A Data Mining Model for Predicting Diarrhea in Afghan Children

Abdullah Zahirzda, Gunt Chanmas, Jonathan H. Chan

*School of Information Technology
King Mongkut's University of Technology Thonburi
Bangkok 10140, Thailand*

ab.zahirzada95@gmail.com, gunt.3648@mail.kmutt.ac.th, jonathan@sit.kmutt.ac.th

Abstract— Childhood diarrheal illnesses are a significant public health concern. Data on disease burden and factors associated with diarrheal diseases are poorly described in low-income countries, precluding successful prevention programs. Therefore, the ability to predict diarrhea is beneficial and an excellent preventative measure. This study aims to implement a predictive diarrhea model for children under the age of five in Afghanistan. The data has been obtained from Afghanistan Demographic and Health Survey 2015. The study's main objective is to identify the most suitable machine learning techniques (i.e., classifiers). The three most popular techniques include Naïve Bayes, Random Forest, and Support Vector Machine (SVM) are selected for this study. Data preprocessing is carried out before implementing the predictive models. It has been carried out through cleaning data and selecting features. The information Gain and Correlation-based Feature Selection algorithms are used to determine the most informative attributes. The classifier in this study comprises of two categories, 'Diarrhea' and 'Non-diarrhea', to ensure well-balanced samples in each group and the preparation of the dataset is performed carefully. The study found that Random Forest is the best classifier with an accuracy of 81.48% and with the Area Under Curve (AUC) of 89.80%, respectively. The study can directly benefit the health departments, influence prevention decisions, and assist policymaking in Afghanistan.

Keywords— *Afghanistan Demographic and Health Survey, Information Gain, Correlation-based Feature Selection, Diarrhea, Random Forest*

I. INTRODUCTION

Diarrhea is the world's second leading cause of infant mortality after pneumonia in children under five [1]. Diarrhea is a significant cause of childhood morbidity and mortality [2]. And it is responsible for killing around 525,000 children every year. There are three clinical types of diarrhea: acute watery diarrhea, acute bloody diarrhea, and persistent Diarrhea [WHO]. Diarrheal disease is often caused by polluted sources of food and water. Seven hundred eighty million people worldwide lack access to better drinking water, and 2.5 billion lack improved sanitation. Diarrhea is common in many developing countries due to the infection among children under three years of age who suffer an average of three diarrhea episodes per year in low-income countries. Each episode deprives the child of the required nutrition for development. As a result, diarrhea is a significant cause of malnutrition, and diarrhea is more likely to make malnourished children sick—[WHO, 2]. The death of small children under the age of five is infant mortality. The child mortality rate (IMR) calculates this death toll [3, 4]. In 1990, nine million infants younger than one year died globally. This figure has almost been halved to 4.6 million child deaths by 2015 [5].

Child mortality, calculated as the under-5 child mortality rate (U5MR), is the child's death before the child's fifth birthday. National statistics also groups these two rates of mortality. Globally, in 2017, 5.4 million children died before their fifth birthday [6]. A reduction of child mortality is now a target in the Sustainable Development Goals (SDG) for Goal Number 3 ("Ensure healthy lives and promote well-being for all at all ages") [7]. In 2012, in terms of the total under-five mortality rate, Afghanistan was ranked 18th globally. Similarly, Afghanistan was ranked 8th among nations with the highest deaths of under-five from pneumonia and diarrhea. Thus, it is not surprising that the two most prevalent health problems claiming the lives of under-five Afghan children are diarrhea and pneumonia.

Furthermore, it is beneficial to identify the factors associated with Diarrhea in Afghanistan through a predictive diarrhea model. Therefore, this study proposes a predictive model for diarrhea children under five with specific Afghanistan characteristics. It also aims to find a suitable classifier for this task. Three popular classifiers, Naïve Bayes, Random Forest, and Support Vector Machine are selected for the study. The dataset used in this study is Afghanistan's Demographic and Health Survey 2015, which comprises records of children whose age group is under five years.

II. CLASSIFICATION SELECTION

Classification is a supervised learning technique whose primary objective is to construct models based on known data and predict new data categories [8,9]. In classification, models are built by splitting a supplied dataset into training and test sets. One or more classification algorithms run through the training set, and the classifier models are subsequently developed. The test set is then used to assess the accuracy of the models [5]. Subsections A to C briefly introduces the classifiers selected for this study, while Section III briefly describes previous related work.

