ISSN 2278-3091
Volume 8, No.3, May - June 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse94832019.pdf
https://doi.org/10.30534/ijatcse/2019/94832019

A Clinical Diagnostic Model Based on Supervised Learning

Oguntimilehin A.[1], Babalola Gbemisola. O[2], Olatunji K.A[3]
[1]Afe Babalola University, Ado-Ekiti, Nigeria, ebenabiodun2@yahoo.com
[2]Afe Babalola University, Ado-Ekiti, Nigeria, gbemibabz@abuad.edu,ng
[3]Afe Babalola University, Ado-Ekiti, Nigeria, olatunjika@abuad.edu,ng

ABSTRACT

Shortages of medical practitioners, medical facilities and complexity of diseases among others have given rise to the need for the use of computer programs to give helping hands in the health sector. Supervised Learning as a method of developing predictive systems has been proved to be powerful on classification problems, mostly when dealing with diseases. It was used on malaria fever datasets collected from a reputable hospital in Ado-Ekiti, Ekiti State, Nigeria in this work to create a classification model using Partial Tree (PART) Technique. The developed model was tested on both the training and the testing sets with the detection rates of 100% and 98.04% respectively, and adjudged promising. The final implementation and deployment of the model will be carryout as a mobile application so as to have a wider coverage in terms of accessibility and usability. It is hopeful it will be of immense benefits in the health sector.

Key words: Diagnosis, Therapy, Supervised Learning, Machine Learning, Malaria Fever.

1. INTRODUCTION

A machine learns whenever it changes its structure, program, or data (based on its inputs or in response to external information) in such a manner that its expected future performance improves. Some of these changes, such as the addition of a record to a data base, fall comfortably within the province of other disciplines and are not necessarily better understood for being called learning. But for example when the performance of a speech-recognition machine improves after hearing several samples of a person's speech, there will be justification that the machine has learned. Machine learning usually refers to the changes that perform tasks associated with artificial intelligence (AI). Such tasks involve recognition, diagnosis, planning, robot control, prediction and so on [8].

There are three major settings in which a function can be learned, supervised learning, unsupervised learning and semi-supervised learning. Supervised machine learning is the search for algorithms that reason from externally supplied instances to produce general hypotheses, which then make predictions about future instances. In other words, the goal of supervised learning is to build a concise model of the distribution of class labels in terms of predictor features. The resulting classifier is then used to assign class labels to the testing instances where the values of the predictor features are known, but the value of the class label is unknown [7].

Supervised learning is the machine learning task of inferring a function from supervised training data. The training data consist of set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervised signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which is called a classifier (if the output is discrete) or a regression function (if the output is continuous). The inferred function should predict the correct output value for any valid input object. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way. The parallel task in human and animal psychology is often referred to as concept learning [6].

A wide range of supervised learning algorithms is available, each with its strengths and weaknesses. There is no single learning algorithm that works best on all supervised learning problems. Supervised learning is powerful when the classifications are known to be correct (for instance, when dealing with diseases, it is generally straight-forward to determine the design after the fact by an autopsy), or when the classifications are simply arbitrary things that we would like the computer to be able to recognize for us. Classification learning is often necessary when the decisions made by the algorithm will be required as input somewhere else. Otherwise, it would not be easy for whoever requires that input to figure out what it means. Both techniques can be valuable and which one you choose should depend on the circumstances, what kind of problem is being solved, how much time is allotted to solving it and whether supervised learning is even possible [2].

The increased in demand for high-quality medical services coupled with the explosive growth of medical knowledge led to the suggestion that computer programs should be used in assisting physicians and other healthcare providers in discharging their clinical roles such as diagnosis, therapy and treatment. Computer tools help to organize, store and retrieve appropriate medical knowledge needed by the practitioner in

dealing with each difficult case and suggesting appropriate diagnosis, prognosis, therapeutic decisions and decision-making technique [12].

Diagnosis is the identification of abnormal condition that afflicts a specific patient, based on manifested clinical data or lesions. If the final diagnosis agrees with a disease that afflicts a patient, the diagnostic process is correct; otherwise, a misdiagnosis occurred [1].

