International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-2 Issue-6, August 2013

Data Mining Application on IVF Data For The Selection of Influential Parameters on Fertility

M. Durairaj, R. Nandha Kumar

Abstract- This paper illustrates the process applying data mining techniques for identifying influential tests for infertility couples to determine the success rate of IVF (In-vitro Fertilization) treatment. The data set used in the experiments contains information recorded during IVF treatment and relevant laboratory tests [1]. It has supportive information for the medical practitioner to identify which are tests have high impact factors in determining the success of infertility treatment. Data mining has so much of techniques that used to finding the data reduction, pre-processing and normalization [3].The reduced data set contain the set of parameters which have an influence on the results that can be used to predict and forecast [2]. The experiment is in a way of study related to the representativeness of the sample and irrelevant features. Out of around 250 million individuals estimated to be attempting parenthood at any given time, 13 to 19 million couples are likely to be infertile. So the couples prefer the IVF treatment compared with other methods of treatment. In India the board of medical council announced the duration of infertility. If a woman was not conceived after his marriage within 6 months they caused infertility. So they must start the initial fertility treatment. Most of them prefer the In-Vitro fertilization compare with other fertility treatments [9]. A survey of the fertility treatment 1 in 20 of all pregnancies conceived by the ivf treatment. But the patients suffer from the negative imagination and they don't know the success level of the treatment.

The prediction of the success rate of IVF treatment has a great economic importance for the couples who undergo treatment for baby [2]. The data set are preprocessed by the supervised filter and the attribute selection algorithm before subject to the prediction. It is very essential to properly analyze the data set and reduce or clean the unwanted data that increases the prediction accuracy [6]. The parameters with high impact factor can be selected by applying the proper reduct algorithm, which removes the parameters that has a lesser role in determining the success rate of particular patients and help the Gynecologists to recommend them for specific treatment of IVF, IUI or ICSI.

Keywords- Attribute selection algorithm, Data mining, IVF, spermatological data, supervised filter.

I. INTRODUCTION

Technique of Data mining is for the automated discovery of knowledge or extracting useful information from the large databases. In data mining, different types of methods and algorithms have been proposed; in this work for mining information a data mining software tool Waikato Environment for Knowledge Analysis (WEKA) is used.

This tool offers the option to select suitable algorithms and techniques for knowledge discovering process. Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization and feature selection. All of Weka's techniques are predicated on the assumption that the data is available as a single flat file or a relation, where each data point is described by a fixed number of attributes [1]. In the Weka tool, preprocess and classify methods are used to extract the data from the large volume.

In before IVF treatment the IUI (Intra Uterine Insemination) treatment has to be done for infertility. But the most important risk of treatment with intrauterine insemination in a stimulated cycle is the risk of multiple pregnancies [11]. The main importance of IVF treatment prediction helps the patient has relief with Psychological and mental pressure. The infertility is one of the globally rising social problems and it affects the people physically and psychologically [5]. So the infertile couple starts a treatment with multiple test and analysis. But they don't know the success level and the stand in inflexible situation. They spend more money, time and health for avoiding infertility [10]. So if anybody produce the success rate of the treatment means the patients may trained and ready for the treatment without fear and negative opinions.

In-Vitro Fertilization is one of the most effective treatments for addressing infertility causes. This study is to find reduced set of the parameters without compromising knowledge and omits the less influential parameters set [4].In the study of fertility data analysis, supervised algorithm is used for preprocessing and feature selection technique is used for identifying influential parameters set from the outcome of IVF functional tests. The IVF tests outcomes such as Age of patient, Duration of infertility, Previous pregnancy, Assisted Reproductive Procedure, Spontaneous conception, Body Mass Index(BMI), Major Psychological factors, Endometriosis, Tubal Infertility(TI), Ovulatory factor, Cervical factor, Unexplained factor, No. of embryos transferred, No. of oocytes retrieved, Male Factors, Semen Ejaculate volume, Gross and microscopic appearance, Sperm morphology and Sperm motility are used in the study and subject to the attribute selection through supervised filter in WEGA. The selected set contains the most significant parameters that can be used to predict IVF success rate [6]. The reduction algorithm used is capable of reducing the attribute sets of the information system without changing basic knowledge of the system [7]. Section I describes the materials and methods used for experiments. The section II describes the experiments carried out.