A. Naïve Bayes

Naïve Bayes learning is based on a Bayesian probabilistic model that accredits a prior-class probability to an instance. The simple Naïve Bayes algorithm uses these probabilities to assign an instance to a class. Naïve Bayes Classifier is very widely used, and it leads to a simple prediction framework that gives decent results in some practical cases. A more definitive term for the underlying probability model would be the "independent feature model." It is also called idiot's Bayes. It relies on the existence (or absence) of a particular class attribute that is irrelevant to the presence (or absence) of any other attribute. Naïve Bayes classifier will converge faster than logistic regression, so it requires less training time. This method is applicable to different applications with

various features. It is good at missing value handling. Naive Bayes has a lower error rate than the other classifier, but it is not always true practically and less accurate than other classifiers. In particular, it cannot learn interaction between features, does not support pruning, and contains sharp decision boundaries [10,11]

B. Random Forest

Like a decision tree, the random forest is an ensemble classifier that can be used to solve classification and regression issues. It uses the principle of multiple random tree generation with training dataset bootstrap, sample bagging, voting scheme, and features being randomly selected in each split decision, which improves predictive ability and results in higher performance.

Most of the time, it produces better performance compared with decision trees. One example of the random subspace approach is the selection of a random subset of features; it is used in various applications. It does not rely on the data, which is ideal for modeling high dimensional data. By overcoming, over-fitting and removing prune trees, random forest runs efficiently, and are able to provide high predictive accuracy. It can deal with missing values, outlier, and preserves accuracy. Random forest interpretability model and prediction accuracy are exceptional among common methods of machine learning. Random Forest requires much less input preparation. They can handle binary features, categorical features, and numerical features, and there is no need for feature normalization. Random Forests are quick to train and to optimize according to their hyperparameters [12, 13].

C. Support Vector Machine

Support vector machine (SVM) introduced for regression, binary classification, or ranking function, and is based on statistical learning. It is widely used by researchers in health care for classification due to its many attractive features, handling complicated nonlinear data points. The support vector machine algorithm's basic concept is to find the optimal hyperplane, which separates two considered classes, by maximizing the distance among the classes' closest points. The middle of the margin is called the optimal hyperplane. SVM is a good classifier, and it does not require prior knowledge, even if the input space is very high [12,14].

III. RELATED WORK

Several studies were conducted to extract the hidden patterns from diarrhea data in the healthcare industry. Their essential aim is to provide useful information for decision-makers and policymakers in the healthcare industry, helping to determine the target population's diarrhea status. At present, data mining and machine learning are significant contributors in terms of predicting and analyzing the factors related to diarrhea; this section focuses on previous related works done by researchers on the modeling of diarrhea.

A work [15] investigated the association between the prevalence of childhood diarrhea and caregiver awareness of the causes and the prevention of diarrhea using statistical methods in a prospective cohort of 952 children < 5 years of age in Cochabamba, Bolivia. The caregiver knowledge survey found that more than 80% of caregivers were unaware that handwashing with soap could prevent

childhood diarrhea. In this cohort, significant risk factors for diarrheal disease were a lack of knowledge of the value of hygiene and sanitation practices for diarrhea prevention. This study's knowledge finding suggests that health promotion in these communities should emphasize increasing understanding of how water treatment, handwashing with soap, proper disposal of child feces, and food preparation relate to childhood diarrhea prevention.

A study in [16] used Demographic and Health Survey data in Nepal to identify the factors associated with diarrhea and their possible influence, a Bayesian logistic regression model was applied. This study found that mothers with no formal schooling background have a higher chance that their children have diarrhea. Autumn and winter seasons have a higher prevalence of diarrhea, children under the age of 12-24 months have the highest risk category of diarrhea, and male children had a higher risk than females.

Research in [17] was a cross-sectional analysis done using a formal questionnaire and a checklist of observations. Using simple random sampling techniques, a total of 546 households with at least one child under five were chosen. Data entry and cleaning was carried out using the Statistical Package for Social Science (SPSS). For descriptive analysis, frequencies and proportions were computed. Finally, multivariable regression was used to conduct further analyses. Therefore, practices focusing on adequate handwashing methods at all appropriate times, proper handling of refuse, nutrition enhancement, and better childcare are also strongly recommended.

