Journal of Advances in Mathematics and Computer Science

31(4): 1-10, 2019; Article no.JAMCS.39029

ISSN: 2456-9968

(Past name: British Journal of Mathematics & Computer Science, Past ISSN: 2231-0851)

Predicting Paediatric Malaria Occurrence Using Classification Algorithm in Data Mining

T. C. Olayinka[1*] and S. C. Chiemeke[2]

[1]Department of Computer Science, Wellspring University, P.M.B. 1230, Benin City, Edo State, Nigeria.
[2]Department of Computer Science, University of Benin, Benin City, Edo State, Nigeria.

Authors' contributions

This work was carried out in collaboration between both authors. Author TCO carried out the research work. Author SCC supervised the entire research. Both authors read and approved the final manuscript.

Article Information

DOI: 10.9734/JAMCS/2019/v31i430118
Editor(s):
(1) Dr. Kai-Long Hsiao, Associate Professor, Taiwan Shoufu University, Taiwan.
Reviewers:
(1) Dr. Bukola Badeji – Ajisafe, University of Medical Science, Ondo State, Nigeria.
(2) Dr. A. V. Krishna Prasad, Osmania University, Telangana, India.
(3) Dr. Manish Mahajan, CGC College of Engineering, Landran, Mohali, Punjab, India.
(4) Tzu-Chuen Lu, Chaoyang University of Technology, Taiwan.
Complete Peer review History: http://www.sdiarticle3.com/review-history/39029

Original Research Article

Received: 15 November 2017
Accepted: 31 January 2018
Published: 01 April 2019

Abstract

This paper gives the current overview of the application of data mining techniques on the haematological and biochemical dataset to predict the occurrence of malaria in children between age zero (0) and five (5). Malaria has been eradicated from the developed countries but still affecting a large part of the world negatively. A larger percentage of malaria is estimated to affect young children in sub-Sahara Africa. In order to reduce mortality from paediatric malaria, there should be an efficient and effective prediction method. In healthcare, data mining is one of the most vital and motivating areas of research with the objective of finding meaningful information from huge data sets and provides an efficient analytical approach for detecting unknown and valuable information in healthcare data. In this study, a model was built to predict the occurrence of malaria in children between age zero (0) and five (5) years, using decision tree classification algorithms on WEKA workbench tool. The classification algorithms used are LMT, REPTree, Hoeffding tree and J48. A J48 algorithm was used for building the decision tree model since it has higher accuracy for performance with least error margin.

Keywords: Healthcare; paediatric; malaria; mortality; data mining; classification.

**Corresponding author: E-mail: tcolayinka@gmail.com;*

1 Introduction

Continuous generation of the exponentially growing data has given rise the concept of data mining for knowledge discovery and decision based on the data. The amount of Electronic Health Records collected by healthcare facilities is also on an exponential progression on a daily basis. It has been the norm for nurses to take responsibility in handling patient data input that was traditionally recorded in paper-based forms. Presently, the whole world is massively digitalized and computerized, where paper-based of recording of a large amount of data, managing the data of effective health care delivery is tedious and futile. Accuracy is extremely important when it comes to patient care and computerizing this massive amount of data enhances the quality of the whole system.

1.1 Data mining

Data Mining is one of the most vital and motivating areas of research with the objective of finding meaningful information from huge data sets. Presently, data mining is becoming popular in healthcare field because there is a need for efficient analytical approach for detecting unknown and valuable information in health data. Data mining involves the analysis of large observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner [1]. It is also known as knowledge discovery in database, which is the non-trivial extraction of implicit, previously unknown and potentially useful information from the data. In real life application, data mining can be used to predict the disease possibility from the health record with an approved feature of an individual. In the health industry, data mining provides several benefits such as detection of the fraud in health insurance, availability of the medical solution to the patients at lower cost, detection of causes of diseases and identification of medical treatment methods. It also helps the health care researchers to make efficient healthcare policies, constructing drug recommendation systems, developing health profiles of individuals and so on [2]. The analysis of health data improves the healthcare by enhancing the performance of patient management tasks [3]. Moreover, it is used to uncover patterns from a large amount of stored information and then used to build predictive models which are being utilized in healthcare programs around the world.

