associated bloodstream infections and surgical intensive care unit mortality rates (Flatow, 2015). Electronic Health Records (EHR) provide secure access to patient data, which improves care quality and productivity (Tharmalingam, 2016). Electronic Health Records (EHR) systems have been used to manage chronic diseases such as diabetes, and it has been discovered that if providers participate in health information exchanges, regular use of the Electronic Health Records (EHR) can reduce data fragmentation and increase provider continuity of care (Rinner, 2016). Using patient data, specialised AI systems assist specialists in their clinical workflow by recognising and diagnosing various diseases (Simi, 2017). In the emergency department (ED), using a decision tree with Electronic Health Records (EHR) improves medical decision making, increases patient quality of life, and is cost-effective (Ben-Assuli, 2016). Another cost-benefit analysis of using Electronic Health Records (EHR) to collect data yielded encouraging results (Beresniak, 2016).

One of the most common diseases affecting humans is infertility. In accordance with World Health Organization (WHO), this issue affects 60 to 80 million people (WHO, 2004), with infertility affecting 17%

of females between the ages of 20 and 24. More than 186 million people worldwide are infertile, with the majority living in developing countries (Bittles, 2010). Female infertility can occur for a variety of reasons. In some cases, the disease could be caused by physiological factors. Sometimes there is no obvious cause for the disease. Ovulation disorders, endometriosis, tube damage, uterine disorders, and even lifestyle and environmental elements can all contribute to infertility (Amoako, 2015).

The excessive time it takes to detect the true reason of infertility is one of the most typical trends. A test to confirm a condition can take up to six months, however this delay in diagnosis can alter the likelihood of total cure or the pace with which the disease is cured. Our research focuses on the early detection of unexplained infertility issues. Because clinicians are often unable to diagnose the causes of unexplained infertility, the couple must undergo a battery of costly tests to determine the cause of infertility. Clinicians can easily predict unexplained infertility using our proposed system, and the couple can opt for assisted reproductive technology (ART). As there is no time lag between detection and treatment, the success rate of ART can be significantly improved.

Predictive modelling for infertility diagnosis is still in its initial phases of development. The majority of articles only predict infertility as certain or uncertain (Idowu, 2016). They don't look into the data's causes or conclusions. The majority of this research was done in hospitals with limited population data sets. In this work, the authors classify a broader range of inferences and identify likely, unlikely, and other probable (but not imminent) cases of infertility. For five types of ensemble learners, the authors predict with greater than 90% accuracy. The researchers expanded the number of variables in our work to include twenty-six variables in total, thirteen of which the researchers used for the first time. Our work also made significant contributions to prediction by adapting random forest (RF) (T.K. Ho, 1995) and J48 (Quinlan, 1993).

In this work, the researchers explored various available predictive techniques for early diagnosis of female infertility problems and proposed an intelligent prediction for female infertility (PreFI). The major contributions of our work are (a) the identification of key variables that contribute to female infertility and (b) expanding the prediction system from binary classification to a problem of prediction among 9 classes, including unexplained infertility. The researchers used 26 variables for the early diagnosis. The researchers determined a subset of these 26 variables as biomarkers.

These biomarkers contribute significantly to better prediction of our problem. Apart from base classifiers, the researchers also explored ensembles of those algorithms. Ensemble methods are learning algorithms that combine a number of base classifiers and then use a weighted average of their predictions to label new data. In this ensemble, classification is performed using the simple average voting based ensemble of RF, J48, and LDA. The researchers determined that the use of 26 variables improves the prediction of female infertility. By using base classifiers, the researchers demonstrated early diagnosis of nine classes. The researchers designed a new predictive system using ensemble methods with biomarkers and improved the performance of the predictive system. The researchers performed comparative analysis for predicting female infertility using base classifiers and our new ensemble techniques and observed that the mispredictions were substantially reduced.

BACKGROUND

In the current study, the accuracy rates of the developed models were greater than 90%. Predicting female infertility has been the subject of considerable research. Simi (2017) compared the accuracy of two different classification algorithms—J48 and RF—to assess women's infertility. The results showed that RF (96.6%) had a higher accuracy rate than J48 (86.5%). Overall, the accuracy rates of the two algorithms were high. According to Liao S (2019), female-related factors (female cause of infertility, female age, and ovarian response) paired with antral follicle count (AFC) can be used to determine the likelihood of clinical pregnancy. The antral stage follicle biomarkers are also significant predictors of ovarian response in fertility treatments.