The work [18] at the Maternal and Child Health Clinic (MCH) in Tanzania, a matched case-control study was conducted during the rainy season to elucidate the risk factors for and etiology of diarrheal diseases in children under five years of age. Precoded questionnaires were completed with demographic data, health history, and physical signs. For bacterial, parasitological, and viral studies, stool samples were obtained. The probability of diarrhea was associated with many siblings, the number of siblings surviving, the birth order, and the distance from the house to the water source.

As stated in the above literature, various statistical, and data mining techniques have been applied to healthcare data, with reference to diarrhea. This is beneficial for developing countries to infer useful knowledge for domain area experts.

Previous works on diarrhea in Afghanistan relied on statistical methods. They also focused their work only in medical-related data. To date, there has not been an attempt to implement a predictive diarrhea model based on both non-medical related data (for example, wealth status) and medical-related data.

IV. METHODOLOGY

Data mining is responsible for discovering unknown and secret patterns in a large amount of data to obtain valuable information. As stated earlier, one of the objectives of this work is to find a suitable classifier among the most popular machine learning techniques. The classification task in this study is binary, where two categories are considered. 'No Diarrhea' indicates that the sample belongs to the category where the infant did not have diarrhea, and the 'Diarrhea' category indicates that the infant had diarrhea.

Fig. 1. A depiction of the overall implementation of the predictive model.

A. Data Understanding

The dataset in this study is the publicly available Afghanistan Demographic and Health Survey 2015 (AfDHS). It contains information from a cross-sectional, population-based, nationally representative survey carried out by the Central Statistics Organization and the Ministry of Public Health in Afghanistan's rural and urban areas in 2015. Altogether, the AfDHS comprises information from 25,650 households where some households contributed more than one sample to the dataset (i.e., more than one record relating more than one child). A detailed description of data in this survey can be found in [19]. The dataset may be obtained from the Demographic and Health Surveys (DHS) website.

B. Data Preprocessing

Like most data mining process, the initial step of data preprocessing is also necessary for this study. This consists of the main two stages, data cleansing and feature selection, as described below:

1) Data Cleansing

Initially, the AfDHS data comprises 32,712 records (i.e., samples). The sample excluded children whose mothers did not know if they had diarrhea in the past two weeks and those whose data were missing. At the end, 5,800 complete and useable samples remained. Table I reports the number of samples in both categories.

TABLE I. NUMBER OF SAMPLES IN BOTH CATEGORIES.

No. in diarrhea category	No. in the non-diarrhea category	No. of total samples
2900	2900	5800

2) Feature Selection

Each sample in the dataset consists of 1,187 attributes. Careful analysis reveals that most attributes are not related and inappropriate for the implementations. Such attributes include the respondent's line number, date of interview, mother names, etc. Therefore, an essential step in this initial

stage is to identify relevant attributes related to diarrhea. This is carried out by consultations with doctors and experts in this area. Different doctors and experts had expressed a different number of attributes that contribute to diarrhea. The study identifies common attributes among these doctors and experts. Finally, it was found that 21 attributes are commonly express among them. Table II reports these 21 attributes, as shown below.

TABLE II. SELECTED ATTRIBUTES USED IN THIS STUDY

No.	Name	Description
1	Region	Region of residence
2	Mother's Education	The education level of the mother
3	Water source	Source of drinking water
4	Toilet type	Type of toilet facility
5	No. of household	Number of household members
6	No. of children	Number of children under five years
7	Cooking fuel	Type of cooking fuel
8	Plain water	Given child plain water
9	Powder or fresh milk	Given child powder or fresh milk
10	Disposal methods	Type of disposal methods
11	Father's Education	Education level of father
12	Child age	Current age of child
13	PBI	Preceding birth interval
14	Breastfeeding	Duration of breastfeeding
15	Given anything	Given child anything else than breast milk
16	Fever	Had child fever?
17	Place of residence	Place of residence (rural and urban)
18	Gender	Sex of child (male or female)
19	Wealth	The economic situation of the family
20	Breastfeeding	Currently child breastfeeding
21	Occupation	Occupation of the mother

Feature selection is an essential factor in the data mining process' success and selecting informative or relevant attributes. Attribute/Feature selection methods are used to reduce the dimensionality of data by removing the redundant and irrelevant attributes in a data set. The feature selection process has many benefits. It allows visualization and understanding of data more quickly, reduces the time and storage required for the mining process, and improves the algorithm's performance by avoiding the curse of dimensionality. Performance is enhanced by eliminating features that do not add any value to the algorithm's efficiency. This study used Information Gain and correlation-based feature selection algorithm (CFS) algorithms to reduce dimensionality in the datasets.

