Vol 4 | Issue 1 | Pages 111-116

Insights of Biomedical Research

ISSN: 2642-4576

Research Article

DOI: 10.36959/584/456

Prediction of Childhood Diarrhea in Bangladesh using Machine Learning Approach

Md. Maniruzzaman[1], Md. Shaykhul Islam[2], Md. Menhazul Abedin[1], Md Amanullah[3] and Sadiq Hussain[4]*

[1]*Statistics Discipline, Khulna University, Khulna-9208, Bangladesh*

[2]*Department of Statistics, University of Rajshahi, Rajshahi-6205, Bangladesh*

[3]*Department of Respiratory Medicine, Sir Run Run Shaw Hospital and Institute of Translational Medicine, Zhejiang University School of Medicine, Hangzhou, Zhejiang 310016, China*

[4]*Dibrugarh University, India*

Abstract

Diarrhea has remained a major health problem among under-five (U5) children that leads high level of morbidity and mortality. The objective of this study is to determine the socio-demographic risk factors of diarrhea as well as predict of diarrhea status using machine learning (ML) based approach among U5 children in Bangladesh. Bangladesh Demographic and Health Survey, 2014 dataset is used in this study. This dataset consisted of 7,538 respondents who had 371 (4.9%) child's diarrhea. Logistic regression (LR) is used to determine the high-risk factors of diarrhea. Then four ML-based approach namely naïve Bayes (NB), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and support vector machine (SVM) was applied to predict the child's diarrhea status and accuracy, sensitivity, and specificity are used to evaluate the performance of these classifiers. Around 4.9% women reported that their children have experienced an episode of diarrhea in two weeks before the survey. LR model showed that the child's age, region (Khulna and Rangpur), mothers who had completed secondary education, and respondents who were rich wealth index, significantly associated risk factors for diarrhea disease. Our findings indicate that SVM with radial basis kernel yielded 65.61% accuracy, 66.27% sensitivity, and 52.28% specificity which are comparatively better than others. The prevalence of diarrhea disease is more common among Bangladeshi children. Our study shows that SVM is capable of predicting child diarrhea status (generally highly imbalanced data). This study allows policy makers towards appropriate decisions to reduce childhood diarrhea in Bangladesh.

Keywords

Diarrhea, Under-five children, Machine learning, Bangladesh

Introduction

Diarrhea is a major health problem in any developed and developing countries like Bangladesh. Globally, 1 in 9 children died due to diarrhea [1] and diarrhea is second leading cause of mortality [2]. The symptoms of diarrhea are passing loose, three or more times watery stools in a 24-hour period [3]. There were about 1.7 million children cases of diarrhea and 525000 under-five children died due to diarrhea in worldwide [2,4].

In the previous literature, there were lots of studies to identify the risk factors of diarrheal disease in Bangladesh [5-7]. Based on our knowledge, there were no studies in Bangladesh to identify and predict the childhood diarrhea using machine learning (ML) based approach. In this study, an attempt has been made to identify the risk factors of diarrhea. Moreover, the ML based approach is used to predict the childhood diarrhea in Bangladesh.

Material and Methods

Data source

The study has been used a representative set of cross-sectional data extracted from the Bangladesh Demographic and Health Survey (BDHS), 2014. The BDHS, 2014 included a household survey of ever-married women 15-49 years [5] having 7886 respondents. The datset contains some missing values and omiiting these missing values. Finally, 7538 respondents are considered for analysis.

***Corresponding author:** Md. Maniruzzaman, Statistics Discipline, Khulna University, Khulna, Bangladesh

Accepted: November 23, 2020

Published online: November 25, 2020

Citation: Maniruzzaman, Islam S, Abedin M, et al. (2020) Prediction of Childhood Diarrhea in Bangladesh using Machine Learning Approach. Insights Biomed Res 4(1):111-116

Potential risk factors

We have divided the child's age into three groups namely (0-11) months, (12-23) months and (24-59) months. Wealth index is divided into five categories (poorest, poorer, middle, richer and richest), but for our calculation, this index is divided into three categories as poor, middle and rich. The main predictor variables are age of the child, sex of child, region, mother's education, wealth index, source of drinking water.