Manuscript published on 30 August 2013.
***** Correspondence Author (s)
Dr. M. Durairaj*, Asistant Professor, Department of Computer Science, Engineering and Technology, Bharathidasan University, Tiruchirappalli, India.
Mr. R. NandhaKumar, Research Scholar, Department of Computer Science, Engineering and Technology, Bharathidasan University, Tiruchirappalli, India.

In this section, the artificial neural network construction, confusion matrix with error rate generation and results obtained are discussed in detail. The last section ends with the conclusion and future work descriptions.

II. MATERIALS AND METHODS

A. Data set

The IVF data set used for this work is collected from the specialty test tube baby hospitals and IVF research centers in Tamilnadu. The data samples directly collected from the infertility couples undergoing treatments and the couples who successfully delivered babies through IVF treatments. Data set contains 250 patients' records and 27 different tests parameters known as attribute sets.

B. Materials and Methods

1. Basic concept of attribute selection algorithm

An attribute selection algorithm can be seen as the combination of a search technique for proposing new feature sets, along with an evaluation measure which scores the different feature subsets. The simplest algorithm is to test each possible subset of features finding the one which minimizes the error rate. This is an exhaustive search of the space, and is computationally interacting for all but the smallest of feature sets. In WEGA tool, the choice of evaluation metric mostly influences the algorithm, and these evaluation metrics distinguishes between the three main categories of feature selection algorithms: wrappers, filters and embedded methods.

Filter methods use a proxy measure instead of the error rate to score a feature set. This measure is chosen to be fast to compute, whilst still capturing the useful of the feature set. Common measures include the Pearson product-moment correlation coefficient, and inter/intra class distance. Filters are usually less computational intensive than wrappers, but it produce a feature set which is not tuned to a specific type of predicting method. Many filters provide a feature rank rather than an explicit best feature set, and the cutoff points in the ranking are chosen via cross-validation and testing [7].

2. Minimum-redundancy-maximum-relevance (mRMR) attribute selection

The mRMR can use either mutual information, correlation, distance/similarity scores to select features. For example, with mutual information, relevant features and redundant features are considered simultaneously. The relevance of a feature set S for the class c is defined by the average value of all mutual information values between the individual feature f_i and the class c as follows:

$$D(S,c) = \frac{1}{|S|} \sum_{f_i \in S} I(f_i;c) \quad (1)$$

The redundancy of all features in the set S is the average value of all mutual information values between the feature f_i and the feature f_j:

$$R(S) = \frac{1}{|S|^2} \sum_{f_i,f_j \in S} I(f_i;f_j) \quad (2)$$

The mRMR criterion is a combination of two measures given above and is defined as follows:

$$mRMR = \max_S \left[\frac{1}{|S|} \sum_{f_i \in S} I(f_i;c) - \frac{1}{|S|^2} \sum_{f_i,f_j \in S} I(f_i;f_j) \right]. \quad (3)$$

Suppose that there are n full-set features. Let x_i be the set membership indicator function for feature f_i, so that $x_i=1$ indicates presence and $x_i=0$ indicates absence of the feature f_i in the globally optimal feature set. Let $c_i=I(f_i;c)$ and $f_{ij}=I(f_i;f_j)$. The above may then be written as an optimization problem:

$$mRMR = \max_{x \in \{0,1\}^n} \left[\frac{\sum_{i=1}^n c_i x_i}{\sum_{i=1}^n x_i} - \frac{\sum_{i,j=1}^n a_{ij} x_i x_j}{\left(\sum_{i=1}^n x_i\right)^2} \right]. \quad (4)$$

It may be shown that mRMR feature selection is an approximation of the optimal dependency feature selection that maximizes the mutual information between the joint distribution of the selected features and the classification variable. However, since mRMR turned a combinatorial problem as a series of much smaller scale problems, each of which only involves two variables, the estimation of joint probabilities is much more robust. In certain situations the algorithm can underestimate the usefulness of features as it has no way to measure interactions between features. This can lead to poor performance when the features are individually useless, but are useful when combined. Overall the algorithm is more efficient than the theoretically optimal dependency reduction, yet produces a low redundancy feature set.

