FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 7 No. 3, June (Special Issue), 2023, pp 288 - 292
DOI: https://doi.org/10.33003/fjs-2023-0703-801

ENHANCED PREDICTIVE MODEL FOR SCHISTOSOMIASIS

Salisu Ahmad, *Umar Iliyasu and Bashir Ahmed Jamilu

Department of Computer Science, Federal University Dutsin-Ma, PMB 5001, Katsina State Nigeria

*Corresponding authors' email: umariliyasut@gmail.com

ABSTRACT

Neglected Tropical Diseases (NTDs) are wide spread diseases found in many countries in Africa, Asia and Latin America, they are mostly found in tropical areas where people have no access to clean water or safer ways to dispose of human waste. Schistosomiasis is one of the NTDs. Data mining is used in extracting rules to predict certain information in many areas of Information Technology, medical science, biology, education, and human resources. Classification is one of the techniques of Data mining. In this work, we used three classifiers namely; Naïve Bayes, Support Vector Machine and Logistic Regression to design a framework for classifying and predicting the status of Schistosomiasis and its complications in a suspected patient using their clinical data. For the purpose of this study, we considered the parameters: Abdominal, Diarrhea, Bloody_stool, Bloody urine, Swim, Dam_river_ use, Urinating_stool_in_water, Boil_water_use. The framework was trained using data acquired from Federal Medical Centre Katsina and NTD unit of Katsina State ministry of health, to test for performance accuracy. The research shown that out of the three classifiers, Logistic Regression performed better by having 97.8% accuracy.

Keywords: Neglected Tropical Diseases (NTDs), Schistosomiasis, Naïve Bayes, Support Vector Machine Logistic Regression

INTRODUCTION

Schistosomiasis, also known as bilharzia was named after Theodor Bilharz, who was the first to identify the parasite in Cairo in 1851. Infection is widespread with a high morbidity rate, causing severe debilitating illness in millions of people. The disease is often associated with water resource developmental projects such as dams, streams and irrigation schemes where the snail is the intermediate hosts of the parasite breed (WHO, 2010). Schistosomiasis is a parasitic disease that is chronic caused by blood flukes (Trematode worms) of genus schistosoma. Out of the total world population, it was estimated that 218 million people suffer from schistosomiasis of which approximately 90% live in Africa (WHO 2017). Two-thirds of these cases are caused by Schistosoma haematobium, the etiologic agent of UGS (Van *et. al.*, 2003). The potential consequences of S. haematobium infection include haematuria, dysuria, nutritional deficiencies, lesions of the urinary bladder, hydronephrosis, stunting (in children) (Saathoff, 2004) and in adults, infertility, cancer, and increased susceptibility to HIV (King *et. al.*, 2008).

A technique called Predictive Analysis incorporates a variety of machine learning algorithms, data mining techniques and statistical methods that uses current and past data to find knowledge and predict future events. By applying predictive analysis on healthcare or clinical data, significant decisions are taken and predictions are be made. Predictive analytics can be done using machine learning and regression techniques. Predictive analytics aims at diagnosing the disease with best possible accuracy, enhancing patient care, optimizing resources along with improving clinical outcomes. Machine learning is considered to be one of the most important artificial intelligence features that supports development of computer systems having the ability to acquire knowledge from past experiences with no need of programming for every case (Gauri *et. al.*, 2017). Existing method for Schistosomiasis detection is through culturing of suspected patients' stool sample in microbial lab for at least a week. However, this method is time wasting and consuming. This research work focuses on building predictive model using machine learning algorithms and data mining techniques for accurate Schistosomiasis prediction.

Related Works

Li *et. al.* (2018) studied the use of an ANN model with a standard feed-forward back propagation (BP) network structure including an input layer of 16 neurons, a hidden layer of 20 neurons and an output layer of 2 neurons to predict the prognosis of patients with advanced schistosomiasis. Sigmoid transfer functions were applied to the hidden and output layers. Gradient descent was used to calculate the synaptic weights. The initial learning rate was defined as 0.07 and the momentum was 0.95. The batch size was defined as 256 and the number of iterations was 200. Ten-fold cross validation was employed. However, there is currently no accepted theory that predetermines the optimal number of hidden layer neurons; the numbers of hidden layer neurons were determined by repeated trial and error test until the best sensitivity and specificity was achieved.