Despite the development of advanced ART for the management of various types of subfertile women, such as tubal factor, unexplained infertility, and so on, the ageing process of the ovary and poor ovarian reserve remain a major challenge for clinicians in achieving a successful pregnancy (Coskun, 2018). In a case series by Zhang (2018) on the prediction of the endometriosis fertility index in patients with endometriosis-associated infertility after laparoscopic treatment, 72.38 percent of patients had infertility for less than three years, while 27.62 percent had infertility for more than three years. Primary infertility affected 48.95% of patients, while secondary infertility affected 51.5%. In the same study, 90.1 percent of patients were under the age of 35, 7.93 percent were between the ages of 36 and 39, and 1.91 percent were over the age of 40.

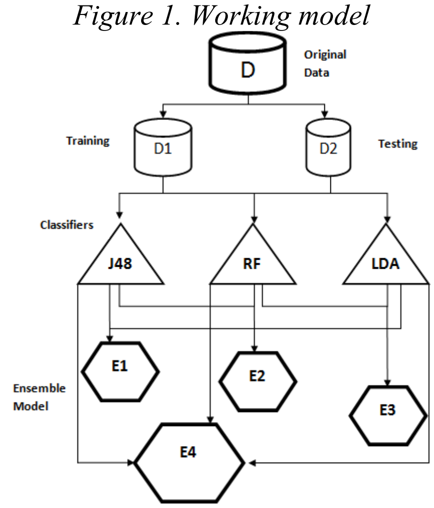
Many studies have been carried out in order to predict the causes of infertility in couples. According to the findings of this study, the support vector machine with a polynomial kernel function predicted with a 76.7% accuracy (Dormahammadi, 2014). Women with PCOS are more likely to become infertile. Infertility is caused by infrequent ovulation, in which the ovary is unable to release a mature egg. This infertility has an impact on a woman's ability to conceive. A study found that 18% of East Indian females have PCOS-related infertility (Palvi Soni, 2018). Cristiana Neto (2021) compares the performance of multiple algorithms, namely, Support Vector Machines, Multilayer Perceptron Neural Network, RF, Logistic Regression, and Gaussian Naïve Bayes. Finally, it was discovered that RF provides the best classification and that using data sampling techniques improves the results, resulting in an accuracy of 0.95 and a precision of 0.96.

MAIN FOCUS OF THE CHAPTER

In this study, the authors use an Ensemble Learner to forecast the cause of infertility in females who are trying to have a baby, decreasing the course of treatment and effectively assisting physicians in planning the course of their treatment. The data was obtained from a fertility clinic by the authors. This has 1678 entries and 26 features. It includes data from patients who were diagnosed between January 2014 and October 2016. The research looked at those under the age of 50. (In layman's terms, premenopausal women.) The authors used mean decrease accuracy (MDA), a variable important in R (Cooper, 2012), as well as clinical doctor recommendations to select features. Features with a high mean decrease in accuracy are more useful in data classification (Louppe, 2013).

The researchers adopted model stacking for developing the Female Infertility Predictor. In fact, the authors also make use of bagging and boosting by applying RF and gradient boosting models, i.e., all three techniques are applied together. The pseudo code for the ensemble is given below, and the working model is also seen in Figure 1.

Figure 1. Working model



PreFI is the ensemble for the prediction of female infertility. Which incorporates the benefits of important variables chosen with the help of MeanDecreaseAccuracy (MDA) variable selection. This ensemble reasonably reduces the time to find the actual cause of infertility, and thus the expenses are also reduced. With infertility, all patients are in the risk zone, whereas 30% of infertility causes are unknown. Patients with unexplained infertility can be easily identified by clinical doctors using this new system, PreFI, and treatment can begin immediately.

Unexplained infertility is currently diagnosed after five to ten years of medical procedures. Since women over the age of 40 have a lower fertility potential than younger women, they are less fertile. In addition, they have significantly lower success rates with fertility treatments such as in vitro fertilization (IVF). If the treatment is started later, the success rate will drop significantly. The success rate of IVF in women over the age of 40 is 10-15%. PreFI will assist those patients who have been diagnosed with unexplained infertility in starting treatment as soon as possible.