3. IVF data preprocessing using supervised filter

A supervised attribute filter that can be used to select attributes. It is very flexible and allows various search and evaluation methods to be combined. The options available for Attribute Selection are Evaluator, Search. It also supported the Missing class values, Nominal class, Numeric, Date, Binary and No class [8].

In Table 1, the set of in-vitro and spermatological data set and the data that has the highest influence on fertilization rate prediction are illustrated.

Table 1 IVF Reduced data set by applying Weka's Attribute Selection Tool

SLNO	Considered parameters	Significant influence of fertility
1	Age(F)	NO
2	Age(M)	NO
3	Duration Of Infertility (Years)	NO
4	Previous Pregnancy	NO
	a. If Yes, Previous Miscarriage	NO
	a1. If Yes Miscarriage Caused	NO
6	Medical Disorders	NO
7	Previous Surgery	NO
8	Body Mass Index (BMI) (F)	NO
9	Pre-Existing Symptoms of Depression	NO
	a. Fear And Negative Treatment Attitude	NO
	b. Psychological And Emotional Factors	NO
	c. Difficulty In Tolerating Negative Emotions For Extended Time	NO
	d. Uncertainty	NO
	e. Strain of Repeated Treatment	NO
10	Endometriosis	YES
11	Tubal Infertility	NO
12	Ovulatory Factor	NO
13	Hormonal Factor	NO
14	Cervical Factor	NO
15	Unexplained Factor	YES
16	Semen Ejaculate Volume	NO
17	Liquefaction Time	NO
18	Gross And Microscopic Appearance	NO
19	Sperm Concentration	YES
20	Sperm Motility	YES
21	Sperm Vitality	NO
22	Sperm Morphology	NO
23	No.of Oocytes Retrieved	NO
24	No.of Embryos Transferred	NO
25	Male Factor Only	YES
	a. Severe Male Factor	NO
26	Female Factor Only	NO
27	Unknown Factor	NO

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

AF-Age female, DI – duration of infertility, PP- previous pregnancy, MiC-miss carriage caused, END-endometriosis, MD-medical disorder, PS-previous surgery, TI-tubal infertility, OF-ovulatory factor, HF-hormonal factor, CF-cervical factor, UF-unexplained factor, SEV-sperm ejaculated volume, LT-liquidation time, SC-sperm concentration, SM-sperm motility, SV-sperm vitality, SPM-sperm morphology, NOR-number of oocytes received, NET-number of embryos transferred, AR-assisted reproductive.

The sample IVF data set collected from different fertility clinics and maternity hospitals are illustrated in Table 2. The table contains the data of individual couples who undergone treatments for test tube baby through various tests.

Table 2. IVF data collected from infertility clinics and research centers (sample)

Patients No (Couples)	AF	DI	PP	MiC	MD	PS	END	TI	OF	HF	CF	UF	SEV	LT	SC	SM	SV	SPM	NOR	NET
001	20-30	3	YES	AR	NO	NO	NO	NO	NO	NO	YES	NO	YES	YES	YES	NO	YES	YES	12	4
002	20-30	4	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO	YES	YES	YES	YES	YES	YES	4	3
003	31-40	12	YES	AR	NO	NO	YES	YES	YES	NO	NO	NO	YES	YES	YES	NO	YES	YES	3	3
004	20-30	7	NO	NO	NO	NO	NO	NO	YES	NO	YES	NO	YES	YES	YES	YES	YES	YES	2	3
005	20-30	7	YES	AR	NO	NO	NO	NO	YES	NO	YES	NO	YES	YES	YES	NO	YES	YES	6	3
006	20-30	6	NO	NO	NO	NO	NO	NO	YES	NO	YES	NO	YES	YES	YES	NO	YES	YES	1	2
007	20-30	2	NO	NO	NO	NO	NO	NO	YES	NO	YES	NO	YES	YES	YES	NO	YES	YES	7	3
008	20-30	3	NO	NO	NO	NO	NO	NO	YES	NO	YES	NO	YES	YES	YES	YES	YES	YES	16	4
009	20-30	5	NO	NO	NO	NO	NO	NO	YES	NO	YES	NO	YES	YES	YES	YES	YES	YES	9	3
010	20-30	3	NO	NO	NO	NO	NO	NO	YES	NO	YES	NO	YES	YES	YES	NO	YES	NO	2	3
.	.	.	.																	
.	.	.	.																	
.	.	.	.																	
250	31-40	7	NO	NO	NO	NO	NO	NO	YES	NO	YES	NO	YES	YES	YES	YES	YES	YES	6	3