Information Gain (IG) is an entropy-based feature evaluation method widely used in machine learning. As Information Gain is used in feature selection, it is defined as the amount of information provided by the text category's feature items. Information gain is calculated by how much of a term can be used to classify information to measure the importance of lexical items for the classification in natural language processing [20]. The formula of the information gain is shown in equation (1).

$$\begin{aligned}
IG(t) = & - \sum_{i=1}^m P(C_i) \log P(C_i) \\
& + P(t) \sum_{i=1}^m P(C_i | t) \log P(C_i | t) \\
& + P(\bar{t}) \sum_{i=1}^m P(C_i | \bar{t}) \log P(C_i | \bar{t}) \quad (1)
\end{aligned}$$

C is a set of document collection, in which there are various features or terms t. Such collection would have m documents. The IG value depends on the probability and conditional probability of each document as indicated. When the IG(t) value is greater, it is more useful for the classification for C. Then this term t should be selected. The value of IG(t) depends on the value of P(t) and P($\bar{t}$).

The CFS algorithm computes the correlation between all features and the output class. It selects the best feature subset (i.e., the subset with features positively correlated with the class variable and has a low correlation with each other features) using a correlation-based heuristic evaluation function [21]. CFS is calculated by equation 2.

$$r_{zc} = \frac{k\bar{r}_{zi}}{\sqrt{k + k(k-1)\bar{r}_{ii}}} \quad (2)$$

where

r_{zc} : correlation between features and the class variable.

k: number of features.

$\bar{r}_{zi}$: the average of the correlation between feature-class.

$\bar{r}_{ii}$: average inter-correlation between features.

CFS and Info Gain were applied to the datasets to determine the informative and high correlated attributes. Tables III and VI reveal the results of the CFS and Info Gain application to the datasets.

TABLE III. ATTRIBUTES SELECTED BY INFO GAIN.

No.	Name	Info Gain values
1	Region of residence	0.316
2	Had fever	0.085
3	Given Plain water	0.039
4	The education level of the mother	0.017
5	Type of cooking fuel	0.014
6	Type of toilet facility	0.013
7	Type of disposal method	0.012
8	Given child anything else than breast milk	0.0098
9	Age of child	0.0084
10	Given powder or fresh milk	0.0082
11	Wealth status	0.0079
12	Place of residence (urban or rural)	0.0069
13	No. of household members	0.0045
14	The education level of the father	0.0037
15	No. of children under five years	0.0036
16	Source of drinking water	0.0026

TABLE IV. ATTRIBUTES SELECTED BY CORRELATION-BASED FEATURE SELECTION.

No.	Name	CFS values
1	Region of residence	0.423
2	Given plain water	0.227
3	Had fever	0.179
4	Given child anything else than breast milk	0.122
5	Age of child	0.111
6	Given powder or fresh milk	0.094
7	Type of disposal method	0.085
8	No. of household members	0.082
9	No. of children under five years	0.065
10	Type of toilet facility	0.053
11	Duration of breastfeeding	0.046
12	Type of cooking fuel	0.038
13	The education level of the father	0.037
14	The education level of the mother	0.037
15	Source of drinking mother	0.017
16	Preceding birth interval	0.016

Referring to Tables III & IV, the different attributes selected by Information Gain are the wealth status and place of residence (rural & urban) attributes. Preceding birth interval and duration of breastfeeding are the different attribute selected by CFS. These differences merit further analysis. Therefore, model building was done using both Info Gain and CFS selected attributes. Tables V and VI reveal the accuracy measures of them.