Algorithm 1. PreFI

Algorithm 1 Pseudo-code of PreFI

C-: number of classifiers

1. **foreach** c=*1,.....,C* **do**
    - select predictive variables A
    - build classifier model c_A with the selected variables A
    - obtain prediction p_1 from classifier

end

2. apply integration method over predictions $p_1,\ldots\ldots,p_c$
3. obtain prediction P

The researchers developed a total of four ensembles for female infertility prediction. The E1 integrates the prediction results of J48 and LDA. For E2, J48 and RF are combined. E3 is the combination of RF and LDA, and E4 was obtained as the result of integrating the three base learners. Tumer Ghosh (1996) shows how correlation among individual classifiers can affect the performance of an ensemble. So we selected three classifiers with appropriate correlation. Ensemble E1 combines prediction of base learner J48 and LDA. The correlation between them is 0.05. The accuracy of the E1 predictor is 89.9%, with a correctly classified instance of 448 (89.95%) and incorrectly classified instances of 50 (10.04%). The Ensemble E2 integrates the prediction of individual classifiers RF and J48. The accuracy of prediction is 92.8%. The number of correctly classified instances is 462 (92.77%) out of 498, and there are 36 (7.23%) misclassifications. The correlation between RF and J48 is 0.13. The ensemble E3 combined RF and LDA of correlation 0.57. The predictive accuracy is 91.9%. The number of correctly classified instances is 458 (91.97%) and the number of incorrectly classified instances is 40 (8.03%).

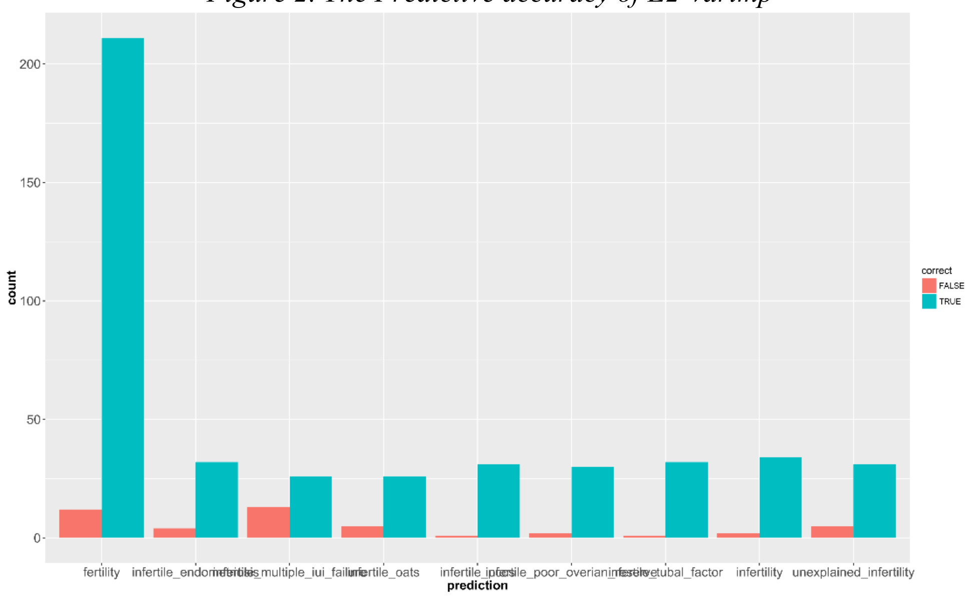
When comparing these three ensembles, it can be seen that Ensemble E2 is the best among the three, with the highest predictive accuracy of 92.8%. So, as the second part implementation, the authors developed Ensemble E2 and E4 with biomarkers as the predictors. Since the biomarkers improve the results of individual classifiers RF and J48, they can also improve the predictive outcome of ensembles. Ensembles E2-varimp used the 12 biomarkers as the predictive variables. The predictive accuracy of E2-varimp is 93.8%, with 467 (93.77%) correctly classified instances and 31 (6.2%) misclassifications out of 498.

Table 1. Class Accuracy of Ensembles E2-varimp

	Class: fertility	Class: infertile_endometriosis	Class: infertile_multiple_iui_failure	Class: infertile_oats	Class: infertile_pcos	Class: infertile_poor_overian_reserve	Class: infertile_tubal_factor	Class: infertility	Class: unexplained_infertility
RF	0.991	0.943	0.733	0.935	0.816	0.968	1	0.841	0.841
J48	0.972	0.914	0.8	0.774	0.789	0.839	0.906	0.818	0.75
StackModel	0.991	0.971	0.733	0.935	0.868	0.968	1	0.886	0.841

Table 1 shows the accuracy of each class of ensemble E2-varimp. Because of the important variable selection, the predictive accuracy of six classes increases rapidly compared to E2 with all variables as predictors. The predictive accuracy of infertile endometriosis goes from 97.7% to 99.1%. Another huge hike is in the case of infertile PCOS, where the accuracy is increased from 78.9% to 86.8%. The variable selection is highly influenced in the case of unexplained infertility. The predictive accuracy attained a huge increase from 65.9% to 84.1%. This is the highest predictive accuracy obtained using a combination of two classifiers along with higher accuracy for other classes. The predictive accuracy of each class is shown in the graph in Figure 2. The x axis stands for prediction, and the y axis represents the count of predictions for each class.

