Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 5 (2023)

# An Explorative Study on Medicaments of Reproductive Health using Data Analytics

[1] Suvika K V, [2] Dr. D G Jyothi

[1]Research Scholar, Dept. of CSE, Bangalore Institute of Technology, kvsuvika@gmail.com
[2] Professor and HOD, Dept. of AIML, Bangalore Institute of Technology, jyothi.bitcse@gmail.com

**Abstract—** The variety of ailments and individual concern are causing the medical data to rise dramatically every day. The examination of these data is absolutely vital for society's advancement. Data Analytics is extensively investigated for its wide applications in various fields. Medical data analytics is considered as the outcome of applying data analytics themesto the real world of medical and clinical services. It is a interdisciplinary and dynamic area of research. We are concernedin reproductive health in this paper. 186 million individuals and 48 million couples worldwide struggle with issues associated to reproductivity, according to the World Health Organization [1]. We have examined the several publications that are complexin this area in aspects of algorithms, procedure, and outcomes. This research paper's purpose is to explore the fundamental benchmarks for reproductive health and to present the findings and outcomes of previous studies that have used data analytics and other techniques to analyse.

**Index Terms—**Medical Data, Data Analytics, Reproductive Health

## 1. Introduction

Reproductive health encompasses a person's overall mental,physical, and social well-being as it relates to their repro- ductive system, behaviours, and processes, and goes beyond simply the absence of sickness or infirmity. One must becapable to reproduce, have the option to decide whether, when,and how often to do so, and be able to have a meaningful and safe sexual experience in order to be in good reproductive health.

Sexually transmitted infections, malignancies of the repro- ductive tract, HIV and AIDS, abortion, and childbirth-related deaths and disabilities are all examples of sexual and reproduc-tive ill health. Reproductive health issues account for at least 20 percent of the overall health problems faced by womenof childbearing age, compared to 14 percent for men. The disparity in sexual and reproductive health between developedcountries and low-income, as well as between wealthy and impoverished individuals inside countries, is most pronounced when evaluating progress using indicators related to sexualand reproductive health. Many countries with high burdens of disease currently fall far short of providing appropriate health services. By better integrating services for Human Immun- odeficiency Virus Infection and Acquired Immune DeficiencySyndrome and reproductive and sexual health, such that they are complimentary rather than competitive, resources could beused more effectively.

Products related to reproductive health are in greater de- mand with awareness and population growth. Wider develop- ment is under danger due to poor sexual and reproductive health as well as the enormous unmet demand for familyplanning, particularly in Africa, where UN forecasts show a population surge from 794 million in 2000 to 2,000 millionin 2050 [2].

Healthcare providers' ability to deliver high-quality careeffectively and efficiently has already been greatly influenced by the integration of data analytics in the field. Despite this, the significance of data analytics in healthcare systems and enhancing patient outcomes continues to expand and increase as more data sources become accessible and new technologiesemerge that simplify the interpretation and use of analytics results for healthcare professionals.

Realizing the promise of data analytics to alter the health- care sector requires an understanding of how the technology may be used to address difficulties encountered by healthcare providers, such as staff utilization and recruitment, opera- tional economies, and improved patient experiences. To deliverpatient-centered care, it is essential to comprehend the wants and needs of patients. This vital information can be uncoveredthrough the use of data analytics.

## 2. Big Data Architecture For Healthcare

Research And Studies On The Use Of Big Data In Other Sectors suggest that a healthcare big data application system should have a four-tier architecture: data collection, data storage, data analysis, and data interchange and sharing.

### A. Data Collection

The goal of data collection is to gather data generated by health and medical organizations, primarily from healthcare institutions, public health organizations, healthcare insurance providers, resident health records, population statistics, and electronic medical records. The collection of big data in healthcare is typically divided into distributed and centralized collection. Ensuring the acquisition of high-quality data that meets requirements and performing data collection, cleaning,conversion, and loading are of utmost importance.