1.2 Paediatric malaria

Malaria is an insect-borne parasitic infection of tremendous global importance. Malaria is a parasitic infection of erythrocytes or red blood cells that can lead to organ failure. There are four species of Plasmodium that infect human, and they are Plasmodium *falciparum*, Plasmodium *vivax*, Plasmodium *ovale* and Plasmodium *malariae,* but Plasmodium *falciparum* is responsible for most malaria fatalities. Malaria is transmitted from one human to another by infected female Anopheles mosquitoes. Sometimes, it can also be transmitted congenitally (from an infected mother to her unborn baby) and by blood transfusions. The World Health Organization (WHO) estimated that there were approximately 650,000 deaths directly attributed to malaria worldwide [4]. The heaviest burden of Plasmodium *falciparum* malaria falls on sub-Saharan Africa, where children under five years old are disproportionately affected by this parasite. Therefore, malaria remains a very common cause of hospital admission in sub-Saharan Africa, where severe malaria is mainly a disease of children under five years of age. It has been estimated that approximately 90% of the world’s severe and fatal malaria affects young children in sub-Saharan Africa [5]. Paediatric malaria is still endemic to a certain part of the world especially less developed and developing nations like Nigeria. Lack of adequate medical experts to diagnose and make prescription is largely responsible for paediatric mortality. To decrease mortality from paediatric malaria, there should be a fast and effective detection method. There is urgent need to stop paediatric malaria occurrence. Data mining can be an appropriate tool to help the general practitioner in detecting the malaria early by obtaining knowledge and information from patient's data.

2 Literature Review

2.1 Review of related work

Ibrahim et al. [6] in their study compared different classification techniques using WEKA for breast cancer. The aim of the study is to investigate the performance of different classification methods for a set of large datasets. The algorithms tested are Bayes Network, Radial Basis Function, Pruned Tree, Single Conjunctive Rule learner and Nearest Neighbours. The best algorithm on the breast cancer data sets is Bayes network classifier with the highest accuracy and lowest average error.

Boris and Milan [7] performed prediction and decision making in healthcare using data mining. They analysed the usefulness of data mining in the healthcare sector and some of the obstacles that disable the effective and efficient prediction.

Sharma et al. [8] in their study presented malaria outbreak prediction model using machine learning. In this study, they stated that the early prediction of malaria outbreak is the key to the control of malaria morbidity. This prediction can help as an early warning tool to identify potential outbreaks of malaria. The machine learning used for the data mining was classification algorithms based on support vector machine (SVM) and artificial neural network (ANN). Also, the total number of Plasmodium falciparum cases and an outbreak occurs in binary values yes or no. Root mean square error (RMSE) and receivers operating characteristics (ROC) were used to measure the performance of the models.

Kapor and Rani [9] employed an efficient decision tree algorithm using J48 and reduced error pruning. In the study, decision trees were utilized to delineate decision-making process. The decision tree builds classification or regression models in the form of the tree structure, which divides the datasets into tinier and tinier subsets. Some of the benefits and limitations of the decision trees where highlighted. The paper introduces a new decision tree algorithm based on J48 and reduced error pruning. The tree obtained is fast decision tree learning and will be based on the information gain or reducing the variance.

Leopard et al. [10] worked on survey and analysis on classification and regression data mining techniques for disease outbreak prediction in datasets. In this study, the need to develop a strong model for the prediction of disease outbreak in various countries using data mining algorithms was discussed. The advantages and disadvantages of the different classification techniques were highlighted and also the accuracy measures in decision trees from previous publications from the year 2001 to 2014 were presented.

Bbosa et al. [11] studied clinical malaria diagnosis: ruled-based classification statistical prototype. In the study, they were able to identify the predictors of malaria, developed data mining, statistically enhanced rule-based classification to diagnose malaria and developed an automated system to incorporate the rules and the statistical models. The prototype was evaluated for efficacy showing a sensitivity value of 70% across the age groups. They also presented tables for malaria prevalence, signs and symptoms of both hospital and diagnosis.

