World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Gynecological disease prediction by machine learning

Ashraful Islam *, Rupa Parvin and Tania Sultana

Department of CSE, Daffodil International University, Dhaka, Dhaka, Bangladesh.

World Journal of Advanced Research and Reviews, 2024, 23(03), 2107–2112

Publication history: Received on 08 August 2024; revised on 17 September 2024; accepted on 19 September 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.23.3.2862

Abstract

The rapid change in climate affects the ecosystem of animal life worldwide. The consumption of processed foods and excess pesticides on crops create different difficulties for the human body. Nowadays especially girls are suffering from various kinds of gynecological diseases. Miscarriage and Anemia are very common among them. Machine learning algorithms are well-liked and widely used for disease forecasting. The detailed information we obtained from the survey established the data set. To predict gynecological diseases, we utilized 5 machine learning algorithms. Three of them are statistical based like Decision Tree, K- Nearest Neighbor (KNN), and Naive Bayes classifier and 2 are hybrid modeling like Decision tree and SVM(DT&SVM), Random Forest, and Naïve Bayes (RF&NB). We took into account the top algorithm for gynecological illness prediction, according to the correct interpretation. The best results were obtained by the Decision Tree ensemble and Naïve Bayes out of all the algorithms, with an accuracy of 86.30% and a Recall score of 87.15%. Thanks to the top completion method, our model has excellent gynecological illness prediction capabilities.

Keywords: Gynecological disease; Machine learning; Data Analysis; Statistical Analysis

1. Introduction

Bangladesh is lagging far behind many other emerging nations. People still tend to be conservative and resistant to new ideas. Many girls in our nation suffer from various diseases because of the prevalence of underage marriage, Malnutrition, and lack of health knowledge. Nearly 60% of Bangladeshi females are married before they turn 18 and 22% before they turn 15. Regrettably, with almost 4.5 million women married or in union before the age of 18, Bangladesh ranks second internationally for both the prevalence of child marriage and the total number of women in such unions.[1] Malnutrition is another concern that is not only harmful only to women but also to their successors.[2] A different study shows that the main influencer for gynecological cancer is malnutrition.[3][4] Another problem is most rural areas girl did not have proper knowledge in their adolescent time. In the age of artificial intelligence now it is very easy to early detect of human disease based on historical data.[5] In this study, we worked on this specific area to mitigate gynecological problems in Bangladesh. There are five features used in this work age, iron deficiency, incompetent cervix, excess menstrual bleeding, frequent pregnancy, and disease is the target where 'yes' and 'no' level is used. We collected our dataset by surveying rural and town areas. After preprocessing finally we got about 1458 rows of data which are perfect for training. There are four traditional and one hybrid model used for training purposes. Most of them performed very similarly but our hybrid model produced better recall than another model. As we need less false negative rate this model is finally selected for real-life implementation.

Section II provides the literature review, which is one of several sections of the report. In Section III, the technique that has been suggested is detailed. In Section 4, we go over the experimental results, and in Section 5, we wrap off by discussing the results and what comes next.

* Corresponding author: Ashraful Islam

2. Literature review

Machine learning is very useful for solving different data mining tasks. Nowadays machine learning is used to predict different things. It has brought unimaginable success in various disease predictions. Some of them are bellowed.