C. Model Building

On the cleaned dataset with sixteen selected attributes by Info Gain and CFS, three classification algorithms, namely Naïve Bayes, Random Forest, and Support Vector Machine (SVM), were applied using the WEKA [22] machine learning tool on the datasets. The datasets were balanced, where the classification categories were approximately equally represented, and the dataset split into 85% training and 15% testing data.

D. Evaluation

To evaluate the performance and effectiveness of the predictive model implemented, evaluation metrics are used. In this study, four commonly used metrics are employed [23]. These are 'Accuracy', 'Precision' Recall', and 'Area Under the Curve (AUC)'. Their brief descriptions are given below:

1) Accuracy

This metric is most used in classification as it is the first important indicator of how well the model performs. It is converted to percentage usually (i.e., 0% to 100%). It can be determined by the equation below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

TP : the number of true positives (samples that are correctly classified in their correct class).

TN : the number of true negatives (samples correctly classified that they do not belong to the target class).

FP: the number of false positives (samples incorrectly labeled as the target class when they are not).

FN: the number of false negatives (samples that are incorrectly labeled as not the target class while they are).

2) Precision

The information retrieval field introduced this metric; however, it has an application in classification, and it is a useful addition in evaluating the performance. It is also expressed in percentage. It can be determined by the equation below:

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

3) Recall

The Recall metric is often used in conjunction with the Precision metric in the information retrieval field. Hence, it can add useful information in evaluating the performance. It is also expressed in percentage. It can be determined by the equation below:

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

4) Area Under the Curve (AUC)

AUC is the area under the receiver operating characteristic curve (ROC). It provides a comprehensive assessment of a model's accuracy by screening the range of threshold values for decision making. The larger area, the more accurate the diagnostic test is. It is also expressed in percentage.

V. RESULTS AND DISCUSSION

Predictive model building is an iterative procedure, and, therefore, it is crucial to perform multiple experiments with different classifiers to select the best model for solving the problem at hand. The experiments were conducted by using three popular classifiers, namely: Naïve Bayes, Random Forest, and Support Vector Machine, on the balanced datasets. Experimentations were carried out for each algorithm, and the percentage split validation was used to estimate each algorithm's performance. Out of all the instances, 85% was utilized for training, and the remaining 15% made up the test set. The data mining algorithm performance was evaluated based on accuracy, precision, recall and the area under the curve (AUC). The AUC serves as an indicator for the overall performance of the algorithm [23, 24]. The area under the receiver operating characteristics curve technique provides a comprehensive assessment of the performance of a predictor of screening the range of threshold values for the decision making. The larger area, the more reliable the diagnostic test is. Therefore, models with the highest accuracy and AUC are deemed as the best.

TABLE V. PERFORMANCE MEASURES OF THREE DATA MINING ALGORITHMS USING ATTRIBUTES SELECTED BY CORRELATION-BASED FEATURE SELECTION.

Classifier	Accuracy	AUC	Precision	Recall
Naïve Bayes	69.10 %	76.70%	69.1%	69.1%
Random Forest	81.48 %	89.80%	82.0%	81.4%
SupportVectorMachine	73.22 %	73.60%	74.8%	73.2%

Fig. 2. Performance measures of three data mining algorithms using attributes selected by correlation-based feature selection

TABLE VI. PERFORMANCE MEASURES OF THREE DATA MINING ALGORITHMS USING ATTRIBUTES SELECTED BY INFO GAIN.

Classifier	Accuracy	AUC	Precision	Recall
Naïve Bayes	69.85%	75.9%	70.3%	69.9%
Random Forest	77.10%	86.4%	77.3%	77.1%
SupportVectorMachine	72.61 %	72.5%	72.7%	72.6%

Fig. 3. Performance measures of three data mining algorithms using attributes selected by Info Gain

Referring to Tables V and VI, it can be seen that the random forest algorithm has higher accuracy, AUC, precision, and recall with the selected attributes by correlation-based feature selection (CFS). But in contrast it has an overall lower accuracy, AUC, precision, and recall with attributes selected by Information Gain, therefore this finding showed that CFS application is suitable with the nature of this dataset.