Figure 2. The Predictive accuracy of E2-varimp



In the case of Ensemble E4, the authors combine all three classifiers with biomarker predictors, and the authors are able to obtain an accuracy of 94.8%. 472 correctly classified instances (94.77%) and 26 incorrectly classified instances (5.2%) out of 498. In fact, this is the possible highest accuracy of our Female Infertility Predictor. Table 2 shows the accuracy of each class about ensemble E4.

Table 2. Class Accuracy of Ensembles E4

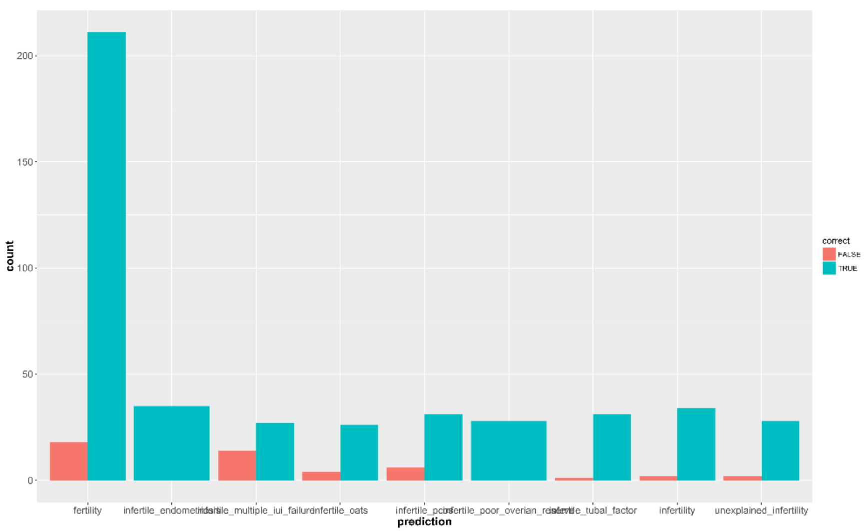
	Class: fertility	Class: infertile_endometriosis	Class: infertile_multiple_iui_failure	Class: infertile_oats	Class: infertile_pcos	Class: infertile_poor_overian_reserve	Class: infertile_tubal_factor	Class: infertility	Class: unexplained_infertility
RF	0.991	1	0.733	0.968	0.816	0.935	1	0.841	0.841
J48	0.972	0.914	0.8	0.774	0.789	0.839	0.906	0.818	0.75
LDA	0.977	0.886	0.467	0.581	0.737	0.806	0.719	0.636	0.545
StackModel	0.991	1	0.8	0.968	0.868	0.935	1	0.864	0.909

The predictive accuracy of each class is shown in the graph in Figure 3. The x axis stands for prediction, and the y axis represents the count of predictions for each class. Detailed accuracy of ensemble E4 with sensitivity, specificity, positively predicted value, negatively predicted value, prevalence, and detection rate of nine classes were described in Table 3.

Table 3. Detailed Accuracy of Ensembles E4

Class Type	Sensitivity	Specificity	Pos. Pred. Value	Neg. Pred. Value	Prevalence	Detection Rate
fertility	0.9906	0.9930	0.9906	0.9930	0.4277	0.4237
Infertile_endometriosis	1.0000	1.0000	1.0000	1.0000	0.0702	0.0702
Infertile_multiple_iui_failure	0.8000	1.0000	1.0000	0.9873	0.0602	0.0481
Infertile_oats	0.9677	0.9850	0.8108	0.9978	0.0622	0.0602
Infertile_pcos	0.8684	0.9913	0.8918	0.9891	0.0763	0.0662
Infertile_poor_overian_reserve	0.9354	0.9978	0.9666	0.9957	0.0622	0.0582
Infertile_tubal_factor	1.0000	0.9978	0.9697	1.0000	0.0642	0.0642
infertility	0.8636	0.9933	0.9268	0.9268	0.0883	0.0763
Unexplained_infertility	0.9090	0.9823	0.8333	0.9911	0.0883	0.0803