Kurt Benirschke et. al. [6] presented that numerous diseases may obstruct pregnancy. They reviewed the most significant of those that influence placental improvement and function. With some of these dilemmas of pregnancy, the placental decisions may be affirmative of these diseases. On the other hand, placental diagnostics may be the primary implication of an anomaly. Arezzo et. al.[7] designed a complex three different decision-making methods like KNN, RF, and Logistic Regression that can detect ovarian cancer based on gynecological ultrasound. There dataset contains 12 features and five-fold cross-validation is used during training. The highest score achieved by RF algorithm rate is accuracy 93.7%, precision 90%, and recall 90%. Paul Fiadjoe et. al. [8] represented a principal diagnostic and remedial challenge about urological diseases in maternal situations. They have said, at the time of pregnancy, the urinary tract bears unusual anatomical and physiological alterations that may occur in several manifestations and neurotic diseases harming both the mother and fetus. By doing proper appraisement and hasty therapy, the forecasting can give a good result. They described urological difficulties in pregnancy, respectively infection, calculus, renal failure, renal tumor, lower urinary tract signs and shock, and their management. Carlos Sotomayor-Beltran et. al. [9] showed that Anemia caused by iron deficiency in pregnant women is one of the malnutrition conditions. They used public data from the National Institutes of Health (INS) from Peru to study the spatial distribution pattern of anemia among pregnant women at the regional level in Peru in 2017. They found that pregnancy-related anemia is common in many areas of the highlands and is considered a serious health problem (prevalence ≥40%). Their maps are made using the local Moran. By using ArcGIS software, they can display many hot spots in the central and southern parts of the high-lands. Adiba Akhtar Khalil et. al. [10] reviewed/tested the hemoglobin and hematocrit in the third trimester of pregnancy, a retrospective study was conducted in the outdoor patient department of the Rawalpindi United Military Hospital for descriptive research. They studied 500 cases. Among 500 cases of complete anemia, 241 cases (48.2%) were anemia. The severity of mild anemia was 39.8%, the severity of moderate anemia was 7.6%, and the severity of severe anemia was 0.8%. The main types of anemia that affect the study samples are iron deficiency anemia and beta-thalassemia features. The percentages obtained are 41.6% and 4.8%, respectively. In their study population, iron deficiency in the third trimester is common. Rajesh Kannan Megalingam et. al. [11] discussed pregnant women from rural areas who do not undergo regular check-ups in the early stages of pregnancy. They found from research that routine examinations can prevent the birth of dis-abled children and greatly reduce fetal mortality. They developed a system and performed an ultrasound scan of the pregnant woman in the system, and then measured some important parameters of the pregnant woman, including vital signs, ECG, temperature, and blood pressure, and saved them on a memory card. Additionally, they created a mobile app that can access data and monitor for emergencies. In the event of an emergency, the medical personnel will be prompted to send a comprehensive message including the patient's details to the doctor by SMS. Kim et. al.[12] introduced a new approach to detecting gynecologic cancer using seven different features of clinical and patient demographics data. They used four traditional ML algorithms and two boosting algorithms. The highest performance was achieved by the Random Forest algorithm. The test AUC rate is 81% and for five-fold cross validation, AUC is 82%. The classification result of this model is an accuracy of 88.88%, sensitivity of 90.12% specificity of 87.65%. M. Mehedi Hasan et. al [13] introduced the prediction of pneumonia disease based on survey data. They used a different machine-learning algorithm to predict pneumonia in children. They tried to find out the reason behind pneumonia's affection for the child. This study mainly focused on maternal condition analysis. They talked about the effect of early marriage also and showed statistical analysis. The highest accuracy is achieved by the decision tree algorithm and the accuracy rate was 90.61%.

For the above discussion, we can differentiate our work by many aspects. Our work is totally survey-type. We also added statistical analysis and tried to provide a visual representation of this analysis.

3. Material and methods

In this work, we followed a total of 5 steps to address the problem. Data collection, Statistical analysis, Data preprocessing, Algorithm Implementation, and Evaluation are the main parts. For explicit analysis, some parts are divided into subparts. Figure 1 represents our overall methodology.

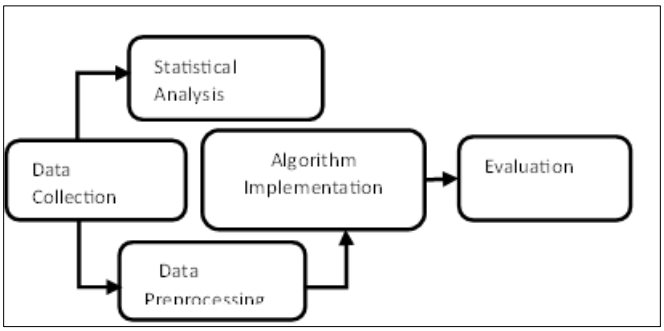

Figure 1 Methodology Diagram

3.1. Data collection

We collected our data by survey from three different district in rural area in Bangladesh. In this survey basically, we focused on women who were affected by anemia or miscarriage. We divided our dataset into three categories based on the woman's age when she got married. Category 1 contains 1-15, category 2 contains 16- 35 and Category 3 contains upper 35 years old women. Most of our question was yes/no type like:

- How old is she?
- Is she suffering from anemia?
- Did she have a miscarriage?
- How old was she when she got married?
- We also collected information about Iron deficiency, Incompetent Cervix, Excess Menstrual Bleeding, and Frequent Pregnancy.