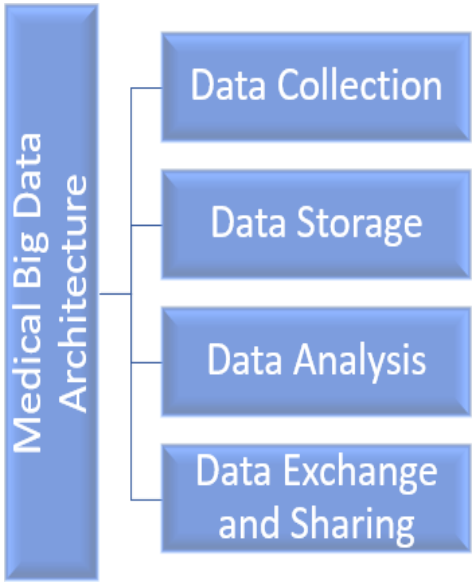


**Fig. 1.** Medical Big Data Architecture

### B. Data Storage

Big data technology depends heavily on data storage. Big data in medicine often ranges from TB to PB in size. Struc- tured, unstructured, and semi-structured storage are all partof the storage pattern. Relational databases (RDBMS) are commonly used to store structured data; examples encompass SQL Server, ORACLE, MySQL, and PostgreSQL.NoSQL technology is required for unstructured storage. Hadoop has the capability to implement semi-structured storage. PACSdata, Electronic medical records and follow-up data are typ- ically present in the form of texts or images in the medical profession.

### C. Data Analysis

The backbone of big data technology is data analysis methodology, which encompasses three main techniques: (1) traditional analytics based on regression analysis, multidi- mensional analysis, feature analysis, association rules, andclassification and clustering algorithms, (2) intelligent analysisutilizing machine learning, NLP, data mining, and semantic search, and (3) user-defined analysis.

Given the large volume of data and the limitations of out- dated analysis methods, the intelligent examination approachis currently the most widely used data analysis technique inthe industry. Additionally, advancements in the medical sector include more sophisticated natural language processing of electronic healthcare records and semantic analysis techniquesfor PACS images.

### D. Data Exchange and Sharing

In addition to facilitating data sharing integration, central- izing data collection, sorting, and

dissemination, data sharing and exchange should enable a distributed SOA architecture and wider data exchange through document, Web Service,database, and other methods. This enables tight integration within organizations and loose integration between services, supports various interface and standard specifications, and enables integration, data exchange, and sharing of key essentialapplication systems and service platforms.

## 3. Data Analytics Techniques And Tools

In general, data size increases daily. In all the shifting domains of innovation, industry, and research, there is now a greater requirement to observe comprehensive, complicated, data-improved informational collections. In today's competi- tive world, the ability to extract important knowledge from these vast quantities of data and to follow up on that knowl- edge is increasingly crucial. Data analytics is the process of using a computer-based data framework, including innovative methodologies, to extract knowledge from data. Predictive models and descriptive models make up the majority of data analytics models. To forecast unknown or upcoming estimatesof many elements, the predictive models frequently use su- pervised learning functions. Conversely, descriptive models frequently use unsupervised learning techniques to identify patterns that describe the data.

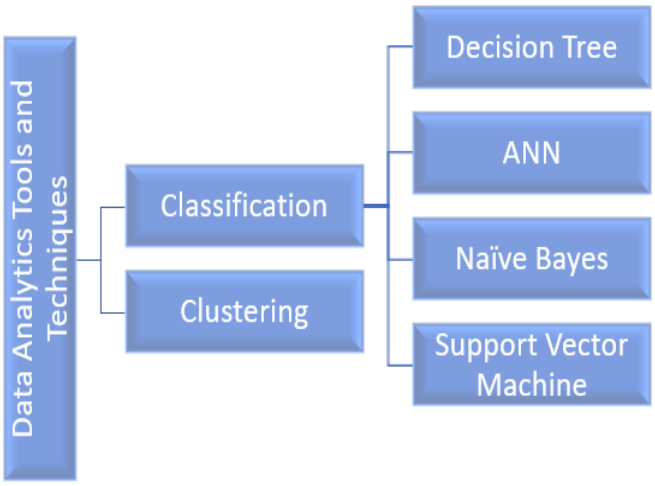


**Fig. 2.** Different Classification Tools and Techniques

### A. Classification

Finding a class model for something through classification involves looking at its attributes. In data when the classes are preset, classification is a supervised learning technique that anticipates each specimen's target class. Prediction is what werefer to when we employ a categorization model to foresee or foretell unknowable things. Classification trees are a crucial prediction technique that look for a reliable relationship amonga group of qualities. A classification model may potentially predict the values of unordered attributes.

