ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 2, April 2017
An International Conference on Recent Trends in IT Innovations - Tec'afe 2017
Organized by
Dept. of Computer Science, Garden City University, Bangalore-560049, India

A Thorough Study on Machine Learning algorithms for Improvised Health Concern

Rakshitha Kiran P, Dr. Naveen N C

Assistant Professor, Department of MCA, Dayananda Sagar College of Engineering, Bangalore, India

Professor, Department of CSE, JSSATE, Bangalore, India

ABSTRACT: Machine Learning is a technique of data analysis which automates building up of an analytical model. The machine learning techniques allows computers to get the hidden insights without being explicitly programmed. The data acquisition is through iterative algorithms. The main aim of machine learning is to develop algorithms which can predict on large data. With the need to increase the awareness towards healthcare, the healthcare industry is playing a very prominent role. Keeping in view of the patients care, Various machine learning algorithms are being introduced. In this paper a study has been made on different machine learning algorithms used in addressing various health issues like Breast Cancer, asthma, tuberculosis and diabetes. Also the paper shows that there is a need for research to help the women fraternity to overcome the infertility problems that they are facing.

KEYWORDS*:* Machine Learning;Analytical Model; fertility;

I. INTRODUCTION TO HEALTHCARE INFORMATICS

Health care informatics is a very broad area which involves the achieving the data of the patient under test and performing various techniques on the data. Patient data will vary in their quality. We obtain data which are clinically validated i.e., data which are recorded by medical transcriptionist, while other data are patient-owned, having been obtained by the patients itself. Complexity arises in the range of patient data, which varies from high to low precisions. The low precision data is the one acquired by manually observing the patient. Clinical data of the patients can be derived from many sources like genomic data, data collected during consultation, data from medical tests etc. Therefore the field of healthcare informatics generates large data set and these data sets can be useful for clinical research. Machine learning techniques can be applied to this clinical data set to obtain desired results.

II. MACHINE LEARNING

In last few decades there has been extreme change in the way how the data is stored, retrieved, analyzed and processed. Very large amount of data is generated each and every second. This data is stored, used and analyzed efficiently so that it can show important insights. A lot of machine learning techniques have been evolved in analyzing the large amount of data. Machine Learning (ML) directs at making available the computational methods for collecting, formatting and upgrading information in intelligent systems. Machine learning mechanisms help us to bring out information from the available data through the algorithms. ML algorithms were designed with the main purpose to analyze medical data sets. Today ML provides several necessary tools for smart data analysis. The revolution digital domain gave a relatively cost effective means to process data. The Modern hospitals are assembled with monitoring patient dataset and also data collection devices. ML technology is right now well suited in understanding and analyzing medical data. Machine learning techniques are helpful in cases where algorithmic answers are unavailable or in the case where there is shortage of formal models and knowledge about the application domain.

III. APPLICATION OF MACHINE LEARNING IN HEALTHCARE

In the paper [1] data mining techniques were applied to predict the frequency of the recurring breast cancer cell. This was done by analyzing data collected from ICBC registry. They used DT (Decision Tree), SVM (Support Vector

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 2, April 2017
An International Conference on Recent Trends in IT Innovations - Tec'afe 2017
Organized by
Dept. of Computer Science, Garden City University, Bangalore-560049, India

Machines) and ANN (Artificial Neural Network) ML algorithms were used to find about the in breast cancer causing cells. Also these algorithms were compared to discover which method performs the best. Here results show that SVM algorithm outperforms both DT and MLP in all the parameters like specificity, sensitivity and accuracy. Support Vector Machines (SVM) is the best predictor algorithm for breast cancer recurrence. The table 1.1 shows the various parameters used by the author to predict the recurrence of breast cancer cells.