Outcome variable

In this study, we have considered childhood diarrhea as a dependent variable. We have defined the dependent variables as:

$$\begin{cases} 1, \text{if the children had diarrhea} \\ 0, \text{Otherwise} \end{cases}$$

Statistical analysis

The Chi-squared test and binary logistic regression (LR) analyses are used as statistical tools. Differences in variables between the children who had diarrhea (yes/no) are analyzed using Chi-Square test for categorical variables. The methodology is used LR based model [8-11], where the goal is to determine the risk variables/factors of childhood diarrhea. Finally, naïve Bayes (NB) [8,12], linear discriminant analysis (LDA) [8,13], quadratic discriminant analysis (QDA) [8,14], and support vector machine (SVM) [8,15] are used to predict the childhood diarrhea. Stata version 14 and Ri86 3.4.2 was used for the analysis.

Result

Baseline characteristics of the respondents

Table 1 shows the baseline characteristics of the respondents. Distribution of the prevalence of diarrhea is 4.9% of the total child had diarrhea in the last two weeks before the study. The highest prevalence of diarrhea (6.6%) is in age group 0-11 months and lowest (3.2%) in age group 24-59 months' children. Male children have affected diarrhea by 5.3% and females are 4.6%. The highest prevalence of childhood diarrhea is in Chittagong division (6.4%) and the lowest in Rangpur region (2.6%). The prevalence of diarrhea is higher in rural areas (5.0%) compared to urban areas (4.9%). Table 1 confirms that a lower level of mother's education (primary) causes the highest rate of diarrhea (6.1%) whereas, the lowest rate of diarrhea (2.4%) with very higher mother' education. The respondents with poor socio-economic status is reported the highest prevalence of diarrhea (5.6%) and the middle and rich were both lower (4.5%). In these study areas, 4.9% children are affected by diarrhea in Muslim families and 4.7% Hindu, and 5.1% in others. Table 1 confirms that 4.7% children are affected by diarrhea which family take water from tube well and 9.2% for pond. The Chi-square test are revealed that children's age, region, mother's education, wealth index and source of drinking water are statistically significant associated (p-value < 0.05) with the prevalence of diarrhea of children.

Risk factors for diarrhea using logistic regression

The LR model is used to assess the net effects of socio-de-

Table 1: A baseline characteristic of the respondents.

Characteristic	Had Diarrhea recently		Overall, N (%)	p-value[1]
	Yes, N (%)	No, N (%)		
Total	371 (4.9)	7167 (95.1)		
Age of child (in months)				
0-11	200 (6.6)	2814 (93.4)	3014 (40.0)	< 0.001
12-23	123 (4.1)	2911 (95.9)	3034 (40.2)	
24-59	48 (3.2)	1442 (96.8)	1490 (19.8)	
Sex				0.162
Male	204 (5.3)	3674 (94.7)	3878 (51.4)	
Female	167 (4.6)	3493 (94.4)	3660 (48.6)	
Type of place				0.812
Urban	120 (4.9)	2276 (95.0)	2396 (31.8)	
Rural	251 (5.0)	4891 (95.1)	5142 (61.2)	
Region				< 0.001
Barisal	50 (5.7)	825 (94.3)	875 (11.6)	
Chittagong	93 (6.4)	1358 (93.6)	1451 (19.2)	
Dhaka	68 (5.1)	1261 (94.9)	1329 (17.6)	
Khulna	29 (3.5)	791 (96.5)	820 (10.9)	
Rajshahi	37 (4.0)	881 (96.0)	918 (12.2)	
Rangpur	24 (2.6)	892 (97.4)	916 (12.2)	
Sylhet	70 (5.7)	1159 (94.3)	1229 (16.3)	