III. EXPERIMENTS

The data set are preprocessed to select only most influential parameters using attribute selection algorithm, which filters the noisy data resulted in the selection of parameters with high impact factors. The selected data set is subjected to the further classification, as illustrated in Fig. 1.

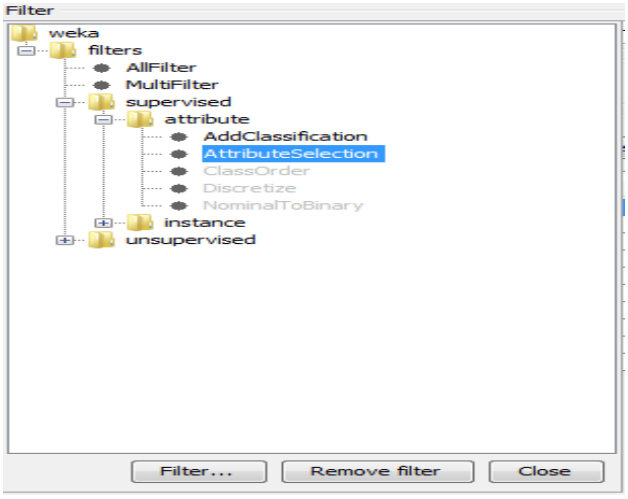

Fig 1. Attribute selection algorithm for data filtering

The attribute selection method reduced the input set and produces four highly influential set of parameters as unexplained, sperm concentration, sperm motility and male factor only (See Fig 2.).

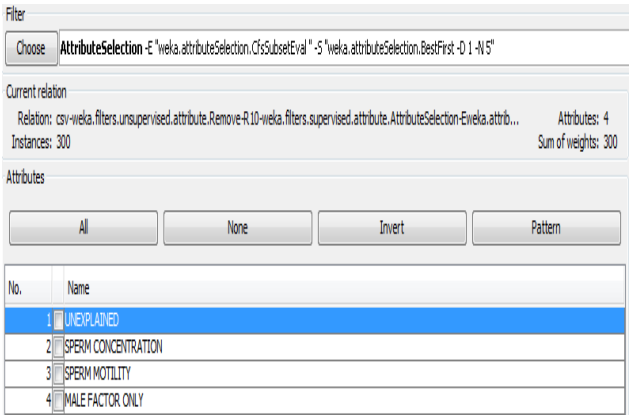

Fig 2. Attribute selection result

These attributes will be the input set to multilayer perceptron network, a type of Artificial Neural Network (ANN), for classification and to find true or false pulse rate and error rate. In this classifier, the multilayer perceptron network architecture is used to train and to determine a mean value of error rate and confusion matrix. The classification process through multilayer perceptron in WEGA tool is as shown in Fig. 3.