In the paper [2] author talks about different techniques of Machine Learning were used to predict the results of TB treatment course. The completion of the TB course is a very crucial stage to assure that the patients will not undergo protracted, lengthened, relapse, infectiousness and more expensive therapy because of multidrug resistance TB. At least 50% of the patients with TB do not complete the treatment course. To solve TB treatment problem, with the help patient and global organization called "global plan to stop TB" considered by WHO planned a model to predict the result of DOTS therapy

Table 1.1: Parameters for predicting the occurrence of breast cancer[1]

No	Variable Name	Definition
1.	Local Recurrence	Yes or No
2.	Age at Diagnosis	≤ 35, 35 to 44, 44-45, 55 ≥ years old
3.	Age at Menarche	≤ 12 to ≥ 12 years old
4.	Age at Menopause	≤ 50 to ≥ 50 years old
5.	Infertility	Yes or No
6.	Family History of Breast Cancer	Yes or No
7.	History of other Cancer (CA)	Yes or No
8.	Location	Upper outer Quadrant (UOQ), Upper inner Quadrant (UIQ), Lower outer Quadrant (LOQ), Lower inner Quadrant (LIQ), Central, Axilla, Upper half, Lateral half, Lower half
9.	Side	Left, Right, Bilateral
10.	Tumor Size	≤ 2cm to ≥ 5cm
11.	LN/Nexion	Lymph node involvement/number of removed nodes after surgery
12.	Metastasis	Bone, Liver, Lung, Brain, others
13.	NPositive	Number of Positive lymph node involvement
14.	B.Pathology	Results of Biopsy Pathology after
15.	Type of surgery	Mastectomy (Preservative or Bilateral)
16.	G (Grade)	1, 2 or 3
17.	Margin of Involvement	Free or ≥ 2cm
18.	Estrogen Receptor	Negative or Positive
19.	Progesterone Receptor	Negative or Positive
20.	Type of Chemotherapy	Adjuvant or Neoadjuvant
21.	Radiotherapy	Yes or No
22.	Hormone Therapy	Tamoxifen, Raloxifen, Femara, Aromasin or Megace
23.	Death	Realted to Breast Cancer or unrelated
24.	Her2	Negative or Positive

Machine learning techniques were used to predict the results of TB therapy. For the analysis models by six machine learning algorithms were developed and validated. The algorithms included ANN (Artificial Neural Network), BN (Bayesian network), DT (decision tree), LR (logistic regression), RBF (radial basis function), and SVM (support vector machine). Figure 1.1 shows schematic representation of applied methodology of validation process and model development.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 2, April 2017
An International Conference on Recent Trends in IT Innovations - Tec'afe 2017
Organized by
Dept. of Computer Science, Garden City University, Bangalore-560049, India

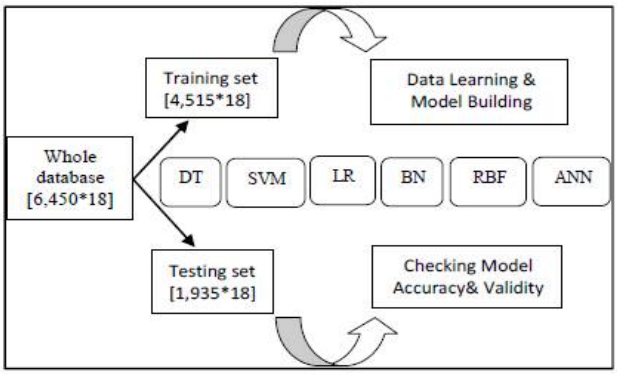

Fig1.1 Schematic Representation of Applied Methodology of Validation Process And Model Development.[2]

It was found that Decision tree DT (C 4.5) was best algorithm with %74.21 prediction accuracy when compared with other algorithms with prediction accuracy as follows: ANN(62.06%), BN(57.88%), LR(57.31%), RBF(53.74%), and SVM(51.36%). The table 1.2 shows various parameters considered for the prediction algorithm

Table 1.2 parameters considered for the prediction algorithm[2]