*1) Decision Tree:* A data analytics method suited for per- forming classification and prediction is the decision tree. The decision tree may iteratively provide results based on several variables, making it useful for assessing the characteristics, parallels, and differences in data. Classification trees, as men- tioned, are a coordinated strategy used to project the potential value of an objective attribute. Through the tree's traversal, the decision tree could generate rules. Decision trees include:

1) Random tree 2) J48 3) Reptree and 4) Random forest

2) *ANN:* An Artificial Neural Network (ANN) is a compu- tational model that simulates the operation of the human brain and is built on the elements of biological neural networks. ANNs use a large number of interconnected artificial neurons to predict new information based on existing information, either by using specific features or the same attributes. The ANN is both hardware and software-based and has been widely used by scholars for data analysis and prediction purposes.

3) *Naive Bayes:* The probabilistic learning method Naives Bayes calculates exact probability for each hypothesis. It makes probabilistic predictions using Bayes Theorem as afoundation. If the hypothesis is correct, each training record can incrementally raise or lower the likelihood. An element's presence in a class does not

always imply the inclusion of otheritems, according to the Naive Bayes classifier. For really big datasets, this strategy is highly advantageous.

4) *Support vector Machine:* A supervised learning tech- nique called support vector machines is based on the struc- tural threat minimization standard and the statistical learning hypothesis. SVM definitely converts the original input space into a high dimensional feature space using the training data. In this method, the margins of class borders are maximised to find the optimal hyper plane. Support vectors are the training points that are closest to the ideal hyperplane. The decision facet can be utilised for additional classification once it has been collected.

### B. Clustering

Unsupervised learning relies on a few clustering techniques,such as grid-based algorithms, partitioning methods, density- based methods, and various hierarchical clustering algorithms [8]. Clustering techniques use algorithms to form meaningful groups of objects with similar properties. The clusters are cre- ated by selecting the cluster center as the centroid, minimizingthe sum of the squared distances between each cluster point and its center.

## 4. Data Analytics Application In Reproductive Health

D K Girija and M S Shashidhara [3] focused on fibroids, tumours that have an impact on reproductive health. The wombwall has fibroids, which range in size from small to large. Age,ethnic origin, eating habits, family history and heaviness were all taken into consideration as the fundamental building blocksfor constructing a decision tree by applying the J48 algorithm.To classify the data, decision tree was created using the WEKA tool.

M Durairaj and Ramasamy Nandhakumar [4] studied treat- ment for problems in reproductive health of people. Thereare many different kinds of conception therapies. However, due to its high success rate, IVF (In-Vitro Fertilization) is the most widely used method. In this study the attribute selection algorithm was used to select the main feature subsets which will result in better success rate by using Weka's selection tool. Later, supervised filter is used to remove redundant information, and the input is then sent to a multilayer per- ceptron network for error and pulse rate checks. Dependencyis calculated among numerous attributes.

Girela JL and et al [5] conducted research and experiments on semen and reproductive health. In this study, environmentalfactors and lifestyle choices were taken into account in addi- tion to semen characteristics. Following the steps below, 100 participants between the ages of 18 and 36 were taken into consideration for the semen study and experiment. Confusion matrix is ultimately created via population study, semen anal- ysis, ANN architecture, and ANN experiment. The findingson the altered sperm parameters were discovered as:

- Reduced Sperm Motility
- Low Sperm Count
- Abnormal Sperm Shape

Idowu, Peter Adebayo, et al [6] took 39 records with 14 attributes and applied a multi-layer perceptron. MLP was usedto create an infertile prediction model. Using the efsSubset Evaluator Method, features were chosen for the provided attributes, and ultimately 6 attributes were chosen since they were given top priority.

Fang Chan, et al [7] researched how to choose an embryo for implantation based on its morphological and developmentaltraits. In this research, a comparison is done between twomethods for choosing a potential embryo: the traditional Standard Scoring System (SSS) and the modern Computer Assisted Scoring System (CASS). To compare and assess,a prediction model was employed. Multivariate adaptive re- gression splines and multivariate logistic regression were usedas prediction models. Between the years of 2008 and 2013, 871 embryos were chosen for the experiment, which thenunderwent evaluation. Improvements in the generalizability ofprediction models were found by the CASS.