Kavakiotis *et. al.* (2017) used 10 fold cross validation as an evaluation method in three different algorithms, including Logistic regression, Naive Bayes and SVM, where SVM provided better performance and accuracy of 84 % than other algorithms.

Kandhasamy *et. al.* (2015) applied KNN, J48, SVM and Random Forest, where J48 machine learning algorithm provides better performance and accuracy than others before preprocessing technique. The classification algorithms did not evaluate using cross validation method.

Meng *et. al.* (2013) used three different data mining techniques: ANN, Logistic regression and J48 to predict the diseases using real world data sets by collecting information through distributed questionnaire. Finally, it was concluded that J48 machine learning techniques provided efficient and better accuracy than others.

Raj and Prasanna (2013) proposed an automatic disease identification model which could be converted into an integrated model to improve on text classification based on Machine Learning principle. The approach uses Natural

Language Processing and Naïve Bayes technique of Machine Learning (ML) and the diseases considered in the study includes; Malaria, Typhoid, Dengue, Tuberculosis and Hepatitis B. However, this system dwell on Medline text as target parameter and does not consider checking accuracy as means of authenticating model.

Masethe and Masethe (2014) conducted an experiment on early detection and prediction of heart disease using different classifications techniques such as Naïve Bayes, J48, CART, Bayesian Network and REPTREE. The study uses k-means clustering on a dataset from South Africa and demonstrates the prototype using Naïve Bayes to predict the chances of the patient getting a heart attack. The result shows a prediction accuracy of 97% using confusion matrix.

Asarnow and Singh (2018) applied Asarnow-Singh algorithm to identify s.mansoni through extraction of images' features by defining threshold values on image's infected foreground and background parasitic areas. Further, support vector machine (SVM) evaluated into training and testing record which provides effective results.

Ashour *et. al.* (2018) specified level of schistosomiasis images and extracted features by giving statistical names of legion area. Further, evaluates SVM, KNN, DT which shows that linear discriminant SVM classifier results are better than quadratic SVM.

Li *et. al.* (2018) classified into two groups (poor and advanced schistosomiasis). Moreover, they divided data into two groups as training and testing record and applied ANN, DT and LR to determine the best results among them. Confusion matrix indicated the best prediction of disease is performed by ANN.

MATERIALS AND METHODS

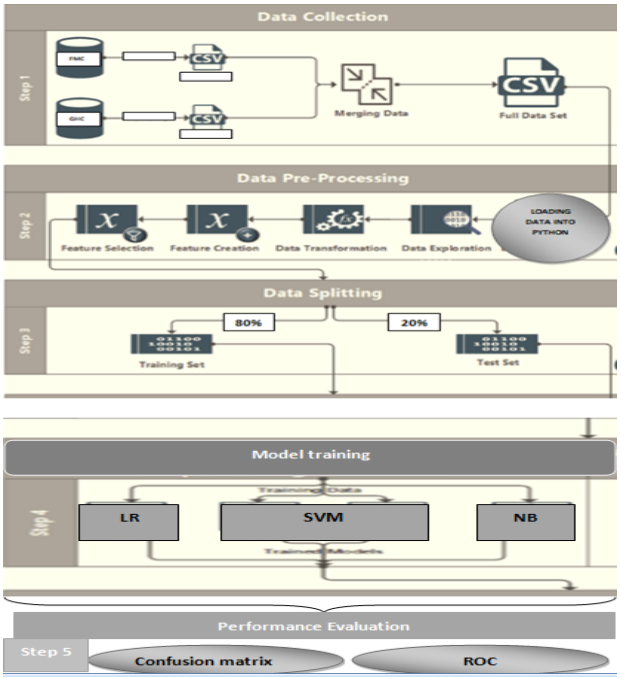

Figure 1: Flow Diagram of Network Design

A framework is shown above that explains the sequence involved in the network design. In principle, data is collected from Federal Medical Centre, Katsina and Hellen Keller International, Katsina.