Variable	Variable Definition	Categories of Values
	Demographic Characteristics	
sex	Gender of TB case	Male (1)/Female (2)
Age	Age of TB patient	(Continuous var.) 0.05 - 99
Weight	Weight of TB patient	(Continuous var.) 4 - 110
Nationality	Nationality of TB patient	Iranian (1), Central Asians (2),
Area	Area of residence	Abroad (1), Mobile (2), Rural
Prison	Current stay in prison	No (1) /Yes (2)
	Clinical Features	
Case type	Type of TB that patient is belonged to	New (1), imported (2), cure aft
Treat cat	Category of treatment that is conducting	A (1)/B (2)
TB type	The part of body that has been affected	Pulmonary (1)/extra-pulmonar
RTBinf	Whether patient has recently TB affected	No (1)/yes (2)
Diabetes	Whether TB patient isaffected by diabete	No (1)/yes (2)
HIV	Whether TB patient has been known as HIV+	No (1)/suspected (2)/yes (3)
Length	Length (month) of being affected by TB	(Continuous var.) 0.03 - 90.77
LBW	Low Body Weight	No (1)/yes (2)
	Social Risk Factors	
Imprisonment	The history of living in prison	No (1)/suspected (2)/yes (3)
IV drug using	Whether patient is using the (IV) drugs	No (1)/suspected (2)/yes (3)
Risky sex	Whether patient has history of risky sex	No (1)/suspected (2)/yes (3)

In paper [3] the author applies machine learning techniques for diagnosing Heart disease in Diabetic patients. Heart disease is one of the leading reasons for death over the past few years. A lot of research has been done by using several ML techniques in the diagnosis for heart disease. Diabetes is a chronic condition that occurs when pancreas cannot produce enough amount of insulin, or when body may not use the insulin produced effectively. Few systems have employed ML methods such as Support Vector Machines and Naïve Bayes algorithms for the purpose of classification. A support vector machine is a modern technique used in different fields of healthcare application. For diabetics' diagnosis, the model showed a positive accuracy and predicts chances of a diabetic patient to suffer from heart disease. The table 1.3 below shows the various attributes considered for the prediction. The data set of 500 diabetic patients was used for experimentation. Out of 500 diabetic patients, 142 records are of those patients who are having heart disease (positive cases) and other 358 records are not having any heart disease (negative cases). These records were pre-processed and are given as input to for SVM classifier.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 2, April 2017
An International Conference on Recent Trends in IT Innovations - Tec'afe 2017
Organized by
Dept. of Computer Science, Garden City University, Bangalore-560049, India

The results showed that the classifier reveals high classification accuracy of 94.60% overall. A very high precision for the positive class 97.52% is obtained with high recall of 83.10% and negative class's classifier exhibits 93.67% precision and high recall of 99.10%. In this paper SVM classifier is used to detect the early detection of problems of a diabetic patient to get heart disease. SVM classifier is the best classification method for diabetic dataset.

In paper [4] the authors tells of the seminal quality prediction using ML Clustering-Based Decision Forests (CBDF). The algorithm is useful for diagnosis of seminal patients or selection of semen donors. Here in this paper a supervised ensemble learning method called Clustering-Based Decision Forests. This algorithm is used to predict quality of semen. During the prediction top five parameters were considered like serious trauma, age, sitting time, high fevers in the last year and the season when the semen sample is produced. Experiments were conducted using different machine learning algorithms like Support Vector Machines, random forests, decision tree, logistic regression and multilayer perceptron neural networks.

The dataset for fertility includes labeled dataset of 100 anonymous young healthy people between the age of 18 and 36 years old. Each specimen in the dataset has nine normalized data variables about life, health status habit and semen quality (concentration). The result labeled as "altered" and "normal". Fertility dataset was with a class ratio of 7:1 and also there are 88 normal & 12 altered specimens in the whole dataset Figure below parameter most important ones in affecting the seminal quality.