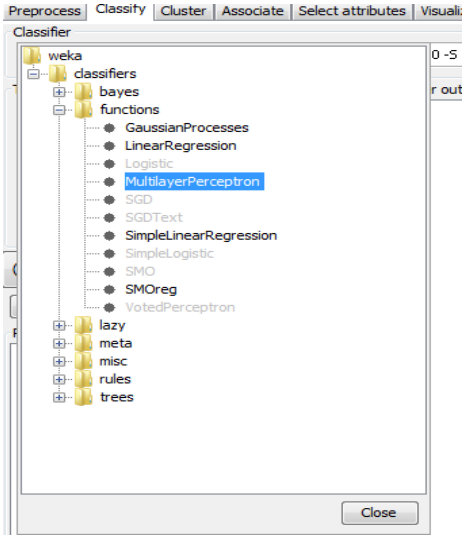

Fig 3.Selecting multilayer perceptron function for classification

The multilayer perceptron neural network is constructed and generated the confusion matrix. In this process of network construction, the valuable errors,YES or NO value, true and false rate are generated, as shown in Fig. 4. The construction network is as shown in Fig. 5.

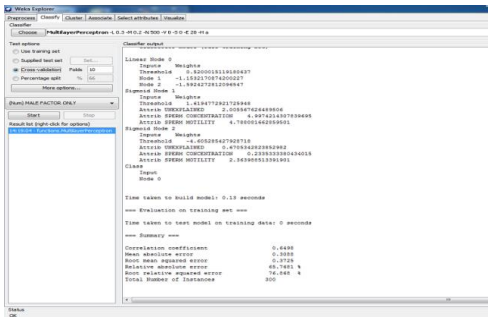

Fig 4. Multilayer perceptron and error rate

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

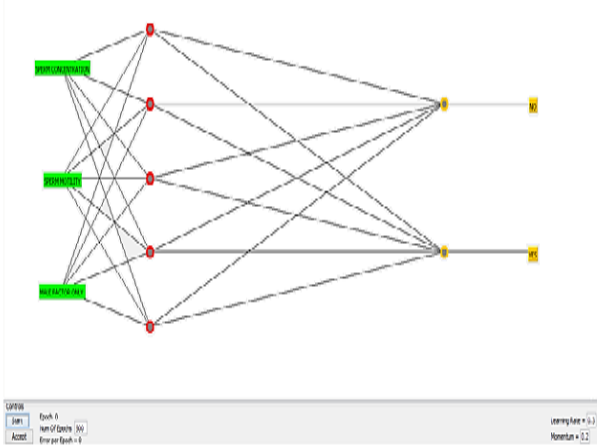

Fig 5. Network construction for multilayer perceptron

The cross validation value of trained network and error rate are explained in this section. The formula of the TP (true pulse) and FP (false pulse) error rate is shown in eqn.(5).

$$X = TP' + E \qquad (5)$$

Where X is the data, T is the component scores and P is the component loadings. E is the residual or error matrix. Because ASCA (Advanced Single Channel Analyzer) models the variation partitions by SCA (Single Channel Analyzer), the model for affect estimates looks like this:

$$A = T_a P_a' + E_a$$
$$B = T_b P_b' + E_b$$
$$AB = T_{ab} P_{ab}' + E_{ab}$$

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.973	0.439	0.798	0.973	0.877	0.619	0.864	0.885	NO
0.561	0.027	0.920	0.561	0.697	0.619	0.864	0.770	YES
Weighted average								
0.825	0.291	0.842	0.825	0.812	0.619	0.864	0.844	

Table 3. Error rate of parameters taken for prediction

TP-true pulse, FP-false pulse	
Correlation coefficient	0.4716
Mean absolute error	0.3576
Root mean squared error	0.4636
Relative absolute error	75.9815 %
Root relative squared error	95.4552 %
Total Number of Instances	300

In table 3, the details of error rate and correlation coefficient generated during the network classification is illustrated. The mean absolute error (MAE) is a common measure of feature selection analysis, where the terms "mean absolute deviation" (MAD) is sometimes used in confusion with the more standard definition of mean absolute deviation. The same confusion exists here generally.

Confusion matrix:

a	b	← classified as
71	2	a = NO
18	23	b = YES

Confusion matrix is a specific layout, which allows visualization of the performance of an algorithm, typically a supervised learning one. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. Based on the error rate and matrix the related graph shows the value influential level and it shows the every individual and grouped attribute for the future enhancement.