It is then preprocessed to solve for the problem of missing values, discretized and label encoded. The network is then built, trained and tested. The flow diagram of the network is as shown in Figure 1 above.

Schistosomiasis Disease Dataset

This research uses the dataset provided by the Federal Medical Centre and Hellen Keller International, Katsina. The data is a schistosomiasis disease dataset that consists of different samples with 8 attributes; 8 numeric inputs named Abdominal, Diarrhea, Bloody_stool, Bloody urine, Swim, Dam_river_ use, Urinating_stool_in_water, Boil_water_use and one output attribute named status. Status contains positive or negative. Table I shows the detailed description of the Schistosomiasis dataset used.

Table 1: Attributes of Schistosomiasis dataset Attribute Description

S/N	ATTRIBUTE	DESCRIPTION	DOMAIN OF VALUE	
1	Abdominal	Do you have abdominal pain?	1 = Yes	0 = No
2	Diarrhea	Do you suffer from diarrhea?	1 = Yes	0 = No
3	Bloody_stool	Is there blood in your stool?	1 = Yes	0 = No
4	Bloody urine	Is there blood in your urine?	1 = Yes	0 = No
5	Swim	Do you frequently swim?	1 = Yes	0 = No
6	Dam_river_ use	Do you use dam or river water for domestic purposes?	1 = Yes	0 = No
7	Urinating_stool_in_water	Do you stool or urinate in dam or river?	1 = Yes	0 = No

8	Boil_water_use	Do you boil water before use?	1 = Yes	0 = No
9	Status	The status of patient	1 = Yes	0 = No

Data Preprocessing

This is the next step in machine learning after data collection. Some of the issues that need to be addressed before any further analysis is making sure that the data is clean without noise or missing values and it is scaled. Although data analysts are continually trying to improve the robustness of machine learning algorithms to be capable of high performance in the presence of missing values or noise, the quality of the results is still affected by the input data.

Missing Value Imputation

We used imputation methods to replace the missing values with new data systematically. Imputing missing values allows us to consider more features rather than removing all the observations with missing values. Imputing missing values keeps the full data set and avoids biased results.

Feature Selection

Feature selection methods provide a subset of the full-size data in which only the relevant features are selected. Dimensionality reduction approaches were used because they are among the most frequently used techniques in machine learning. These techniques can be divided into two categories: feature selection and feature extraction. Learning from a smaller subset not only increases the learning speed and makes the process less computationally expensive, it leads to better performance with improved learning accuracy and the model is more interpretable. Apart from feature selection, feature extraction techniques such as Principal Component Analysis try to reduce the dimensionality by creating a new set of features that can capture the variations in the data and reduce the dimensionality without compromising the performance of the classification algorithms. Both feature selection and feature extraction techniques are essential steps in preparing the data for classification. They enhance the performance of the classifiers and decrease the computational complexity which reduces the time and storage required to build and run the model. In general, feature selection methods are divided into three categories: filter method, wrapper method and embedded methods. However, we used filter method and wrapper methods for effective selection.

Data Splitting

After completing the preprocessing task, the datasets were split into two datasets, training and test sets. The training sets were used to construct the models, while the test set were used to evaluate the performance of the models. In this phase, 80% of the data was allocated to the training set, and the remaining 20% was allocated to the test set.

Classification Algorithms

Three core classifiers were applied to assess the predictive performance for labelling of SD classes including Naïve Bayes, Support Vector Machine and Logistic regression. These were selected due to the positive results received when applied to the disease prediction problem domain and also due to the variety in each approach which yields greater balance for experiment conditions. The classification is focused on three SD classes of 'Low', and 'High' making it a binary dense analysis problem that is used to provide indication of disease likelihood rather than multi variant or not for improving disease likelihood analysis. If these algorithms perform well during testing then they can be considered for modifying to increase classification effectiveness with this research. Classifiers used in each of the experiments in this study assess the distribution and density of schistosomiasis vector levels.

.

