CLASSIFICATION AND FEATURE SELECTION OF SYMPTOMATIC AND CLIMATIC BASED MALARIA PARASITE COUNTS

R.G. Jimoh[1], O.A. Abisoye[2], M.M.B. Uthman[3]

[1]Department of Computer Science, Faculty of Communication and Information Science, University of Ilorin, Nigeria.
jimoh_rasheed@unilorin.edu.ng, jimoh_rasheed@yahoo.com
[2]Department of Computer Science, School of Information and Communication Technology, Federal University of Technology, Minna, Niger State, Nigeria.
o.abisoye@futminna.edu.ng, opeglo@yahoo.com.au
[3]Department of Epidemiology and Community Health, Faculty of Clinical Sciences, College of Health Sciences, University of Ilorin, Nigeria.
uthman.mb@unilorin.edu.ng, uthmanmb@yahoo.com

ABSTRACT

Dynamics of Malaria parasite transmission is complex and been widely studied. Research is needed to find a subset of the original features, that will generate a classifier with the highest possible accuracy. Feature selection improves classifier performance; because some machine learning algorithms are known to degrade in performance when faced with many irrelevant/noisy features. In this paper, Support Vector machine (SVM) with One_against_all algorithm is employed to select optimal features for the multiclass symptomatic and climatic malaria parasite-count. Monthly surveys of malarial incidences cases were collected from sampled health centers in Minna Metropolis, Niger State, Nigeria and served as input variables. Linear, radial basis and polynomial kernel functions were employed but SVM with radial basis kernel function produced better performance result of 85.60% Accuracy, 84.06% Sensitivity and 86.09% Specificity at optimum threshold value of 0.60. SVM selected optimal features to improve prediction performance and reduces time complexity. The experimental results show the robustness and reliability of the proposed model compared to the previous related models.

Key Words: Malaria, support vector machine, feature selection, symptomatic, climatic and multiclass

1.0 INTRODUCTION

Malaria transmission is site specific due to variations of climatic conditions of a region. Temperature, rainfall, relative humidity variations affects the life cycle of malaria parasite [1]. Other non-climatic factors, such as human/behavioural factors can also affect the spread of malaria transmission and severity [2].

Recent researches focuses on dynamics and complexities of Malaria parasite transmission. Research is ongoing on how the risk of asymptomatic and symptomatic malaria infection changes[3,4]. Malaria parasite count diagnosis can be asymptomatically or symptommatically low, mild and high. Sometimes, many symptoms of different patient may even overlap. Malaria patient cases may even have characteristics of other diseases. Therefore, medical problems cannot be generalized and analyzed by imagination. Acknowledge intensive program should be conducted to integrate this complex network of problems and devise individualized solutions[5].

Consequently, a huge amount of malaria cases which is hard to understand and to interpret 7 by humans are collected every year [6]. So difficulties arises on how to analyse the data and interpret it to reduce or possible eradicate subsequent occurrences. Then, the need for a Machine Learning (ML) method arises. ML processes the data and automatically learns from the data. The knowledge generated from the extracted infection cases can be used to solve the problem at hand.

Problems being solved by machine learning methods involves classifying observations, predicting values, structuring data (e.g. clustering), compressing data, visualizing data, filtering data, selecting relevant components from data when faced with many irrelevant/ noisy features., extracting dependencies between data components, modeling the data generating systems, constructing noise models for the observed data, integrating data from different sensors, using classification and drawing inferences.[7,8]

But most ML methods have weaknesses of over-fitting due to large number of parameters to fix resulting to computational complexity and prone to local minimum error. Therefore, a feature selection algorithm SVM was proposed. SVM strength lies in ability to handle prediction, pattern recognition and classification. It can also handle small and large dataset well using a predefined activation function. SVM solves the problems of over-fitting by optimizing the model parameters to feature selection. But SVM weaknesses lie in handling only binary prediction, pattern recognition, classification, and regression analysis. It also needs a good kernel function to perform effectively and chooses appropriately hyper parameters that will allow for sufficient generalization performance. SVM incorporates one-against-all algorithm to handle multiclass nature of malaria cases.

Feature Selection aim is to select features that leads to a large between class distance and small within class variance in the feature vector space[9]. It finds a subset of the original features, that will generates a classifier with the highest possible accuracy. There are quantitative (continuous), ordinal and categorical (nominal/ discrete) types of features. Some classifiers like Naïve Bayes, decision trees, treat categorical and quantitative features differently.

Feature selection gives a better understanding of the data and the classification rule [10, 11]. It avoids computational complexity by reducing the number of features to a sufficient minimum. It also improves classifier performance; because some machine learning algorithms are known to degrade in performance. The theoretical justification to retain the highest weighted features for feature selection was ascertained[12].

This paper proposed a Machine Learning (ML) method, Support Vector Machine linear, radial basis, and polynomial kernel function (SVM-rbf) to make control trade-offs between large datasets, sparsity of data representation and select relevant features from data. This will help to reduce space use when working with a limited amount of system memory.

This paper subsequent section is organized as follows: Section two (2) presents related review of feature selection and classification. Section three (3) describe the method and materials used for the model, Section four (4) described how feature selection algorithm was used, Section five (5) explain the result and gives discussion of the result. Finally, Section six (6) sums up the paper with concluding remarks.

2.0 LITERATURE REVIEW

In Sindhwani *et al.,* study, theoretical justification for retaining the highest weighted features has been independently derived in a somewhat different context [12].