Simi M S, et al [10] done prediction using two best algorithms J48 and Random forest algorithm. 26 important variables were considered and 8 variant were identified for early detection of infertility by using Mean Decrease in Accuracy method. The conclusion was drawn that Random forest is good with high accuracy.

Suriya Prabha T., et al [11] proposed a method for de-tection of PCOS by considering hormones as main aspect. The hormonal imbalance can be caused due to many reasons which includes diabetes, hyperthyroidism, hypothyroidism etc.They collected follicular fluid and plasma samples from 100 women whose

age ranges from 23-35 years and divided into PCOS and non-PCOS by considering FSH, LH, Triglycerides,LDL-C, HDL-C etc as other parameters. Raman Spectroscopywas used to study chemical bonding of biological molecules and ML classifiers were used for prediction. The AdaBoost classifier gives the higher accuracy for prediction.

Bo Zhang., et al [12] proposed a method where a patientcan make optimal decisions based on the results. The patient can decide whether to user her own oocyte or donor oocyte, how many pickup cycles she may need, whether to use frozen embryo or not etc. The three different approaches were made namely clustering, SVM and C-SVM for predicting pregnancy rate in multiple cycles of a patient. The C-SVMwas considered best among all.

Sara Alshakrani., et al [13] proposed a hybrid machinelearning model for early detection of Polycystic Ovary Syn- drome Detection. This study also showed that using hybrid models are more effective. The models used are LSVM (LinearSVM Classifier), LGBM (Light Gradient Boosting Machine), XGBRF (Extreme Gradient Boosting and Random Forest),CatBoost. CatBoost and HRFLR showed more accuracy.

## 5. Open Issues And Challenges

Even if there are many effective treatments for reproductivehealth, there are still numerous obstacles to overcome as follows:

1. Medical jargon consists of idioms from Chinese andother languages as well as phrases from both domestic and international medical professions. Medical language and data are difficult in terms of literal semantics and expression dueto issues like uneven and imprecise terminology standards andrapid updating speeds, notably in the discipline of Chinese medicine in China.
2. Medical data is created from multiple dimensions during the hospital care-seeking process, with an emphasis on the pa-tients. For instance, the doctor creates data from the diagnosis and treatment aspect, the medical technician generates data from the inspection and testing side, and the nurse generates data from the nursing aspect. The same medical activity has different forms of data representation because of the various formats and specifications.
3. Medical data will eventually be incomplete, whether it is recorded manually or electronically, due to a variety of reasonsfor missing records or poor data recording.
4. The heterogeneity issue results in data diversity, including different data sources, management methods, and standards. The major data sources are private medical practices, hospitals, regional medical information platforms, new rural cooperative medical insurance, community health organizations, medical insurance, third-party testing companies, networks and indi- vidual users. The variety of management techniques is affectedby factors such as variations in operating systems, databases, and technologies used by different management systems.
5. Medical data contains sensitive patient information like demographics and health status, which are dispersed or con- cealed in various locations. Data analysis and mining, which reveal the confidential nature of medical data, can provide a comprehensive understanding of patient privacy. In additionto individual medical data, privacy information also includesa wealth of details about hospital operations, diagnostic and treatment methods, and drug effectiveness. These complexitiesare delicate and may impact business, resulting in medical organizations being hesitant to exchange data and some data- analysis departments lacking data.

| Year | Author | Method/ Algorithm Used | Remarks |
|---|---|---|---|
| 2012 | Girija D K et al [3] | J48 Algorithm | To find the tumours on fibroids based on various factors |
| 2013 | M Durairaj et al [4] | Multilayer Perceptron network | To calculate the dependency between attributes for IVF treatment |
| 2013 | Jose L Girela et al [5] | ANN architecture | The semen experimentation was done to find various factors w.r.t sperms |
| 2016 | Peter et al [6] | MLP – Predictive model cfsSubset Evaluator Method | To identify the attributes responsible for infertility |
| 2016 | Fang Chen et al [7] | Multivariate Logistic Regression | To select potential embryo for implantation |
| 2017 | Simi M S et al [10] | J48, Random Forest Algorithm | To identify the early detection of infertility |
| 2019 | Bo Zhang [12] | SVM and C-SVM | To take optimal decision by a patient for choosing IVF. |
| 2022 | Suriya Prabha T et al [11] | Raman Spectroscopy, AdaBoost classifier | To identify early detection based on hormones |
| 2022 | Sara Alshakrani et al [13] | LSVM, LGBM, XGBRF, HRFLR | To identify early detection of Polycystic Ovary Syndrome. |