Naïve Bayes: is a probabilistic algorithm which aims to classify data instances without bias based on the vector class properties. The Naïve Bayes classifier was found to provide consistent performance across the prediction domain. However, it worked well but did not achieve the highest classification accuracy results when compared to Support Vector Machine.

Support Vector Machine: Support Vector Machine classification splits the data using a hyper plane which then deduces the class and instance it should reside in. Support Vector Machine has shown to provide increased classification accuracy over Naïve Bayes in disease prediction research. Modified versions of SVM have been widely used with success in the area of disease predictions in epidemiology studies thus, it is deemed suitable and was applied in our study together with the previous method. Experiment results show that SVM perform well when compared with the other selected algorithms in many instances with higher classification accuracy percentages.

RESULTS AND DISCUSSION

Accuracy: Accuracy is defined as the ratio of the number of correctly classified instances to all the cases. It is equal to the sum of TP and TN divided by the total number of instances.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision: Precision is defined as the proportion of true positive instances which are classified as positive. It shows how close predicted values are to each other (Max, 2013).

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall: Recall is defined as the proportion of positive instances that are correctly classified as positive. Recall is also often called sensitivity.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

F1 Score: F1 score is a measure that combines both precision and recall and tries to find a balance between both.

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

Classification Report

Logistic Regression Classification Report

Here, the precision value when false is 0.95, recall is 1.00, F1-Score is 0.98 and the support is 191, but when true, i.e. status is positive, precision is 1.00, recall is 0.96, f1-score is 0.98 while the support is 229, for the accuracy f1-score is having 0.98 while the Macro average and weighted average are also 0.98

Table 2: Logistic Regression classification report table

Report	Precision	Recall	F1-score	Support
0	0.95	1.00	0.98	191
1	1.00	0.96	0.98	229
Accuracy			0.98	420
Macro avg	0.98	0.98	0.98	420
Weighted avg	0.98	0.98	0.98	420

Naive Bayes Classification Report

Here, the precision value when false is 0.88, recall is 0.96, F1-Score is 0.91 and the support is 191, but when true, i.e. status is positive, precision is 0.96, recall is 0.89, f1-score is 0.92 while the support is 229, for the accuracy f1-score is having 0.92 while the Macro average and weighted average are also 0.92

Table 3: Naïve Bayes Classification report table

Report	Precision	Recall	F1-score	Support
0	0.88	0.96	0.91	191
1	0.96	0.89	0.92	229
Accuracy			0.92	420
Macro avg	0.92	0.92	0.92	420
Weighted avg	0.92	0.92	0.92	420

Support Vector Machine Classification Report

Here, the precision value when false is 0.95, recall is 1.00, F1-Score is 0.98 and the support is 191, but when true, i.e. status is positive, precision is 1.00, recall is 0.96, f1-score is 0.98 while the support is 229, for the accuracy f1-score is having 0.98 while the Macro average and weighted average are also 0.98

.

Table 4: Support Vector Machine Classification Report table

Report	Precision	Recall	F1-score	Support
0	0.95	1.00	0.98	191
1	1.00	0.96	0.98	229
Accuracy			0.98	420
Macro avg	0.98	0.98	0.98	420
Weighted avg	0.98	0.98	0.98	420

In this work we developed and presented a predictive model for schistosomiasis disease that is capable of accurately classifying humans with Schistosomiasis larval parasites. The accuracy of our model matched and after using the classifiers, it was discovered that the model that performed most accurately is Logistic regression out of the three classifiers, where result shows that Logistic Regression gave the accuracy of 97.85714285714285, Naive Bayes gave the accuracy value of 91.9047619047619 while SVM displayed no skill.

CONCLUSION

In conclusion, this research work demonstrates the predictive model for classifying schistosomiasis disease in patients. Computer vision model has been applied using three classifiers which are Support Vector Machine, logistic regression and Naïve Bayes. Thorough cleaning and feature selection were done so as to have improved and accurate result. The results obtained reveal that out of the three models, Logistic Regression performed more accurate by considering precision, accuracy and F1 score.