Malaria is a mosquito borne infectious diseases caused by a eukaryotic protist of the genus plasmodium. It is wide spread in tropical and subtropical regions, including parts of the American, Asia and African. Five species of the plasmodium parasite can infect humans, the most serious form of the disease are caused by plasmodium falciparum. Malaria caused by plasmodium vivax, plasmodium ovale and plasmodium malariae causes milder disease in humans that is not generally fatal. A fifth species, plasmodium Knowlesi, is a zoonosis that causes malaria in macaques but can also affect humans [10].

Since Charles Levran first visualized the malaria parasite in blood in 1880, the mainstay of malaria diagnosis has been the microscopic examination of blood. Areas that cannot afford even simple laboratory diagnostics tests often use only a history of subjective fever as the indication to treat for malaria [1].

2. LITERATURE REVIEW

Some of the existing works on the subject matter were carefully reviewed so as to build on their strengths and weaknesses. Some of them are presented below:

A Decision Support System for Diagnosing Tropical Diseases Using Fuzzy Logic was presented in [13]. The researchers embarked on this work because tropical diseases are associated with a high level of mortality rate. Data were gathered by interacting with various medical doctors who are experts in diagnosing tropical diseases to gain heuristic knowledge on the diseases. The system was developed to diagnose ten tropical diseases including malaria. Diagnosis was carried out by weighing each symptom with respect to the disease in question using generalized fuzzy soft set (GFSS). This system lacks performance evaluation, carries out diagnosis without therapy, not mobile based and use of small dataset. A Knowledge-Based Data Mining System for Diagnosing Malaria Related Cases in Healthcare Management was developed in [14]. The need of computer based approaches in health sector led to this work. Data collection was obtained by survey from four hospitals in Lagos metropolis of Nigeria. Visualization and knowledge representation techniques were used to present the mined knowledge to the user. The components of the knowledge based data mining system are: knowledge base, inference engine, rules and decisions. The implementation of the system was carried out using C#.NET programming language and Microsoft SQL Server 2005. Lack of performance evaluation of the developed system is one of the major setbacks.

In [18], Computer Automation for Malaria Parasite Detection using Linear Programming was developed. Malaria causes more than 1 million deaths arising from approximately 300-500 million infections every year, manual microscopy is not a reliable screening method when performed by non-experts and need of an automated system aims at performing diagnosis without human intervention are some of the motivations for this work. Formulation of a linear programming model based on the given data, solving and displaying the result using graphical method approach for detecting parasite was the method adopted. Challenges of the work include no measurement of accuracy.

Decision Support System for Malaria and Dengue Disease Diagnosis (DSSMD) was presented by Priynka *et al* in [15]. This work was motivated as a result of reasons which include -malaria and dengue remain to be the most vital cause of morbidity and mortality in India and in many other tropical countries with complete 2 to 3 million new cases arising every year, malaria is a major health problem in the world, oldest chronic and most widespread fatal disease, unavailability of pathological and imaging based medical diagnosis tool in remote areas. The system was developed using MATLAB. The overall classification was done using fuzzy logic toolbox. The system has three modules; GUI interface showing the symptoms, Knowledge Base where fuzzification takes place, and Inference Engine where the fuzzified value is defuzzified in the decision support system model. More than 200 fuzzy rules were generated by the system for diagnosis. No clear indication of good accessibility.

A Knowledge Based Expert System for Symptomatic Automated Healthcare was developed in [16]. Healthcare distribution is being transformed by developments in e-health, the revolution of medical science to computer technology has made life easy and patients or end users are not bound to the doctors or other resources of medical science are the motivations for this work. The system is an Expert System having three client modules – user interface, inference engine and knowledge base. Patients or users remotely interact with the system and find out the disease by giving some symptoms to the computer. In this way, the system makes feasible diagnosis of patients and also suggests particular treatment regarding the disease. Forty-eight diseases, including malaria, were diagnosed by the system. The bases on which the rules were generated was not mentioned, considering 48 diseases may jeopardize the effectiveness of the system and detection rate of the system was not measured.