2.2 Process of data mining

Data mining can better be regarded as a process and not just a set of tools. Fig. 1 depicts the life cycle of a data mining, as defined by the Cross Industry Standard Process for Data Mining (CRISP-DM) reference model [12]. The first phase is the business understanding phase, which involves investigating the business objectives and requirements, deciding whether data mining can be applied to meet them, and determining what kind of data can be collected to form a deployable model. The second phase is the data understanding, where an initial dataset is established and considered to see whether it is suitable for further processing. The third phase is Preparation. Preparation in data mining involves preprocessing the raw dataset, to produce a model using machine learning algorithms. Also, preparation is a structural description of the information that is implicit in the data and model building activities. The fourth phase is Evaluation, where structural descriptions inferred from the datasets have any predictive value. In case, the evaluation phase shows that the model is poor there will need to reconsider the entire project from the business understanding phase to identify other business. On the other hand, when the accuracy of the model is high, then, its process can proceed to the final phase, called Deployment. Deployment involves integrating the model into a larger software system, where it can be implemented.

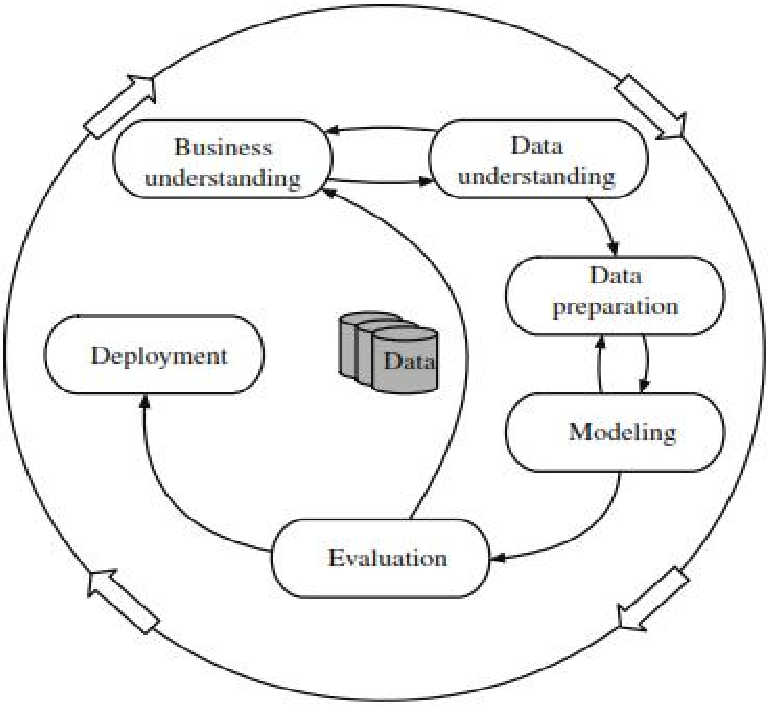

Fig. 1. Life cycle of a data mining project (Shearer, 2000)

2.3 Impact of data mining in healthcare

Data overload: From computerized health records, there is a wealth of knowledge to be gained. Yet the overwhelming bulk of data stored in these databases makes it extremely difficult, if not impossible, for humans to sift through it and discover knowledge.

Evidence-based medicine and prevention of hospital errors: Whenever medical organizations apply data mining to their existing data, they can discover new, useful and potentially life-saving knowledge that otherwise would have remained inert in their databases.

Policy-making in public health: Data mining can be used to analyze similarities between community health centres. Using data mining, it is possible to discover patterns among health centres which steered to policy recommendations for public health. Data mining and decision support methods, including novel visualization methods, can lead to better performance in decision-making.

More value for money and cost savings: Data mining allows medical organizations to get more out of existing data at minimal extra cost. Data mining have been applied to discover fraud in credit cards and insurance claim and to detect anomalous patterns in health insurance claims.

Early detection and prevention of disease: The use of classification algorithm in data mining will help in the early detection of heart disease, a major public health concern all over the world. The use of data mining as a tool will aid in monitoring trends in the clinical trials of vaccines [13]. With the use of data mining, medical experts could find patterns and variances better than just looking at a set of tabulated data.

2.4 Constraints of data mining in healthcare

One of the biggest difficulties in data mining in the healthcare sector is that the raw medical data is voluminous and heterogeneous [14]. These data can be gathered from various sources such as an interview with patients, laboratory results, review and interpretation of doctors. All these mechanisms can have a major impact on diagnosis, prognosis and treatment of the patient, and should not be ignored. The scope and complexity of medical data are one of the barriers to successful data mining. Moreover, missing, incorrect, inconsistent or non-standard data such as pieces of information saved in different formats from different data sources create a major obstacle to successful prediction in data mining [7]. Most time, it is very difficult for a researcher to process gigabytes of records, although working with images is relatively easy, because doctors are able to recognize patterns, to accept the basic trends in the data, and formulate rational decisions. Information stored in the database becomes less useful if they are not available in the easily apprehensible format.