Attribute	Description
Sex	A classification of the sex of the person
Age	Age of the patient
Family Heredity	Previous history (Father / Mother)
Weight	Patient's weight
BP	Blood pressure
Fasting	Sugar level after fasting
PP	Post Prandial blood glucose level
A1C	HbA1c level Glycosylated Last 4 months sugar level
LP Tot Cholesterol	Total cholesterol level

Attribute Role	Attribute Name	Attribute Type	Description
Regular	Sex	binominal	Sex of the patient. Takes the following values: Male, Female
Regular	Age	integer	Age of the patient
Regular	Fam/Heri	polynomial	Indicates whether the patient's parents were affected by diabetes. Takes the following values: Father, Mother, Both
Regular	Weight	numeric	Weight of the patient
Regular	BP	polynomial	Blood Pressure of the patient
Regular	Fasting	integer	Fasting Blood Sugar
Regular	PP	integer	Post Prondial Blood Glucose
Regular	A1C	numeric	Glycosylated Hemoglobin Test
Regular	LDL	integer	Low Density Lipoprotein
Regular	VLDL	integer	Very Low Density Lipoprotein
Label	Vulnerability	nominal	Indicates the vulnerability of the patients to heart disease. Takes the following values: High, Low

Table 1.3 attributes considered for the prediction of heart disease

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 2, April 2017
An International Conference on Recent Trends in IT Innovations - Tec'afe 2017
Organized by
Dept. of Computer Science, Garden City University, Bangalore-560049, India

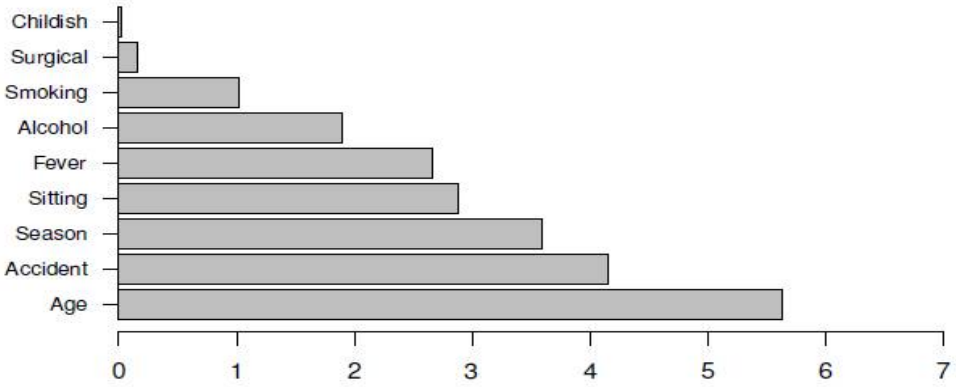

Fig1.2: variables affecting semen concentration

In this research paper comparison of the performance of Clustering-Based Decision Forests (CBDF) with other five popular machine learning algorithms like RF, SVM, CART, MLP and LR were done and also evaluated those machine learning algorithms in predicting Semen concentration results CART, RF, SVM, MLP and LR algorithms were highly precise classifiers and also find a lot of great applications in various fields. The experiment results by CBDF on Fertility Dataset have showed better results. The studies have propounded that age, sexual abstinence, smoking, food intake, obesity, psychosocial stress are also associated with quality of semen . Semen accident or serious trauma will highly effect the quantity of semen.

In the paper [5] the author shows the various possibilities of using ANN algorithm (Artificial Neural Networks) to reduce the amount of time and simulation scores to obtain a better result. The procedure of prediction of infertility matching is time-consuming, costly and difficult. This paper provides a gateway for infertility matching by using a trained neural simulation model and ANN. The simulator provides the necessary training and testing data to train and validate network. Then the resulting network is employed to predict with help of data set. The figure1.3 shows the dataset considered for the prediction.

This study provides the guidelines for the development of Artificial Neural Networks for infertility matching.