It is a scientific belief that the bigger the data used in a research, the better the coverage and accuracy [11]. Measuring detection rates also build confidence in the system

and the intending users. These among other reasons called for further researches.

3. MATERIALS AND METHOD

Data Collection and Description of Data Sets

Data were collected from a reputable health services provider in Nigeria - Adetoyin Specialist Hospital, Ado-Ekiti, Nigeria. A total of one thousand two hundred and twenty five malaria fever instances diagnosed through symptomatic method were collected for a period of six months and used as the training set for the model, while another set of four hundred and eight (408) instances were collected for another period of six months and was used as a testing set. There are nineteen (19) malaria fever symptoms (conditional attributes) under consideration from the data set. They are: Weakness (WKN), Abdominal Pain (ABP), Cough (COH), Body Pain (BOP), Fever (FVR), Rigour (RGR), Cold (COD), Anorexia (ANR), Headache (HEC), Catarrh (CAH), Insomnia (ISN), Yellow Urine (YEU), Vomiting (VOM), Joint Pain (JOP), Dizziness (DSN), Ill-looking (ILL), Convulsion (COV), Temperature (TMP) and Diarrhoea (DIA).

Using the symptoms of each patient a record was formed using a relational table in Microsoft Excel. A symptom is either perceived to be High or Low or None based on patient's feelings. Using the available symptoms the medical practitioners classified a case of malaria as one of the five tuples (Very High, High, Moderate, Low, Very Low) which means there are five classes in the datasets (both training and testing) which amounts to five decision attributes. For case of programming, the conditional attributes High, Low and None were converted to 2, 1, and 0 respectively while the decision attributes Very High, High, Moderate, Low and Very Low were converted to 5, 4, 3, 2 and 1 respectively. For example a record set is of the form below:
{Headache = High, Insomnia = High, fever = High, Anorexia = High, Yellow urine = Low, then malaria diagnosed = Very High} which is represented in programming as:
{Headache = 2, Insomnia = 2, Fever = 2, Anorexia = 2, Yellow Urine = 1; malaria diagnosed = 5}. A model was built on the training set using Partial Tree Algorithm (PART)

Description of Partial Tree (PART) Technique

Partial Tree (PART) is a Supervised Learning technique which uses Gain Ratio as node splitting criteria to build a partial decision tree and makes the best leaf into a rule. PART does not generate a decision tree but decision list (rules) [9].

The information gain measure is biased toward tests with many outcomes. That is, it prefers to select attributes having a large number of values. An extension to information gain known as gain ratio [17], which attempts to overcome this bias. It applies a kind of normalization to information gain using a split information value defined analogously with *Info*(*D*) as

$$\text{Split info}_A(D) = -\sum_{j=1}^{v} \frac{|Dj|}{|D|} * \log_2\left(\frac{|Dj|}{|D|}\right) \quad (1)$$

This value represents the potential information generated by splitting the training data set, *D*, into *v* partitions, corresponding to the *v* outcomes of a test on attribute *A*. Note that, for each outcome, it considers the number of tuples having that outcome with respect to the total number of tuples in *D*. It differs from information gain, which measures the information with respect to classification that is acquired based on the same partitioning. The gain ratio is defined as:

$$\text{Gain Ratio} = \frac{Gain(A)}{Split\,\text{info}(A)} \quad (2)$$

The attribute with the maximum gain ratio is selected as the splitting attribute. Note, however, that as the split information approaches 0, the ratio becomes unstable. A constraint is added to avoid this, whereby the information gain of the test selected must be large-at least as great as the average gain over all tests examined [5,6].

PART however used Reduced Error Pruning (REP) method for pruning the decision list. For more information on REP, consult [4, 3].