Mother's education				
No education	57 (4.9)	1107 (95.1)	1164 (15.4)	
Primary incomplete	73 (6.1)	1124 (93.9)	1197 (15.9)	
Primary complete	50 (5.6)	848 (94.4)	898 (11.9)	0.045
Secondary incomplete	142 (4.8)	2831 (95.2)	2973 (39.4)	
Secondary complete	12 (4.6)	485 (97.6)	497 (6.6)	
Higher	37 (2.4)	772 (95.4)	809 (10.7)	
Wealth index				
Poor	170 (5.6)	2890 (94.4)	3060 (40.6)	
Middle	65 (4.5)	1389 (95.5)	1454 (19.3)	0.034
Rich	136 (4.5)	2888 (95.5)	3024 (40.1)	
Religion				
Islam	342 (4.9)	6583 (95.1)	6925 (91.9)	
Hindu	26 (4.7)	528 (95.3)	554 (7.3)	
Others	3 (5.1)	56 (94.9)	59 (0.8)	0.966
Source of drinking water				
Tube well	285 (4.7)	529 (95.3)	814 (7.5)	
Tab	35 (6.2)	5808 (93.8)	6093 (80.8)	
Pond	12 (9.2)	119 (90.8)	131 (1.7)	
Others	39 (5.2)	711 (94.8)	750 (9.9	0.048

[1]p-value is obtained from Chi-Square test

Table 2: Logistic regression for the effect of variables on children who had diarrhea.

Characteristics	**OR**	**p-values**	**95% CI for OR**	
			Lower	**Upper**
Age (in years)				
0-11	1.000	--	--	--
12-23	0.585	< 0.001	0.464	0.738
24-59	0.458	< 0.001	0.331	0.633
Sex				
Male	1.000			
Female	0.859	0.159	0.695	1.061
Type of place				
Urban	1.000			
Rural	0.921	0.538	0.707	1.198
Region				
Barisal (*Ref*)	1.000	--	--	--
Chittagong	1.045	1.000	0.685	1.594
Dhaka	1.015	1.000	0.676	1.523
Khulna	0.550	0.046	0.310	0.979
Rajshahi	0.653	0.108	0.392	1.087
Rangpur	0.412	0.002	0.232	0.732
Sylhet	0.951	0.858	0.590	1.535
Mother's education				
No education (*Ref*)	1.000	--	--	--
Primary incomplete	1.294	0.162	0.901	1.858

Primary complete	1.124	0.563	0.756	1.670
Secondary incomplete	1.064	0.716	0.761	1.489
Secondary complete	0.505	0.041	0.263	0.972
Higher	1.042	0.863	0.654	1.660
Wealth index				
Poor (*Ref*)	1.000	--	--	--
Middle	1.020	0.929	0.788	1.321
Rich	0.773	0.023	0.623	0.959
Religion				
Islam (*Ref*)	1.000	--	--	--
Hindu	1.060	0.784	0.699	1.606
Others	0.782	0.684	0.239	2.557
Source of water				
Tube well (*Ref*)	1.000	--	--	--
Tab	0.736	0.152	0.484	1.119
Pond	1.331	0.458	0.626	2.832
Others	0.775	0.331	0.463	1.296

Table 3: Kernel optimization for support vector machine.