**Fig. 3.** Comparison Table of various types of research done

## 6. Conclusion

In this paper, the therapeutic options currently availablefor male and female reproductive health are studied through various research publications. Additionally, the difficulties encountered with gathering medical data are examined. Futureresearch will present an optimal technique that takes into account both male and female feature.

## References

[1] Boivin J, Bunting L, Collins JA, et al. International estimates of infertility prevalence and treatment-seeking: potential need and demand for infertility medical care. Human reproduction (Oxford, England) 2007;22(6):1506-12. doi: 10.1093/humrep/dem046 [published Online First: 2007/03/23]

[2] https://gsdrc.org/document-library/sexual-and-reproductive-health-and- rights-a-position-paper

[3] D.K Girija. and Shashidhara, M.S, Classification of Women Health Disease (Fibroid) Using Decision Tree Algorithm. International Journal of Computer Applications in Engineering Sciences 2(3):205 – 209,2012

[4] M., Durairaj Ramasamy, Nandhakumar. (2013). Data Mining Applica- tion on IVF Data For The Selection of Influential Parameters on Fertility.

[5] Girela JL, Gil D, Johnsson M, Gomez-Torres MJ, De Juan J. Semen parameters can be predicted from environmental factors and lifestyle using artificial intelligence methods. Biol Reprod. 2013 Apr 18;88(4):99.doi: 10.1095/biolreprod.112.104653. PMID: 23446456.

[6] Idowu, Peter Adebayo, et al. ”Data Mining Approach for Predicting the Likelihood of Infertility in Nigerian Women.” Handbook of Research on Healthcare Administration and Management, edited by Nilmini Wickra- masinghe, IGI Global, 2017, pp. 76-102. https://doi.org/10.4018/978-1- 5225-0920-2.ch006

[7] Chen F, De Neubourg D, Debrock S, Peeraer K, D'Hooghe T, Spiessens
Selecting the embryo with the highest implantation potential using a data mining based prediction model. Reprod Biol Endocrinol. 2016 Mar 3;14:10. doi: 10.1186/s12958-016-0145-1. PMID: 26936606; PMCID: PMC4776393.

[8] Simi, M. S. and K. Sankara Nayaki. "Data analytics in medical data
: A review." 2017 International Conference on Circuit ,Power and Computing Technologies (ICCPCT) (2017): 1-4.

[9] Das, Nivedita and Rautaray, Siddharth and Pandey, Manjusha. (2018). Big Data Analytics for Medical Applications. Interna- tional Journal of Modern Education and Computer Science. 10. 10.5815/ijmecs.2018.02.04.

[10] Simi, M.S., Nayaki, K.S., Parameswaran, M., Sivadasan, S. (2017). Exploring female infertility using predictive analytic. 2017 IEEE Global Humanitarian Technology Conference (GHTC), 1-6.

[11] S. P. T, R. S and E. R, ”Early Diagnosis of Poly Cystic Ovary Syndrome (PCOS) in young women: A Machine Learning Approach,” 2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), Singapore, Singapore, 2022, pp. 286-288, doi: 10.1109/ISMAR-Adjunct57072.2022.00065.

[12] B. Zhang, Y. Cui, M. Wang, J. Li, L. Jin and D. Wu, ”In Vitro Fertilization (IVF) Cumulative Pregnancy Rate Prediction From Basic Patient Characteristics,” in IEEE Access, vol. 7, pp. 130460-130467, 2019, doi: 10.1109/ACCESS.2019.2940588.

[13] S. Alshakrani, S. Hilal and A. M. Zeki, ”Hybrid Machine Learning Algorithms for Polycystic Ovary Syndrome Detection,” 2022 Inter- national Conference on Data Analytics for Business and Industry (ICDABI), Sakhir, Bahrain, 2022, pp. 160-164, doi: 10.1109/ICD- ABI56818.2022.10041525.