Development of a Diagnostic Model from Partial Tree (PART)

The partial Tree (PART) used the one thousand two hundred and twenty five (1225) training instances to generate a classification model for the diagnosis. The model generated serves as the engine room or inference engine for the intending mobile application so as to have a wider coverage in terms of usability. The model is presented below:

INS = 2 AND WKN = 1 AND DSN = 1: 4 (453.0)
ANR = 2 AND INS = 1 AND COV = 1: 4 (121.0)
HEC = 2 AND INS = 2 AND COH = 2: 5 (24.0)
HEC = 2 AND DSN = 1 AND ILL = 1 AND INS = 1 AND FVR = 2 AND BOP = 1: 3 (124.0)
HEC = 2 AND DSN = 1 AND ILL = 1 AND INS = 1 AND COH = 2 AND YEU = 2: 3 (48.0)
HEC = 1 AND ANR = 1 AND COV = 1 AND INS = 1 AND COH = 2 AND YEU = 1: 2 (86.0)
HEC = 1 AND ANR = 1 AND FVR = 1 AND COV = 1 AND COH = 1: 1 (61.0)
HEC = 1 AND FVR = 1: 5 (49.0)
DSN = 2: 5 (25.0)
ANR = 2 AND HEC = 1: 5 (24.0)
ILL = 1 AND HEC = 2 AND INS = 1 AND COH = 1 AND BOP = 2: 3 (26.0)
ILL = 1 AND HEC = 2 AND COH = 1 AND WKN = 2: 4 (24.0)
ILL = 1 AND COH = 1 AND FVR = 2 AND WKN = 1 AND BOP = 1: 1 (13.0)

ILL = 1 AND COH = 1: 2 (37.0)
VOM = 1 AND ILL = 1 AND ABP = 1 AND CAH = 1: 3 (36.0)
HEC = 2 AND YEU = 1 AND BOP = 1: 4 (24.0)
ABP = 2 AND WKN = 1: 4 (13.0)
BOP = 2: 3 (13.0)
WKN = 2: 5 (12.0)
Number of Rules : 20

4. RESULT AND DISCUSSION

The diagnostic model was tested on the one thousand, two hundred and twenty five training instances as well as the four hundred and eight (408) testing sets the confusion matrices of the results are presented in Table 1 and Table 2.

Table 1: Confusion Matrix for the Training Set

Predicted as Actual	Very Low	Low	Moderate	High	Very High
Very Low(12)	134 (100%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Low (22)	0 (0.00%)	625 (100%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Moderate (40)	0 (0.00%)	0 (0.00%)	247 (100%)	0 (0.00%)	0 (0.00%)
High (104)	0 (0.00%)	0 (0.00%)	0 (0.00%)	135 (100%)	0 (0.00%)
Very High (20)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	74 (100%)

TP = Class group correctly classified
TN = Class group wrongly classified

$$\text{Detection Rate} = \frac{TP}{TP + TN} = \frac{134+635+247+135+74}{1225 + 0} = \frac{1225}{1225} = 100\%$$

Table 2: Confusion Matrix for the Testing Set

Predicted as Actual	Very Low	Low	Moderate	High	Very High
Very Low(12)	33 (100%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Low (22)	8 (3.01%)	258 (96.99%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Moderate (40)	0 (0.00%)	0 (0.00%)	42 (100%)	0 (0.00%)	0 (0.00%)
High (104)	0 (0.00%)	0 (0.00%)	0 (0.00%)	49 (100%)	0 (0.00%)
Very High (20)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	18 (100%)

TP = Class group correctly classified
TN = Class group wrongly classified

$$\text{Detection Rate} = \frac{TP}{TP + TN} = \frac{33+258+42+49+18}{408 + 0} = \frac{400}{408} = 98.04\%$$

5. CONCLUSION AND RECOMMENDATION

This research work has again demonstrated the possibility of synergy between computer science or information Technology professionals and the medical practitioners in bringing hospitals to homes, thereby reducing or eliminating most of the shortcomings in the health sector which include shortage of medical practitioners and unavailability of medical serves mostly in rural areas. Full implementation of the malaria fever diagnostic model generated from this work as mobile application will be of immense benefits to the health sector and individuals in the malaria belt of the world.