Kernel types	**Performance measures**		
	Accuracy (%)	**Sensitivity (%)**	**Specificity (%)**
Linear	64.26	64.88	51.18
RBF	**65.61**	**66.27**	**52.28**
Polynomial-2	64.36	64.96	43.15
Sigmoid	58.58	59.35	51.57
Laplace	64.67	65.34	52.28

RBF: Radial Basis Function

mographic variables on childhood diarrhea. Odds ratios (OR) are used to compare different groups with 95% confidence interval (CI) of OR presented in Table 2. Out of eight independent variables, four, viz. age of the child, region, mother's education and wealth index are statistically significant at 5% levels of significance. The child who were in age group 12-23 months (OR, 0.585; 95% CI, 0.464 -0.738) less prevalence of child diarrhea but the child who were in age group 24-59 month's had (OR, 0.458; 95% CI, 0.331-0.633) less prevalence of child diarrhea as compared to the children of age group 0-11 months. A child who was in Khulna division had (OR, 0.550; 95% CI, 0.310-0.979) and Rangpur division had (OR, 0.412; 95% CI, 0.232-0.732) less prevalence of child diarrhea as compared to the children in Barisal division. Mother's with secondary complete or higher had (OR, 0.505; 95% CI, 0.263-0.972) less prevalence of child diarrhea as compared to mother's had no education. Again respondents in rich wealth index were 22.7% (OR, 0.773; 95% CI, 0.623-0.959) less prevalence of child diarrhea as compared to children with poor wealth index.

Kernel optimization

Support vector machine with five different kernels namely, linear, polynomial, sigmoid, Laplace, and radial basis function (RBF) are adopted in this study. To evaluate the performances of these models, the 10-fold cross-validation protocol is performed and chooses the best kernel which gives the highest classification accuracy. By tuning all of these kernels the best performance are summarized in Table 3. It is observed that the RBF kernel gives the highest classification accuracy of 65.61% along with 66.27% sensitivity and 52.28% specificity. Therefore, the RBF kernel is select for SVM for predicting childhood diarrhea.

Comparison of the classifiers

The comparison of the performances of these classifiers is presented in Table 4. In this study, we have used four classifiers as SVM, NB, LDA, QDA for predicting childhood diarrhea in Bangladesh. The accuracy, sensitivity, and specificity are used to evaluate the performance of these classifiers. It is observed that NB and LDA have greater accuracy and sensitivity than SVM but they failed to detect a single one observation from the small group that is they has zero specificity so that for this data NB and LDA are totally avoided. On the other hand accuracy and sensitivity of QDA is higher but specificity is very low compared to SVM. Hence according to our objectives SVM classifier is the best classifier compared to others. It may conclude that SVM is the best classifier for predicting

Table 4: Comparison of the performance of different classifiers.

Classifier types	Performance measures		
	Accuracy (%)	Sensitivity (%)	Specificity (%)
NB	95.26	1.00	0.00
LDA	95.24	1.00	0.00
QDA	88.57	92.60	7.61
SVM	**65.61**	**66.27**	**52.28**

Table 5: Validations of the results of SVM for the simulated dataset.

Classifier types	Accuracy (%)
NB	50.85
LDA	51.80
QDA	51.15
SVM	**100.00**

childhood diarrhea.

Validation of the results of SVM

In order to validate of the performances of SVM, we have used a simulated dataset. We have simulated/generated 800 observations from which 400 observations are in class 1 and the rest of the observations in class 2. This data set simulated using the normal distribution with different mean and standard deviation. The performance of SVM is presented in Table 5. It is noted that the SVM classifier gives the highest classification accuracy compared to others.

Discussion

This also shows that child's age, region, mother's education and wealth index are found to be significantly associated with childhood diarrhea. Our findings show that child's age has the significant impact on diarrhea. It is observed that the children who ages are 12-23 and 24-59 moths have the lower prevalence of diarrhea. It is also observed that Khulna region and Rangpur region have the lower prevalence of diarrhea compared to Barisal region. The mothers who have completed secondary education is an important factor of child's had diarrhea which were similar finings with the previous studies [16,17]. Because an educated mothers is more conscious about his own life and their children life.

It is noted that wealth index had a significant impact on diarrhea. The children who are come from rich family had the lower prevalence of diarrhea compared to poor [17]. In the previous studies, there was no study for the prediction of childhood diarrhea disease. Best our knowledge, this is the first time, we applied four ML-based approach as NB, LDA, QDA, and SVM to predict the childhood diarrhea. Our study shows that SVM with RBF gives better performance scores compared to others.