This survey was applied to around 10000 women. We found around 1500 women who are affected by different gynecological disease

3.2. Statistical analysis

In this part, we actually tried to find out the probability and effects of miscarriage and anemia. We also tried to show our survey knowledge by graphical representation. Figure 2 represents the percentage of two diseases of 1500 women. From this graph, we can see that 65.3% of women are affected by Anemia and 34.7% of women are affected by miscarriage. That means the miscarriage rate is comparatively less than anemia.

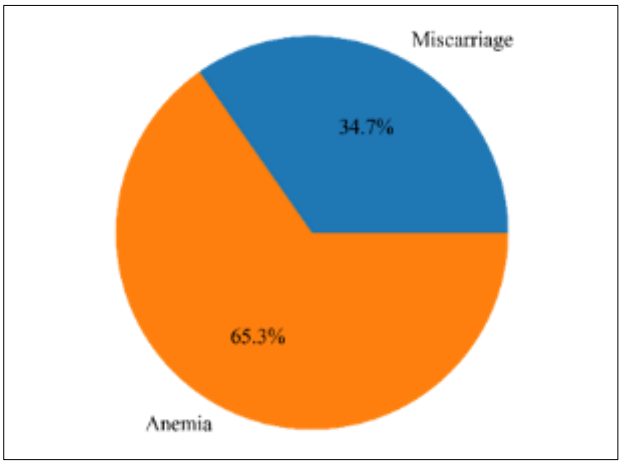

Figure 2 Disease rate

3.3. Age analysis

Figure 3. Represents the marriage age and gynecological disease rate. In this graph, blue color, orange and green color represent the marriage age range of 14-19, 20-35, and 35 years up correspondingly. For miscarriage, we can see that if women get married at the age of 14-19 and have a child then the a possibility to be affected by miscarriage. At the age

of 20-35, there is less possibility of getting a miscarriage. For anemia we can see at the age of 14 -19 there is a very high possibility to be affected by anemia and those whose age is older than 35 are also at higher risk.

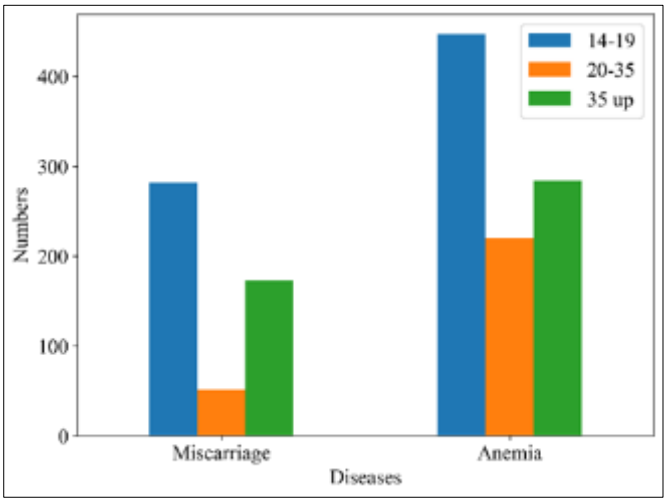

Figure 3 Disease rate

3.4. Data preprocessing

Name, Location, Iron Deficiency, Incompetent Cervix, Excess Menstrual Bleeding, Frequent Pregnancy, and Disease Name are in our raw data. In the preprocessing stage, Name and Location aren't needed to train our model and remove incomplete responses. Therefore, we eliminated these characteristics. We converted all the data to integers because Machine Learning algorithms cannot read strings. Thus, all yes/no questions are represented by 0 or 1. Ages 15-19, 20-35, 35 are 1, 2, 3. Gynecological illness positive = 0 and negative = 1.