3 Methodology

3.1 Waikato environment for knowledge analysis (WEKA) tool

Waikato Environment for Knowledge Analysis (WEKA) workbench is a collection of machine learning algorithms and data preprocessing tools. WEKA was developed at the University of Waikato in New Zealand [15]. It runs on almost any platform and has been tested under Linux, Windows, and Macintosh operating systems. WEKA is designed such that the user can quickly try out existing methods on new datasets in flexible ways and provides extensive support for the whole process of experimental data mining. The workbench of WEKA includes methods for the main data mining problems such as classification, regression, clustering, association rule mining, and attribute selection.

3.2 Decision tree

A decision tree is a classification algorithm. Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts. The models are derived based on the analysis of a set of training data, that is, data objects for which the class labels are known. The model is used to predict the class label of objects for which the class label is unknown [16].

3.2.1 J48 algorithm

J48 is an extension of Iterative Dichotomiser 3 (ID3). The ID3 is a greedy learning decision tree algorithm that is based on Hunts algorithm. The additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges and derivation of rules [17]. In WEKA data mining tool, J48 is an open source Java implementation of the C4.5 algorithm.

3.2.2 LMT (logistic model trees) algorithm

Logistic Model Trees (LMTs) was born out of the combination of the complementary advantages and disadvantages of linear logistic regression and tree induction. Linear logistic regression fits a linear model to the data, and the process of model fitting is quite stable, resulting in low variance but potentially high bias. [18]. The algorithm can deal with binary and multi-class target variables, numeric and nominal attributes and missing values.

3.2.3 REPTree algorithm

Reduced Error Pruning Tree (REPTree) is fast decision tree learning algorithm, and it uses information gain to build decision tree thereby reducing the variance. RepTree uses the regression tree logic and creates multiple trees in different iterations [15].

3.2.4 Hoeffding tree algorithm

A Hoeffding tree is an incremental, anytime decision tree induction algorithm that is capable of learning from massive data streams, assuming that the distribution generating examples does not change over time [19]. Hoeffding trees exploit the fact that a small sample can often be enough to choose an optimal splitting attribute.

3.3 Pre-processing of dataset

A laboratory test was carried out on children between age zero (0) and five (5) for one hundred and fifty (150) children. Each dataset consists of both haematological and biochemical parameters analysed from the blood samples of children under study. Haematological parameters in each dataset include Haemoglobin (HB) (g/dl), Packed Cell Volume (PCV)(%), Platelet (Plt) count (10^9/l) and White Blood Cell (WBC) (10^9/l). White Blood Cell parameters are Neutrophil (%), Lymphocyte (%), monocyte (%), eosinophil (%), basophil (%). Biochemical parameters in each dataset are Glucose (mmol/l), Protein (g/l), Albumin (g/l) and Globulin (g/l). In the classification algorithm, these parameters form the attributes for consideration in the classification of malaria status (positive or negative). The dataset consists of 100 malaria positive children and 50 negative malaria children. The format of the dataset acceptable for Waikato Environment for Knowledge Analysis (WEKA) workbench is Attribute-Relation File Format (ARFF) [20]. WEKA can convert CSV file format from Microsoft Excel to ARFF dataset format. The WEKA Explorer will use these automatically if it does not recognize a given file as an ARFF file, the preprocess panel has facilities for importing data from a database, and for preprocessing this data using a filtering algorithm. These filters can be used to transform the data and make it possible to delete instances and attributes according to specific criteria.

3.3.1 Training datasets using WEKA tool

The feature selection was carried out on the Biochemical and Haematological parameters measured from the blood samples were analysed for the malaria prediction. This was done by visualizing all the biochemical and haematological parameters using WEKA tool with the training datasets when children are positive or negative. From the visualization, the feature selection was carried out. Only glucose out of all the biochemical parameters that have the attribute for predicting malaria because all children who were malaria positive also exhibit low glucose level and other biochemical parameters (albumin, globulin and total protein), malaria positive cases appear across board hence they have little or no attribute in classifying malaria. Hence age and sex are not attributes to consider for malaria classification. The WBC, Lymphocyte and Neutrophil have little or no attribute in classifying malaria since positive malaria cases occur across their range. Based on the feature selection, these parameters were selected for training dataset: HB, PCV, Platelets, Monocyte, Eosinophil, Basophil and Glucose and to model the decision tree because they contribute much in the decision-making process.