Conclusions

This study investigate the risk factors of childhood diarrhea and also suggests a prediction model to predict childhood diarrhea. This shows that child age, mother's education, region and wealth index are significant impact on childhood diarrhea. In this study, LDA, QDA, NB, and SVM-based classifiers are used to predict the childhood diarrhea status. SVM with radial basis kernel gives better performance compared to others.

Conflict of Interest

The authors declare that they have no conflict of interest.

Ethical Approval

The study was supported by the Ethics committee in Bangladesh.

Funding

We have no funding for this study.

References

1. Liu L, Johnson HL, Cousens S, et al. (2012) Global, regional, and national causes of child mortality: An updated systematic analysis for 2010 with time trends since 2000. Lancet 379: 2151-2161.
2. Diouf K, Tabatabai P, Rudolph J, et al. (2012) Diarrhoea prevalence in children under five years of age in rural Burundi: An assessment of social and behavioural factors at the household level. Glob Health Action 7: 24895.
3. Jamison DT, Breman JG, Measham AR, et al. (2006) Disease control priorities in developing countries. World bank publications, Washington, DC.
4. Shine S, Muhamud S, Adanew S, et al. (2020) Prevalence and associated factors of diarrhea among under-five children in Debre Berhan town, Ethiopia 2018: A cross sectional study. BMC Infectious Diseases 20: 1-6.
5. Farthing M, Salam MA, Lindberg G, et al. (2013) Acute diarrhea in adults and children: A global perspective. J Clin Gastroenterol 47: 12-20.
6. Haque R, Mondal D, Kirkpatrick BD, et al. (2003) Epidemiologic and clinical characteristics of acute diarrhea with emphasis on Entamoeba histolytica infections in preschool children in an urban slum of Dhaka, Bangladesh. The American Journal of Tropical Medicine and Hygiene 69: 398-405.
7. Robert E Black, Simon Cousens, Hope L Johnson, et al. (2010) Global, regional, and national causes of child mortality in 2008: A systematic analysis. Lancet 375: 1969-1987.
8. Maniruzzaman M, Rahman MJ, Hasan MAM, et al. (2018) Accurate diabetes risk stratification using machine learning: Role of missing value and outliers. Journal of Medical Systems 42: 92.
9. Maniruzzaman M, Suri HS. Kumar N, et al. (2018) Risk factors of neonatal mortality and child mortality in Bangladesh. J Glob Health 8.
10. Tabaei BP, Herman WH (2002) A multivariate logistic regression

equation to screen for diabetes: Development and validation. Diabetes Care 25: 1999-2003.

11. Hosmer DW, Lemeshow S (2000) Applied logistic regression. JohnWiley & Sons, USA.
12. Ting SL, Ip WH, Tsang AH (2011) Is Naive Bayes a good classifier for document classification. International journal of software engineering and its applications 5: 37-46.
13. Sapatinas T (2005) Discriminant analysis and statistical pattern recognition. Journal of the Royal Statistical Society: Series A (Statistics in Society) 168: 635-636.
14. Zhang MQ (1997) Identification of protein coding regions in the human genome by quadratic discriminant analysis. Proceedings of the national academy of sciences 94: 565-568.
15. Cortes C, Vapnik V (1995) Support-vector networks. Machine learning 20: 273-297.
16. Kandala NB, Ji C, Stallard N, et al. (2007) Spatial analysis of risk factors for childhood morbidity in Nigeria. The American Journal of Tropical Medicine and Hygiene 77: 770-779.
17. Islam MR, Hossain MK, Khan MN, et al. (2015) An evidence of socio-demographic effects on child's diarrhoea in Bangladesh. Journal of Health Science 5: 1-5.

DOI: 10.36959/584/456