3.5. Algorithm implementation

To predict anemia or miscarriage we used classification algorithms. We employed six machine learning algorithms with the following parameters by which algorithms produced the best accuracy. Each parameter is given in Table 1.

Table 1 Hyperparameter Tuning process

Algorithms	Details
KNN	K=3, p=2, randomstate=42, weight= Distance
Decision Tree	Random state=1, splitter= best, criterion = entropy
Naïve Bayes	Random state = 42, classifier = GaussianNB
Random Forest and SVM	Max depth= None, min samples leaf= 1, Min samples split= 2, n estimators= 50 C= 0.1, gamma=scale, kernel=linear
Decision Tree and Naïve Bayes	Max depth= None, min samples leaf=1, Min samples split= 2, var_smoothing=0.657933224657568

4. Experimental result

In order to prepare our historical data for analysis, five different machine learning techniques were employed. We used a variety of classifiers and methods, including K-Nearest Neighbor (KNN), Naive Bayes, Decision Tree, a mix of random forest and support vector machine (SVM), decision tree, and naïve Bayes. We classified these algorithms according to their recall, accuracy, precision, and F1 score. From the accuracy table, we can deduce that the majority of algorithms worked better when tested with 30% test data. The KNN algorithm achieved its maximum accuracy of 85.27% with 40% test data. Additionally, DT, NB, RF&SVM, and DT&NB all attained varying degrees of accuracy: 86.28%, 85.30%, 86.30%, and 81.28%.

Table 2 Accuracy comparison table

Data usages	KNN	NB	DT	RF & SVM	DT & NB
30%	84.47	81.28	85.30	86.30	86.28
40%	85.27	78.94	84.20	85.24	85.27
50%	84.93	77.67	83.23	84.45	84.93
60%	78.77	76.71	83.32	84.34	84.82
70%	81.51	77.20	79.45	80.75	80.54

As most of the algorithm performed better using 30% test data so we have calculated score matrices by using 30% test data only. As we work with disease detection so we need to find out less false negative rate(better recall) performance. From Table 4 we can see that RF&SVM hybrid algorithm produced better recall than another algorithm. Analyzing the results of all algorithms, we have chosen the RF&SVM for its excellent performance.

5. Discussion

In evaluation stage we used test data which are never seen by our model before. Then we calculated precision recall and f1 score.

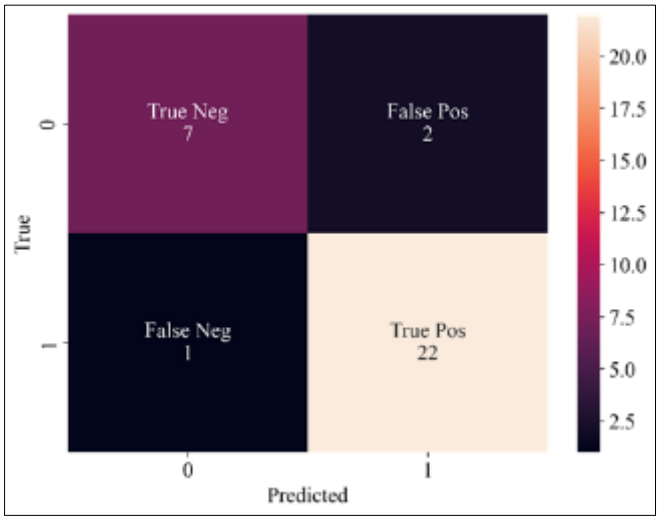

Figure 4 Confusion matrix

$$\text{Accuracy: } \frac{TP+TN}{TP+FP+FN+TN} = \frac{22+7}{1+7+2+22} = 0.91 * 100 = 91\%$$

$$\text{Error: } 1 - 0.91 = 0.09*100 = 9\%$$

$$\text{Precision rate: } \frac{TP}{TP+FP} = \frac{22}{22+2} = 0.92*100 = 92\%$$

$$\text{Recall rate : } \frac{TP}{TP+FN} = \frac{22}{22+1} = 0.96*100 = 96\%$$

We validated our prediction by the confusion matrix in Figure 4. For validation, dataset accuracy is 91%, and precision, and recall rates are 92%, and 96% correspondingly.