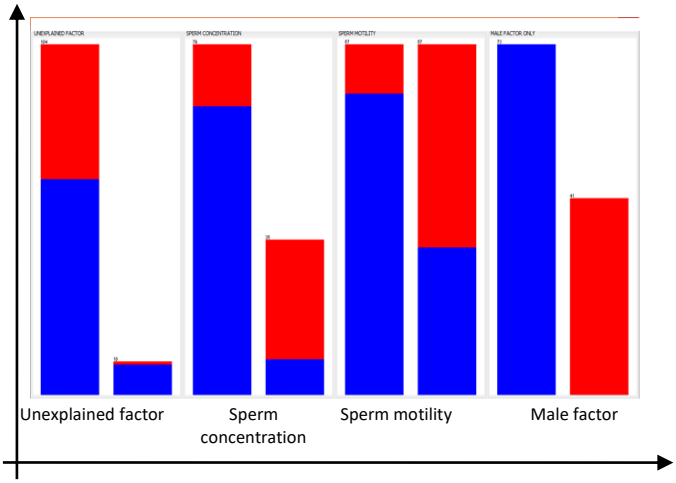

Fig 6. Graph analysis for the TF and FP rate

In the above graph (See Fig. 6), the True rate and False rate of influential attributes are depicted to distinguish. The graph indicates a high level of true pulse and minimum of false pulse rate. The plotter area displays the impact level of True and False rate, highly concentrated area indicates that the respective attributes have more influences on IVF treatments. The plotter matrix of attributes selected and classification process is as shown in Fig. 7.

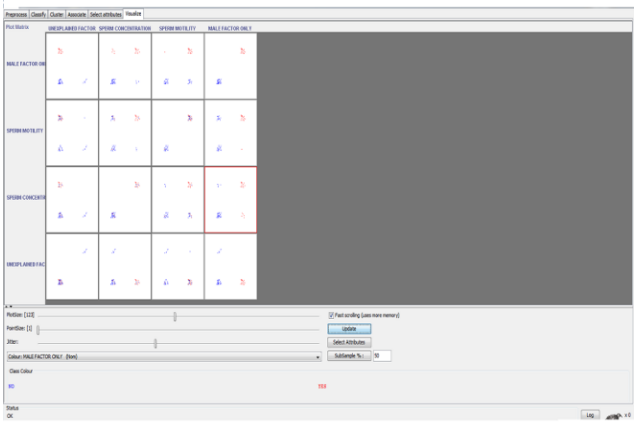

Fig 7. Plotter matrix of attribute classification

These processes of attribute reduction yield the resulted set of attributes known as highly influential attributes on IVF fertility prediction. The repeated application of attribute reduction process in different combination of attributes produces the reduced set of attributes that can enhance the prediction accuracy [8].

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

IV. RESULT AND DISCUSSIONS

The experimental results show that the filter and classifier tool using data mining techniques employed to evaluate and produce the minimum set of data which have most influence on estimating the success rate of IVF treatment. In the experiment, the true and false pulse error, plotter and confusion matrix are described. The correlation coefficient between the attributes to estimate dependency between the attributes and percentage of influences are correctly calculated.

V. CONCLUSION

In this paper, a data mining method of data analysis, classification is proposed for the In-Vitro Fertilization data analysis, and multilayer perceptron network for classification or prediction. From the experiments, the observation is made in the attribute selection analysis and it helps to find the most influential IVF parameters to predict the successful rate of IVF treatment. The proposed technique is useful for finding the minimum set of influential parameters in order to predict a success rate of IVF, which enable the Gynecologists to prescribe the treatment to the couples. By knowing the success rate prior to the treatment, the couples get psychological boost, which increases their chances of getting successful pregnancy.

ACKNOWLEDGEMENT

The authors would like to thank Janani Fertility Centre, Tiruchirappalli, Janani Fertility Centre, Trichy, Devi Hospital, Perambalur, Sri Ramakrishna Hospital, Coimbatore, Ishwarya Fertility Centre, Salem, R.J. Clinic, Coimbatore, Smile Hospital, Mettupalayam, Abi Polyclinic & Fertility Centre, Salem, for providing valuable comments and making available the data. This research work has been funded by the University Grants Commission, New Delhi.