The training dataset is sampled for 100 children which are loaded into WEKA workbench thorough the explorer interface. The decision tree algorithms used for training datasets are J48, LMT, REPTree and Hoeffding tree. The result from different algorithms is then cross-validated using performance classifier measure, and the performance of each algorithm is then compared to each other. To implement the architecture in WEKA, Java Runtime Environment (JRE) must be installed as pre-requisite for running WEKA. Fig. 2 depicts the architecture for decision tree algorithm.

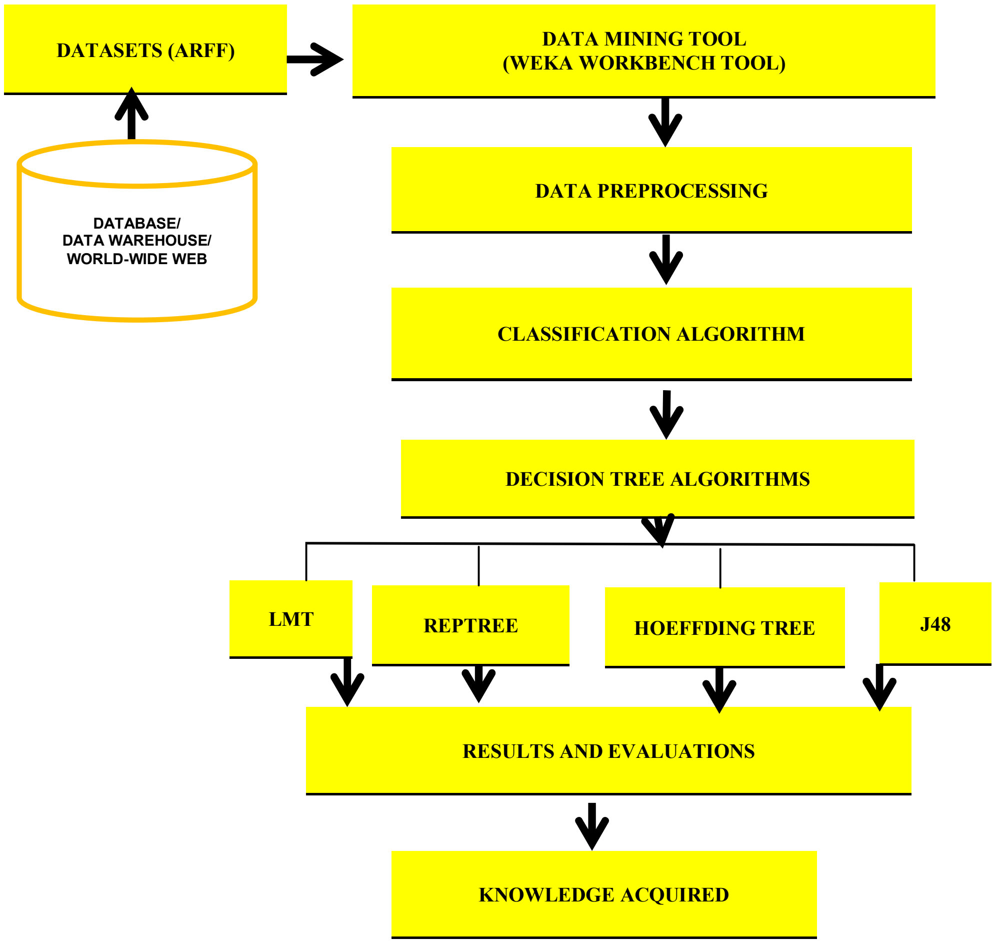

Fig. 2. Architecture for decision tree algorithm

4 Results and Discussion

The dataset obtained from the feature selection of the haematological and biochemical parameters were subjected to four (4) decision tree classifier algorithms: LMT, REPTree, J48 and Hoeffding tree, to build a model for better prediction of paediatric malaria cases. The classifier results are analysed for performance and absolute error. Table 1 depicts the comparison of decision tree classifier algorithms used. From the table, it can be seen that J48 is best classifier because it has least errors (mean absolute error, root mean squared error, relative absolute error and root relative squared error). LMT has 98% correctly classified instances as against 97% for J48, the higher error margins in LMT makes it an inferior choice to J48 classifier. The main disadvantage of LMT is that the time required building the