1.	*Data set: INFERT: has data of a case-control study that investigated infertility after spontaneous and induced abortion.*
2	*Data set Details:* *(a) 248 observations* *(b) 83 women, who were infertile (cases)* *(c) 165 women, who were not infertile (controls).* *(d) Induced - number of prior induced abortions.* *(e) Spontaneous - spontaneous abortions. (Both variables take possible values 0, 1, and 2 relating to 0, 1, and 2 or more prior Abortions).* *(f) Age - in years* *(g) Parity: number of births*
3	*Training of neural networks - The function neural net used for training a neural network provides the opportunity to define the required number of hidden layers and hidden neurons according to the needed complexity*

Fig 1.3: dataset considered for the prediction.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 2, April 2017
An International Conference on Recent Trends in IT Innovations - Tec'afe 2017
Organized by
Dept. of Computer Science, Garden City University, Bangalore-560049, India

IV. MACHINE LEARNING FOR WOMEN INFERTILITY

Infertility in India has become one of the common problems which both men and women are facing. According to the survey [6] there has been a steep rise in infertility problem in India. The research shows tells that 20%-30% rise in infertility in last 5 years. In the article published by Dr. Sharmistha Dey [7] tells that out of every 100 couples around 40% of male suffers from infertility and 50% of women suffers from infertility and rest 10% infertility is because of both of them. From the above study it is clear that there is lot of scope of research in the field of infertility in women. Women infertility is one of the major problems in today's world. There are lots of reasons for women infertility like age, change in life style, obesity, smoking, food habits and so on. Machine learning technique can be applied on predicting women infertility and chances of preventing the problem. The table below shows the various machine learning algorithms applied over various diseases. Different machine learning techniques shows their application over variety of disease. From the observation of the table it is clear that there is a scope for research on women infertility problem using machine learning techniques.

Disease / ML techniques	Breast cancer	TB	Heart disease in Diabetic patients	seminal quality prediction	Women infertility	Diabetes	Astana
Decision Tree (DT),	✓	✓		✓			✓
Support Vector Machines (SVM)	✓		✓	✓		✓	
Artificial Neural Network *(ANN)*	✓	✓			✓		
logistic regression (LR)		✓		✓			✓
radial basis function (RBF)		✓					
Bayesian networks (BN)		✓					
Naïve Bayes (NB)			✓				
Clustering-Based Decision Forests (CBDF)							
Random Forests (RF0				✓			✓
Multilayer Perceptron Neural Networks (MPNM)				✓			
AdaBoost							✓
Gaussian Process (GP)						✓	✓

Table 1.4: Machine learning techniques applied on various diseases

V. CONCLUSION

In this paper it is observed that how machine learning techniques are applied on various diseases to predict and to overcome the diseases. Machine learning techniques are very useful in healthcare. Women infertility is one the major problem faced by lot of people in this generation. Machine learning and to predict to overcome infertility among ladies has a lot of scope for research

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 2, April 2017
An International Conference on Recent Trends in IT Innovations - Tec'afe 2017
Organized by
Dept. of Computer Science, Garden City University, Bangalore-560049, India

REFERENCES

[1] Ahmad LG*, Eshlaghy, AT, Poorebrahimi A, Ebrahimi M and Razavi AR, "Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence" published in Journal of Health & Medical Informatics, ISSN: 2157-7420 Published April 24, 2013

[2] Sharareh R. Niakan Kalhori1,2*, Xiao-Jun Zeng, "Evaluation and Comparison of Different Machine Learning Methods to Predict Outcome of Tuberculosis Treatment Course" Journal of Intelligent Learning Systems and Applications, 2013, 5, 184-193 .

[3] G. Parthiban S.K.Srivatsa" Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients" Inter national Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 3– No.7, August 2012 – www.ijais.org

[4] Hong Wang, Qingsong Xu and Lifeng Zhou, article "Seminal Quality Prediction Using Clustering-Based Decision Forests" ISSN 1999-4893 Algorithms 2014, 7, 405-417; doi:10.3390/a7030405

[5] Prof. A.K. Soni1, Abdullahi Uwaisu Muhammad, "TRAINING NEURAL NETWORKS WITH INFERTILITY PARAMETERS" International Journal of Science, Technology & Management www.ijstm.com Volume No.04, Issue No. 04, April 2015

[6] http://www.dailyo.in/lifestyle/infertility-on-the-rise-in-women-men-lifestyle-stds-sexual-health-sperm-count-vd-menstrual-cycles/story/1/8839.html

[7] http://news.xinhuanet.com/english2010/world/2010-07/16/c_111963155.htm