For the prediction of diarrhea in children under five years in Afghanistan, it can be seen that random forest gives a higher overall prediction accuracy of 81.48%, AUC 89.80%, precision 82%, and recall 81.4% compared to other classification methods. Also, the area under the curve (AUC) is identified in the random forest to be higher than other methods, indicating that random forest is the best model to determine the children's diarrhea in this study. On the other end, Naïve Bayes gave the lowest overall prediction accuracy of 69.10% in this study.

As this work only used a subset of the collected data and preprocessed it to be a balanced dataset, validation may be done using the remaining data in the next step.

VI. CONCLUSIONS AND FUTURE WORK

Diarrhea is the second leading cause of infant mortality after pneumonia in children under five years. To better understand the crucial factors positively associated with diarrhea and having the most suitable predictive model for Diarrhea in Afghanistan, it is very beneficial to mitigate the problem mentioned. This study is the first attempt to determine the most suitable tool to implement the diarrhea predictive model. The study applied Info Gain and CFS in feature selection and identified the attributes with high correlation/more informative attributes. The different attributes selected by Info Gain were wealth status and place of residence, but in contrast, different selected attributes by CFS were preceding birth interval and duration of breastfeeding. Implementation of the model was done using attributes selected by CFS and Info Gain. Random Forest was found as the most suitable tool amongst the three popular data mining techniques consider, with Accuracy, AUC, Precision, and Recall of 81.5, 89.8, 82, and 81.4%.

Future study can be carried out which adopts 16 attributes and relies on CFS to identify the importance of each selected attribute. More detailed analysis can be carried out to ensure each attribute's influence in diarrhea. Model validation will then be done using the remaining data extracted from the survey. Furthermore, a comparative analysis using the full imbalanced data may be undertaken using the approach by Momand et al. [25] and may incorporate this work into their malnutrition application developed for the Afghan government for detection and monitoring of the health status of pre-school age children.

ACKNOWLEDGMENT

The authors wish to express their gratitude to the Higher Education Development Program (HEDP), Ministry of Higher Education of Afghanistan, and King Mongkut's University of Technology Thonburi (KMUTT), Thailand, for the Scholarship support of Mr. Abdullah Zahirzada. The moral support from Assoc. Prof. Dr. K. Lavangnananda during preparation of this paper is much appreciated.