model, which is due mostly to the cost of building the logistic regression models at the nodes. Fig. 3 and Fig. 4 depict the graphical representations of the correctly classified instances (%), mean absolute error and root mean squared error respectively. Fig. 5 displays the decision tree model for predicting paediatric malaria. The decision tree was generated from the J48 classifier for classifying malaria in the test dataset. The J48 pruned tree consists of seven (7) number of leaves and the tree size of thirteen (13). The analysis of the decision tree shows that when Eosinophil is > 3 (%), Platelets > 246 (10^9/l), Monocyte < = 1 (%), HB > 11.2 (g/dl), and Glucose > 3.61 (mmol/l), the child inclines to be malaria negative. Moreover, when the Eosinophil is < = 3 (%), Platelets < = 246 (10^9/l), Monocyte > 1 (%), HB < = 11.2 (g/dl) and Glucose < = 3.61 (mmol/l), the child inclines to be malaria positive.

Table 1. Comparison of decision tree classifier algorithms

	Hoeffding Tree	LMT	REPTree	J48
Correctly Classified Instances (%)	94	98	94	97
Mean absolute error	0.0769	0.077	0.1074	0.0539
Root mean squared error	0.229	0.1658	0.2317	0.1641
Relative absolute error	17.3465	17.3728	24.2275	12.1554
Root relative squared error	48.6804	35.2655	49.2832	34.9084

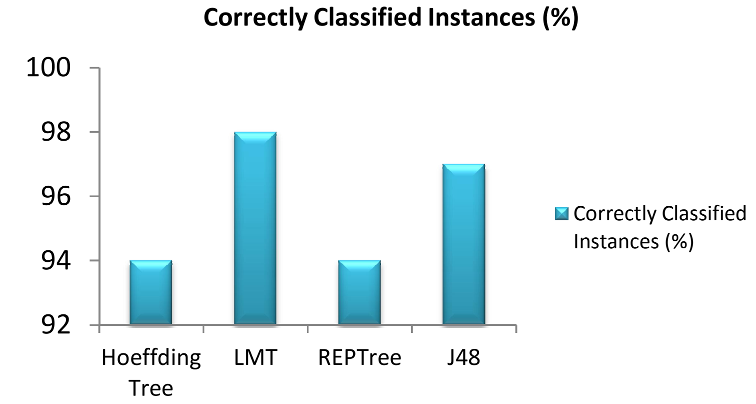

Fig. 3. Percentage correctly classified instances for different classifiers

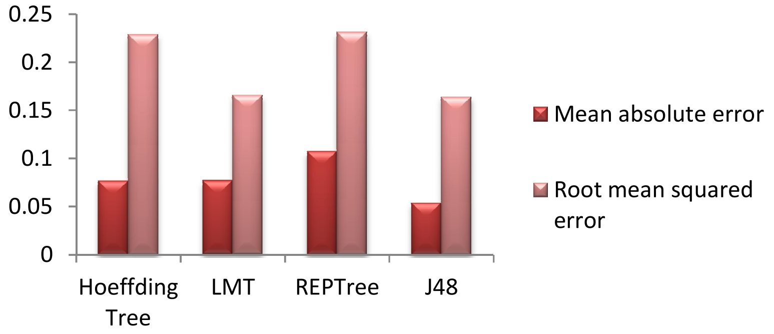

Fig. 4. Root mean squared error and mean absolute error for different classifiers