6. Conclusion

Prevention is better than cure. In this case, awareness is very important for everyone. Bangladesh's Government has already taken many steps against the reduction of these diseases. But people do not care about child marriage and its negative sides. Mostly in rural areas, it is very usual to see these types of cases because of their illiteracy rate. As a result, the death ratio of premature babies is increasing day by day. Death from premature birth is now the principal problem in our country. In our research, we have applied 5 Machine Learning algorithms for early detection of

gynecological disease. Among all of these Algorithms, Random Forest has given the best accuracy of 86.56%. In the future, we will develop a system that will help increase awareness for everyone. It will reduce the number of gynecological diseases by giving proper information.

Compliance with ethical standards

Disclosure of conflict of interest

The authors declare that there are no conflicts of interest regarding the publication of this paper. All authors have reviewed the manuscript and agree to its publication.

References

[1] G. n. Brides, "Child Marriage Rates," Girls Not Brides, 2017. [Online]. Available: https://www.girlsnotbrides.org/child-marriage/bangladesh. [Accessed: May 3, 2024].

[2] H. Mehedi, I. Sutradhar, A. Shahabuddin, and M. Sarker, "Double Burden of Malnutrition among Bangladeshi Women: A Literature Review," Cureus, vol. 9, no. 12, Art. no. e1986, Dec. 2017.

[3] B. Laky, M. Janda, S. Kondalsamy-Chennakesavan, G. Cleghorn, and A. Obermair, "Pretreatment malnutrition and quality of life-association with prolonged length of hospital stay among patients with gynecological cancer: a cohort study," BMC Cancer, vol. 10, pp. 1-6, 2010.

[4] M. Morton, J. Patterson, J. Sciuva, et al., "Malnutrition, sarcopenia, and cancer cachexia in gynecologic cancer," Gynecologic Oncology, vol. 175, pp. 142-155, 2023.

[5] G. A. Saleh et al., "The role of medical image modalities and AI in the early detection, diagnosis and grading of retinal diseases: a survey," Bioengineering, vol. 9, no. 8, Art. no. 366, 2022.

[6] K. Benirschke, G. J. Burton, and R. N. Baergen, "Maternal diseases complicating pregnancy: diabetes, tumors, preeclampsia, lupus anticoagulant," in Pathology.

[7] F. Arezzo et al., "A machine learning approach applied to gynecological ultrasound to predict progression-free survival in ovarian cancer patients," Arch Gynecol Obstet, vol. 306, pp. 2143–2154, 2022.

[8] P. Fiadjoe, K. Kannan, A. J. E. J. O. Rane, "Maternal urological problems in pregnancy," Gynecology and Reproductive Biology, vol. 152, no. 1, pp. 13-17, 2010.

[9] C. Sotomayor-Beltran, G. W. Zarate Segura, and D. Tarazona, "Anemia During Pregnancy in Peru in 2017: A Geographic Information System Study," in IEEE 38th Central America and Panama Convention (CONCAPAN XXXVIII), San Salvador, 2018, pp. 1-5.

[10] A. Khalil, T. Jabbar, S. Akhtar, and S. Mohyuddin, "Frequency and Types of Anemia in an Antenatal Clinic in the Third Trimester of Pregnancy," PAFMJ, vol. 57, no. 4, pp. 273-278, Dec. 2007.

[11] E. Krisnanik, K. Tambunan, and H. N. Irmanda, "Analysis of Pregnancy Risk Factors for Pregnant Women Using Analysis Data Based on Expert System," in 2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), Jakarta, Indonesia, 2019, pp. 151-156.

[12] B. W. Kim et al., "Machine Learning for Recurrence Prediction of Gynecologic Cancers Using Lynch Syndrome-Related Screening Markers," Cancers, vol. 13, no. 22, Art. no. 5670, 2021.

[13] M. Mehedi Hasan et al., "Prediction of Pneumonia Disease of Newborn Baby Based on Statistical Analysis of Maternal Condition Using Machine Learning Approach," in 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2021, pp. 919-924.