Figure 3. The Predictive accuracy of E4

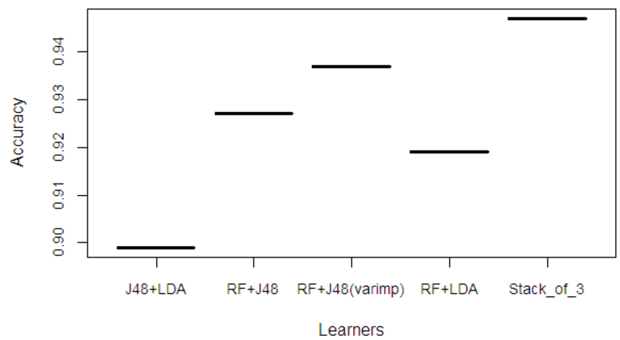


From Table 2, it is clear that for the combination of ensemble rules RF, J48, and LDA, prediction is better than the other combinations. Furthermore, our ultimate aim is to reduce the time lag in the case of unexplained infertility. The causes of unexplained infertility are different from person to person, and the clinical doctors are unable to point out the causes most of the time. Here the E4 shows a significant improvement of more than 90% predictive accuracy in the case of unexplained infertility classes. The authors compared the results of four ensembles. The combination of three rules (E4) performs better, with a predictive accuracy of 94.8%. And finally, the unexplained infertility class is able to predict with an accuracy of 91%.

From Table 2 and Figure 3, the authors observe that for seven classes, there is an improvement in the predictive accuracy and this is the highest among all other combinations. This means that the E4 performs much better than the base learners and the other combinations. And our proposed method could improve the performance of prediction. Which indicates that our method could be used to deal with female infertility detection in an effective manner with no time lag. The combinations have no change in the accuracy of the two classes.

In summary, the authors could conclude as follows: the ensemble with the combination of three base learners with biomarkers as predictors could be able to obtain the highest predictive accuracy and the highest class-wise accuracy, especially in the case of unexplained accuracy. This means that the combinations perform significantly better than the base learners. And our proposed method could improve the performance of prediction. Which indicates that our method could be used to deal with female infertility detection in an effective manner with no time lag.

Figure 4. Summary of all learners



CONCLUSION

In this paper, the authors have presented an ensemble method for dealing with an early female infertility diagnosis. In a variety of applications, combining or integrating the predictions of several classifiers has improved performance. This paper provides an analytical framework for evaluating the benefits of combining classification results. The researchers discovered that combining individual classifiers improved female fertility prediction accuracy. The results of this study make it easier to understand the relationships between variables, classifiers, and combinations in output space. The results show that the proposed ensembles are adequate, as they outperform the results of the individual classifier. This means that with three classifier algorithms as base classifiers, the proposed method could deal with female infertility prediction with a predictive accuracy of 94.8%, which is also better than the conventional method of diagnosis.

ACKNOWLEDGMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCE

A. A.,Amoako, & A.H., Balen (2015), Female Infertility: Diagnosis and Management, Endocrinology and Diabetes, pp 123-131.

A.H, Bittles., & M. L, Black. (2010), The impact of consanguinity on neonatal and infant health, Early Hum Dev 86:737–741.

Akande, V, A., Hunt, L, P., Cahill, D, J., Caul, E, O., Ford, W, C, L., Jenkins, J, M. (2003) Tubal damage in infertile women: prediction using chlamydia serology, *Hum Reprod*, vol. 18 (pg. 1841-1847)

Amadi, L., Onwudiegwu, U., Adeyemi, A., Nwachukwu, C. and Abiodun, A. (2019): Usefulness of Chlamydia serology in prediction of tubal factor infertility among infertile patients at Federal Medical Centre, Bida, North Central Nigeria. *Int J Reprod Contracept Obstet Gynecol*. 8(2): 412-419.

Ben-Assuli, O., Ziv, A., Sagi, D., Ironi, A., & Leshno, M. (2016), Cost-effectiveness evaluation of EHR: simulation of an abdominal aortic aneurysm in the emergency department. *Journal of Medical Systems. 40(6):141*.

Beresniak, A., Schmidt, A., Proeve, J., Bolanos, E., Patel, N., Ammour, N., Sundgren, M., Ericson, M., Karakoyun, T., Coorevits, P., & Kalra, D. (2016), Cost-benefit assessment of using electronic health records data for clinical research versus current practices: Contribution of the Electronic Health Records for Clinical Research (EHR4CR) European Project. Contemporary Clinical Trials. 46:85–91.