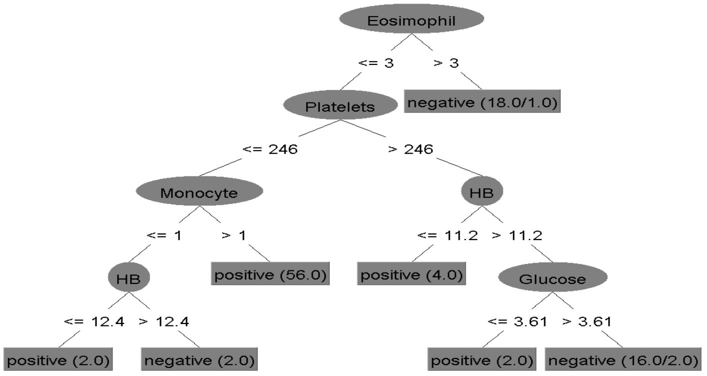

Fig. 5. Decision tree model for predicting paediatric malaria

5 Conclusion

In this study, paediatric malaria was predicted on haematological and biochemical datasets using classification algorithms in data mining. The classification algorithms used for training datasets are, LMT, RepTree, Hoeffding tree and J48, for the occurrence of malaria in children between age zero (0) to five (5) years. J48 algorithm was used for building the model because it has higher accuracy for performance with least error margin, which predicted all cases. The model built to predict paediatric malaria cases is an efficient and effective tool for early detection of malaria occurrence in children to reduce the mortality rate.

Competing Interests

Authors have declared that no competing interests exist.

References

[1] David H, Mannila H, Smyth PC. Principles of data mining. MIT Press, Cambridge; 2001.

[2] Koh HC, Tan G. Data mining application in healthcare. Journal of Healthcare Information Management. 2005;19(2).

[3] Tomar D, Agarwal S. A survey on data mining approaches for healthcare. International Journal of Bio-Science and Bio-Technology. 2013;5(5):241-266.

[4] World Health Organization. World Malaria Report: Geneva; 2012.

[5] Roca-Feltrer A, Carneiro I, Armstrong Schellenberg JR. Estimates of the burden of malaria morbidity in Africa in children under the age of 5 years. Trop Med Int Health. 2008;13:771–83.

[6] Ibrahim F, Abu N, Osman A, Usman J, Kadri NA. Comparison of different classification techniques using weka for breast cancer. Biomed 06, IFMBE Proceedings. Springer-Verlag Berlin Heidelberg Publisher. 2007;15:520-523.

[7] Boris M, Milan M. Prediction and decision making in health care using data mining. Kuwait Chapter of Arabian Journal of Business and Management Review. 2012;1(12).

[8] Sharma V, Ajai K, Lakshmi P, Ganesh K, Anuradha L. Malaria outbreak prediction model using machine learning. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET). 2015;4(12).

[9] Kapor P, Rani R. Efficient decision tree algorithm using j48 and reduced error pruning. International Journal of Engineering Research and General Science. 2015;3(3): 2091-2730.

[10] Leopard H, Cheruiyot KW, Kimani S. A survey and analysis of classification and regression data mining techniques for diseases outbreak prediction in datasets. The International Journal of Engineering and Science (IJES). 2016;5(9):01-11.

[11] Bbosa F, Ronald W, Peter J. Clinical malaria diagnosis: Ruled-based classification statistical prototype. Publisher: SpringerPlus. 2016;5:939.

[12] Witten HI, Frank E, Hall MA, Pal CJ. Data mining: Practical machine learning tools and techniques. Fourth Edition Morgan Kaufmann (Elsevier); 2017.

[13] Cao X, Maloney KB, Brusic V. Data mining of cancer vaccine trials: A bird's-eye view. Immunome Research. 2008;4:7.
DOI: 10.1186/1745-7580-4-7

[14] Cios KJ, Moore GW. The uniqueness of medical data mining. To Appear in Artificial Intelligence in Medicine Journal; 2002.

[15] Witten IH, Eibe F, Christopher JP, Mark AH. The weka workbench “data mining: Practical machine learning tools and techniques”. Morgan Kaufmann, Fourth Edition. 2016;7.

[16] Han J, Kamber M, Pei J. Data mining concepts and techniques third edition by Elsevier Inc. 2012;18-19:622-624.

[17] Quinlan R. C4.5: Programs for machine learning. Morgan Kaufmann Publishers, San Mateo, CA; 1993.

[18] Sumner M, Eibe F, Mark H. Speeding up logistic model tree induction. In: 9th European Conference on Principles and Practice of Knowledge Discovery in Databases. 2005;675-683.

[19] Hulten G, Laurie S, Pedro D. Mining time-changing data streams. In: ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining. 2001;97-106.

[20] Hall M, Eibe F, Geoffrey H, Bernhard P, Peter R, Witten H. The weka data mining software: An update. SIGKDD Explorations. 2009;11(1).

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://www.sdiarticle3.com/review-history/39029