ACKNOWLEDGMENT

The Authors appreciate the contributions of the Chief Medical Director - Dr. Adeife Erinfolami and other medical practitioners of Adetoyin Specialist Hospital, Ado-Ekiti, Ekiti State , Nigeria for their contribution to the success of this research. The Authors whose works were sited are also appreciated

REFERENCES

[1] Adetunmbi A.O, Oguntimilehin A., and Falaki S.O (2012), Web-Based Medical Assistant System for Malaria Diagnosis and Therapy, GESJ: Computer Science and Telecommunications No 1(33), Pg. 42-53.
[2] AIHorizon (2010), Machine Learning, Part1: Supervised and Unsupervised Learning, AIHorizon, www.aihoriszon .com/essays/generai/index.htm. Retrieved 28/07/2017.
[3] Esposito F., Donato M., Glovanni S. and Valentina T (1999), The Effects of Pruning Methods on the Predictive Accuracy of Induced Decision Trees, Applied Stochastic Models in Business and Industry, 15, Pg. 277-299. https://doi.org/10.1002/(SICI)1526-4025(199910/12)15:4<277::AID-ASMB393>3.0.CO;2-B
[4] Haizan W.M., Mohd-Najib M.S. and Abdul H.O (2012), A Comparative Study of Reduced Error Pruning Method in Decision Tree Algorithms, IEEE Conference on Control, Computing and Engineering (978-4673-3141-8), Malaysia.
[5] Ian H.W, Eibe F. and Mark A.H (2011), Data Mining: Practical Machine Learning Tools and Techniques (3rd Edition), Morgan Kaufman Publishers (ELSEVIER), USA.
[6] Jiawei H., Micheline K., and Jian P (2012), Data Mining: Concepts and Techniques (3rd Edition), Morgan Kaufman Publishers (ELSEVIER), USA.
[7] Kotsiants S.B (2007), Supervised Machine Learning: A review of Classification Techniques, International Journal of Computing and Informatics, 31, Pg. 249-268

[8] Nils J.N (1998), Introduction to Machine Learning, Amazon Press. http:// www.scibd.com/doc/ 40480065/Nils-J. Retrieved 26/02/2018.
[9] Nikita P. and Saurabh U (2012), Study of various Decision Tree Pruning Methods with their Empirical Comparison, International Journal of Computer Applications, Vol 60(12), Pg. 20-25.
https://doi.org/10.5120/9744-4304
[10] NTG(2012), Malaria, Centre for Disease Control, Northern Territory Government(NTG), www.nt.gov.au/hea lth, Australia. Retrieved 04/05/18.
[11] Oguntimilehin A. and Ademola E.O. (2014), A Review of Big Data Management, Benefits and Challenges, Journal of Emerging Trends in Computing and Information Sciences, Vol. 5, No. 6 June 2014, Pg. 433-438.
[12] Oguntimilehin A., Adetunmbi A.O. and Abiola O.B (2013), A Machine Learning Approach to Clinical Diagnosis of Typhoid Fever, International Journal of Computer and Information Technology, Vol 2(4), Pg. 671-676.
[13] Olabiyisi S.O, Omidiora E.O, Olaniyan M.O and Derikoma (2011), A Decision Support System Model for Diagnosing Tropical Diseases Using Fuzzy Logic, African Journal of Computing & ICT,Vol4(2)2, Pg. 1-6.
[14] Olugbenga O, Uzoamaka O. and Nwinyi O (2010), A knowledge- Based Data Mining System for Diagnosing Malaria Related Cases in Healthcare Management, Egyptian Computer Science Journal Vol.34(4), Pg.1-10.
[15] Priynka S., Singh D., Manoj K.B and Nidhi M (2013), Decision Support System for Malaria and Dengue Disease Diagnosis (ASSMD), International Journal of Information and Communication Technology, Vol 3(7), Pg. 633-640.
[16] Soomro A.A., Memon N.A, and Menopn M.S (2011), Knowledge Based Expert System for Symptomatic Automated HealthCare, Sindh University Research Journal (Science Series), Vol 03 (I-A), Pg. 79-84.
[17] Vadivu Senthil P. and Bharathi D (2014), Survey on Students' Academic Failure and Dropout using Data Mining Techniques, International Journal of Advances in Computer Science and Technology, vol3, No5, May 2014, Pg 318-324.
[18] Vipul P, Pooja P. and Archana S (2013), Computer Automation for Malaria Parasite Detection using Linear Programming, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering , Vol.2(5), Pg.1984-1988.