Cooper, A. (2012) what is analytics? Definitions and essential characteristics. *JISC CETIS Analytics Series, 1(5).*

Coskun, B., Dilbaz, B., Karadag. B., Coskun, B., Tohma, Y, A., Dur, R. (2018), The role of anti-Mullerian hormone in predicting the response to clomiphene citrate in unexplained infertility. *Taiwan J Obstet Gynecol*, 57:713–7.

DORMAHAMMADI, S., ALIZADEH, S., ASGHARI, M. & SHAMI, M. (2014). Proposing a prediction model for diagnosing causes of infertility by data mining algorithms. *Journal of Health Administration*, 57(17), 46-57.

Flatow, V. H., Ibragimova, N., Divino, C. M., Eshak, D. S., Twohig, B. C., Bassily-Marcus, A. M., and Kohli-Seth, R. (2015), Quality outcomes in the surgical intensive care unit after electronic health record implementation. *Applied Clinical Informatics*. 6(04):611–618.

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

Infecundity, infertility, and childlessness in developing countries (2004) Calverton, Maryland, USA: ORC Macro and the World Health Organization, World Health Organization, DHS Comparative Reports No 9.

Liao, S., Xiong, J., Tu. H., Hu, C., Pan, W., Geng, Y., (2019), Prediction of in vitro fertilization outcome at different antral follicle count thresholds combined with female age, female cause of infertility, and ovarian response in a prospective cohort of 8269 women. *Medicine (Baltimore)* , 98(41):e17470–e17470.

Louppe, G., Louis, W., Antonio, S. & Pierre, G. (2013), Understanding variable importance in forests of randomized trees*, In Advances in Neural Information Processing Systems, pages 431–439.*

M.S., Simi, K.S., Nayaki, M. Parameswaran, & S. Sivadasan (2017), *Exploring female infertility using predictive analytic*, Paper presented at the IEEE Global Humanitarian Technology Conference (GHTC), IEEE (2017), pp. 1-6.

Neto, C., Silva, M., Fernandes, M., Ferreira, D., Machado, J., (2021) Prediction models for Polycystic Ovary Syndrome using data mining. In: Antipova T. (eds) Advances in Digital Science. ICADS 2021. *Adv Intell Syst Comput.,* 1352.

P.A., Idowu, J.A., Balogun, & O.B., Alaba (2016), Data Mining Approach for Predicting the Likelihood of Infertility in Nigerian Women, *Handbook of Research on Healthcare Administration and Management, chapter6,pp 76-104*.

P, Soni. & S, Vashisht., (2018), "Exploration on Polycystic Ovarian Syndrome and Data Mining Techniques", *2018 3rd International Conference on Communication and Electronics Systems (ICCES)*, pp. 816-820.

Quinlan, J. R (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers.

Raghupathi, W, Raghupathi, V. (2010) Data Mining in Health Care. *Healthcare Informatics: Improving Efficiency and Productivity*. Edited by: Kudyba S,Taylor & Francis, 211-223.

Raghupathi, W, Raghupathi, V(2014) Big data analytics in healthcare: promise and potential, *Health Information Science and Systems; 2:3.*

Rinner, C., Sauter, S. K., Endel, G., Heinze, G., Thurner, S., Klimek, P., and Duftschmid, G. (2016), Improving the informational continuity of care in diabetes mellitus treatment with a nationwide Shared EHR system: Estimates from Austrian claims data. *International Journal of Medical Informatics. 92:44–53.*

Simi, M. S. & Nayaki, K. S. (2017, April). *Data analytics in medical data: A review*. Paper presented at the IEEE International Conference on Circuit, Power and Computing Technologies, (ICCPCT), India.

Tharmalingam, S., Hagens, S., and Zelmer, J. (2016), The value of connected health information: perceptions of electronic health record users in Canada. *BMC Medical Informatics and Decision Making. 16(1):93.*

T.K., Ho (1995). Random Decision Forests, *Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 1416, pp. 278-282.*

Tumer, K., Ghosh, J. (1996), Analysis of decision boundaries in linearly combined neural classifiers, *Pattern Recognition, Volume 29, Issue 2, Pages 341-348.*

Zhang, X., Liu, D., Huang, W., Wang, Q., Feng, X., Tan, J. (2018) Prediction of endometriosis fertility index in patients with endometriosis-associated infertility after laparoscopic treatment. *Reprod Biomed Online*, 37:53–9.