REFERENCES

[1] J.Bazan, A.Skowron, P.Synak, "Dynamic reducts as a tool for extracting laws from decision tables, Proc.Symp.on Methodologies for intelligent systems," Charlotte, USA, 1994, 346-355.
[2] M. Durairaj, K. Meena and S. Selvaraju, "Applying a data mining approach of rough sets on spermatological data analysis as predictors of in-vitro fertility of bull semen", International Journal of Computer Mathematical Sciences and Applications, Serials Publications, ISSN: 0973-6786, Vol. 2(3), pp. 221-231, Dec 2008.
[3] M. Durairaj and K. Meena, "Application of Artificial Neural Network for Predicting Fertilization Potential of Frozen Spermatozoa of Cattle and Buffalo", International Journal of Computer Science and System Analysis, Serials Publications, Vol. 2, No. 1, Jan-Jun 2008, pp. 1-10.
[4] Kaufmann, S.J., Eastaugh, J.L., Snowden, S., Smye, S.W. and Sharma, V. Theapplication of neural networks in predicting the outcome of in-vitro fertilization. Human Reproduction, (1997) vol.12 no. 7 pp. 1544-1457.
[5] Larsson, B. and Rodriguez-Martinez, H. Can we use in vitro fertilization tests to predictsemen fertility? Anim. Reprod. Sci. (2000) 60-61: 327-336.
[6] Thangavel, K, Jaganathan, P, Pethalakshmi, A and Karnan, M. "Effective Classification with Improved Quick Reduct for Medical Database Using Rough System", *BIME Journal*, Vol. 05, Issue (1), pp. 7-14, 2005.
[7] Guoqiang Peter Zhang (2000) "Neural Networks for Classification: A Survey" IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, pp.451-462.
[8] K.Srinivas, G.RaghavendraRao, A.Govardhan (2012) "Analysis of Attribute Association in Heart Disease Using Data Mining Techniques" International Journal of Engineering Research and Applications (IJERA) pp.1680-1683.
[9] S.J.Kaufmann, J.L.Eastaugh, S.Snowden, S.W.Smye and V.Sharma (1997)"The application of Neural Networks in predicting the outcome of in-vitro fertilization "Human Reproduction vol.12 no. 7 pp.1454–1457.
[10] Kay Elder, & Brian Dale., 2000, "In- Vitro Fertilization", Second Edition, United Kingdom at the University Press, Cambridge.
[11] Edwards, R. G. (2001) "The Bumpy Road to Human In-Vitro Fertilization. Nature Medicine" 7:1091-1094.
[12] http://www.medaccessindia.com/IVF-pregnancy-success-rates.html
[13] http://www.ehow.com/about_4760556_advantages-vitro-fertilization .html
[14] https://www.centerforhumanreprod.com/ivf-success-rates.html
[15] http://www.drmalpani.com/book/chapter25h.html

He is currently working as a Assistant Professor, Dept. of Computer Science, Engineering &Technology, Bharathidasan University, Trichy, Tamilnadu, India. He completed his Ph.D. in Computer Science as a full time research scholar at Bharathidasan University on April, 2011. Prior to that, he received master degree (M.C.A.) in 1997 and bachelor degree (B.Sc. in Computer Science) in 1993 from Bharathidasan University. Prior to this assignment of Assistant Professor in Computer Science at Bharathidasan University, he was working as a Computer Programmer at National Research Centre on Rapeseed-Mustard (Indian Council of Agricultural Research), Rajasthan, India, and as a Technical Officer (Computer Science) at the National Institute of Animal Nutrition and Physiology (ICAR), Bangalore for 12 years. He has published 20 research papers in national and international journals.

He received the Master Degree (M.C.A) in 2012 from Anna University, Chennai and Bachelor degree (B.C.A) in 2009 from Periyar University, Salem. He is currently pursuing as a M.Phil Research Scholar in the Department of Computer Science, Engineering and Technology at the Bharathidasan University. His areas of interest are Artificial Neural Network, Rough Set Theory and Data Mining.