REFERENCES

Asarnow, D. and Singh, R. (2018) "Determining Dose-Response Characteristics of Molecular Perturbations in Whole Organism Assays Using Biological Imaging and Machine Learning." IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE., pp. 283-290

Ashour, A. S., Hawas, A. R. and Guo, Y. (2018) "Comparative study of multiclass classification methods on light microscopic images for hepatic schistosomiasis fibrosis diagnosis.," Health information science and systems, vol. 6, no. 1, pp. 1-12.

Gauri, D. K., Shivananda, R. P. and Nagaraj, V. D. (2017)" Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop", International Conference On I-SMAC,978-1-5090-3243-3

Kandhasamy, J. P. and Balamurali, S. (2015) "Performance analysis of classifier models to predict diabetes mellitus." Procedia Computer Science 47 pp. 45-51

Kavakiotis, I., Olga, T., Athanasios, S., Nicos, M., Ioannis, V. and Ioanna, C. (2017), "Machine learning and data mining methods in diabetes research." Computational and Structural Biotechnology Journal

King C. H. and Dangerfield-Cha, M. (2008) "The unacknowledged impact of chronic schistosomiasis," Chronic Illness , vol. 4, no. 1, pp. 65–79.

Li, G., Zhou, X., Liu, J., Chen, Y., Zhang, H. and Chen, Y. (2018) Comparison of three data mining models for prediction of advanced schistosomiasis prognosis in the Hubei province. PLoS Negl Trop Dis 12(2): e0006262. https://doi. org/10.1371/journal.pntd.0006262

Li, G. X., Zhou, J., Liu, Y., Chen, H., Zhang, Y., Chen, J., Liu, H., Jiang, J., Yang, A. and Nie, S. (2018) "Comparison of three data mining models for prediction of advanced schistosomiasis prognosis in the Hubei province.," PLoS neglected tropical diseases, vol. 12, no. 2, pp. 1 19.

Masethe, H. D. and Masethe, M. A. (2014) "Prediction of heart disease using classification algorithms". In Proceedings of the world Congress on Engineering andcomputer Science (Vol. 2, p. 2224)

Meng, X. H., Huang, Y. X., Rao, D. P., Zhang, Q. and Liu, Q. (2013). "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors". The Kaohsiung journal of medical sciences, 29(2), 93-99.

Raj, T. F. M. and Prasanna, S. (2013) "Implementation of ML using naïve bayes algorithm for identifying disease-treatment relation in bio-science text". Research Journal of AppliedSciences, Engineering and Technology, 5(2), 421–426

Ramya S. and Radha N. (2016) Diagnosis of Chronic Kidney Disease using ML Algorithm. International Journal of Innovative Research in Computer and Communication Engineering. Vol. 4, Issues 1,

Ranhotra, S. S. (2017) "An alternative approach to detect the presence of schistosoma haematobium infection in affected regions of benue state-Nigeria.," IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI-IEEE)., pp. 2113- 2117

Rani, A. S. and Jyothi, S. (2016) "Performance analysis of classification algorithms under different datasets." In Computing for Sustainable Global Development (INDIACom), 3rd IEEE International Conference, pp. 1584-1589.,.

Saathoff, E. A. O., Magnussen, P., Kvalsvig, J. D., Becker, W. and Appleton, C. C. () "Patterns of Schistosoma haematobium infection, impact of praziquantel treatment and re-infection after treatment in a cohort of school children from rural KwaZulu-Natal/South Africa", BMC Infectious Diseases, vol. 4, article 40

Van, M. J., Der, W., De Vlas, S. J. and Brooker, S. (2003) "Quantification of clinical morbidity associated with schistosome infection in sub-Saharan Africa," Acta Tropica , vol. 86, no. 2-3, pp. 125– 139

WHO (2017) "Schistosomiasis," Fact Sheet No. 115, Geneva, World Health Organisation, http://www.who.int/mediacentre/factsheets/fs115/

World Health Organization Expert Committee. (2010): "Prevention and control of schistosomiasis and soil transmitted helminthiasis". WHO Technical Report Series No. 912, World Health Organization, Geneva, Switzerland: WHO.