REFERENCES

- [1] H. Hussein, "Prevalence of Diarrhea and associated risk factors in children under five years of age in Northern Nigeria: A secondary data analysis of Nigeria and Demographic and Health Survey 2013," 2017.
- [2] A. Sanyaolu, R. Groetz, J. Gillam, P. Patel, and O. Oyeleke, "Global trends of diarrhea diseases in children," Cloudfront.net. [Online]. Available: https://d1wqtxts1xzle7.cloudfront.net/55864093/Diarrhoea_diseases_publication.pdf?1519233841=&response-content-disposition. [Accessed: 29-Nov-2020].
- [3] GBD 2016 Diarrhoeal Disease Collaborators, "Estimates of the global, regional, and national morbidity, mortality, and aetiologies of diarrhoea in 195 countries: a systematic analysis for the Global Burden of Disease Study 2016," *Lancet Infect. Dis.*, vol. 18, no. 11, pp. 1211–1228, 2018.
- [4] D. Nanani, F. White, I. Azam, H. Afsar, and S. Hozhabri, "Evaluation of a water, sanitation, and hygiene education intervention on diarrhoea in northern Pakistan," *Bull. World Health Organ.*, vol. 81, no. 3, pp. 160–165, 2003.
- [5] Z. A. Bhutta et al., "Interventions to address deaths from childhood pneumonia and diarrhoea equitably: what works and at what cost?," *Lancet*, vol. 381, no. 9875, pp. 1417–1429, 2013.
- [6] A. Weisz, G. Meuli, C. Thakwalakwa, I. Trehan, K. Maleta, and M. Manary, "The duration of diarrhea and fever is associated with growth faltering in rural Malawian children aged 6-18 months," *Nutr. J.*, vol. 10, no. 1, p. 25, 2011.
- [7] A. Bhardwaj, "Identification and optimization of ISSR primers for analysis of genetic variability among different landrace collections of Indian sponge gourd (*Luffa cylindrica* L.)," *Plant Physiol. Rep.*, vol. 24, no. 3, pp. 296–304, 2019.
- [8] H. Kaur and S. K. Wasan, "Empirical study on applications of data mining techniques in healthcare," *J. Comput. Sci.*, vol. 2, no. 2, pp. 194–200, 2006.
- [9] S. Gupta, "Data mining classification techniques applied for breast cancer diagnosis and prognosis," *Psu.edu*. [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.300.5279&rep=rep1&type=pdf>. [Accessed: 29-Nov-2020].
- [10] D. Senthilkumar, "Prediction of low birth weight infants and its risk factors using data mining techniques," *Ieomsociety.org*. [Online]. Available: http://ieomsociety.org/ieom_2015/papers/134.pdf.
- [11] Encyclopedia of bioinformatics and Computational Biology: ABC of bioinformatics. Elsevier, 2018.
- [12] P. Roßbach, "Neural Networks vs. Random Forests – Does it always have to be Deep Learning?," *Frankfurt-school.de*. [Online]. Available: <https://blog.frankfurt-school.de/wp-content/uploads/2018/10/Neural-Networks-vs-Random-Forests.pdf>. [Accessed: 29-Nov-2020].
- [13] M. W. Ahmad, M. Mourshed, and Y. Rezgui, "Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption," *Energy Build.*, vol. 147, pp. 77–89, 2017.
- [14] "ShieldSquareCaptcha," *Iop.org*. [Online]. Available: <https://iopscience.iop.org/article/10.1088/1742-6596/1282/1/012010/meta>. [Accessed: 29-Nov-2020].
- [15] C. M. George et al., "Risk factors for diarrhea in children under five doneyears of age residing in peri-urban communities in Cochabamba, Bolivia," *Am. J. Trop. Med. Hyg.*, vol. 91, no. 6, pp. 1190–1196, 2014.
- [16] R. Li, Y. Lai, C. Feng, R. Dev, Y. Wang, and Y. Hao, "Diarrhea in under five year-old children in Nepal: A spatiotemporal analysis based on Demographic and health survey data," *Int. J. Environ. Res. Public Health*, vol. 17, no. 6, p. 2140, 2020.
- [17] B. Melese, W. Paulos, F. H. Astawesegn, and T. B. Gelgelu, "Prevalence of diarrheal diseases and associated factors among under-five children in Dale District, Sidama zone, Southern Ethiopia: a cross-sectional study," *BMC Public Health*, vol. 19, no. 1, p. 1235, 2019.
- [18] J. Gascón et al., "Diarrhea in children under 5 years of age from Ifakara, Tanzania: a case-control study," *J. Clin. Microbiol.*, vol. 38, no. 12, pp. 4459–4462, 2000.
- [19] Central Statistics Organization/Afghanistan, Ministry of Public Health/Afghanistan, and ICF. 2017. Afghanistan Demographic and Health Survey 2015. Kabul, Afghanistan: Central Statistics Organization. [Assessed: 24-Mar-2020]
Available at <https://dhsprogram.com/pubs/pdf/FR323.pdf>
- [20] T. A. Alhaj, M. M. Siraj, A. Zainal, H. T. Elshoush, and F. Elhaj, "Feature selection using information gain for improved structural-based alert correlation," *PLoS One*, vol. 11, no. 11, p. e0166017, 2016.
- [21] K. B. S. Al Janabi, "Data reduction techniques: A comparative study for attribute selection methods," *Ripublication.com*. [Online]. Available: https://www.riplication.com/ijacst18/ijacstv8n1_01.pdf. [Accessed: 29-Nov-2020].
- [22] E. Frank, M. A. Hall, and I. H. Witten, In: *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*, Morgan Kaufmann, Fourth Edition, 2016.
- [23] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International Journal of Data Mining & Knowledge Management Process* 5, no. 2 (2015): 1.
- [24] K. J. Danjuma, "Performance evaluation of machine learning algorithms in post-operative life expectancy in the lung cancer patients," *arXiv [cs.LG]*, 2015.
- [25] Z. Momand, P. Mongkolnam, P. Kositpanthavong and J. H. Chan, "Data Mining Based Prediction of Malnutrition in Afghan Children," 2020 12th International Conference on Knowledge and Smart Technology (KST), Pattaya, Chonburi, Thailand, 2020, pp. 12-17, doi: 10.1109/KST48564.2020.9059388.