Their experiments on text categorization compare the effectiveness of the SVM-based feature selection with that of more traditional feature selection methods. Experimental results indicate that, at the same level of vector sparsity, feature selection based on SVM normals yields better classification performance than odds ratio- or information gainbased feature selection when linear SVM classifiers are used[12]. SVM was also used as a classifier that outperforms most of other classification methods on text data[13, 14]. The limitation of the research was the evaluation of their approach on other data sets, perhaps on domains outside text categorization.

Olivier and Sathiya in 2008 evaluated new embedded methods on a number of text classification problems. The study that embedded methods are superior to a baseline filter method that uses information gain[15]. In parallel works of *Obozinsky et al.*[16] and *Argyriou et al.*[17] a similar model for L1 regularization was developed. They models were applied on multi-task learning and use a block coordinate-wise optimization technique for training.

A research on Support Vector Machine-Firefly Algorithm for malaria diagnosis was conducted in India to classify malaria cases. The motivation was that the performance of SVM mainly depends on its appropriate parameters selection which is very complex in nature and quite hard to solve by conventional optimization techniques. The results indicate that the proposed SVM-FFA model provides more accurate prediction compared to the other traditional techniques. The limitation to the study was that the lead times (such as bi-monthly, quarterly or yearly prediction) were not considered [18].

3.0 MATERIALS AND METHODS

Monthly surveys of malarial incidences were collected from sampled health centers in Minna Metropolis, Niger State. Climatic data consisting of Monthly averages of rainfall, temperature and relative humidity were collected from Nigerian Environmental and Climate Observation Programme (NECOP) Weather

Station, Bosso Campus, Federal University of Technology, Minna, Niger state. Each patient has a set of symptoms and MP count known as Patients' malaria data symptoms and lab test results. This Climatic data combined with monthly malaria incidences were considered as input variables was trained and simulated using Microsoft Excel and libSVM in MATLAB 2015a

Sampled hospitals laboratories, Giemsa staining was used for the laboratory tests. The Red blood cells (RBCs), Plasmodium spp, platelets and other artifacts were identified. This Plasmodium spp is measured in count being called Malaria Parasite Count (MPcount).

a. One Against all Algorithm

SVM is a binary classifier but the algorithm can be used to solve multiclass problem by introducing One-Against-All Algorithm that captures single handedly each class of the target and compare it with the other classes.

Table 1: One-Against-All

Input: Training Malaria Datasets
Output: Optimal Features

1. *Begin*
2. ***For** counter= 1 to Size(target,1)*
3. ***if** Target(counter) = 0*
   Target(counter) == 0
   Else
   Target(counter) == 1
   End
4. *End*

b. Feature Selection Algorithm

The SVM feature selection algorithm was thresholded as shown in Table 2 to get the optimum threshold value that will yield best result for the model using the One-Against-All algorithm.

Table 2: SVM feature Selection Algorithm

Input: Training Malaria Datasets
Output: Optimal Features

1. *Begin*
2. *Input the Malaria Data Features*
3. *Preprocess the data by using the most suited normalization method*
4. *Divide the data into Training Malaria Datasets and Testing Malaria Datasets in ratio 70:30*
5. *Perform One_Against_All(OAA) algorithm to convert Multiclass to Binary class in preparation for feeding into SVM*
6. ***WhileThreshold_Value**>=0.100 Step 0.05 Do*
7. ***WhileAccuracy_Instances**<= No_of _Runs*
8. *Train an SVM*
9. *Simulate SVM*
10. *Recall Simulated SVM*
11. *Simulate with Transposed Testing Malarial Datasets*
12. *Get Simulation Results*
13. *Compute Optimal Features,Accuracy ,Performance Evaluation*
14. ***EndWhile***
15. ***EndWhile***
16. ***End***

3.1 Feature Selection

Given a number of features, wrapper method and Support Vector Machine were used to select subset of features that have the greatest predictive power and still carry their class discriminatory properties. The dataset has these prevalent features: Headache (Hd), Fever (F), Dizziness (D), Body Pain (Bp) and Vomiting (V_m). The climatic factor; temperature, relative humidity and rainfall contributing factors to being having malaria are also the combined features This research features is thus restricted to five (5) predominant malarial symptoms and climatic factors

4.0 RESULTS

The Multiclass malaria data was handled by one-against-all algorithm. The result of various classes of SVM Feature Selection with 1200 malaria cases; 840:180:180 were used for Training, Testing and Validation respectively is presented in *Table 3(a), 3(b) and 3(c).* Also the Graph of the Support vectors Vs. Accuracy for SVM_0, SVM_1, SVM_2 are depicted in Figures 1(a), *1(b) and 1(c),* respectively.

Class 0, Class1 and Class 2 malaria cases were trained, tested and validated with linear, radial basis and polynomial function single handedly. Their results were depicted in *Table 3(a), 3(b) and 3(c).*

Table 3(a): SVM_0 Feature Selection Results

SVM_0	SVM_1	SVM_2	Performance
83.89	80.55	66.67	Accuracy (%)
282	435	378X	Support Vectors
63	60	63	True
88	85	57	True
15	17	45	False
14	18	15	False

SVM_0	SVM_1	SVM_2	Performance
0.8077	0.7692	0.8077	Sensitivity (TP_R)
0.8627	0.8333	0.5588	Specificity (TN_R)
0.1311	0.1667	0.4412	(FR_R)
0.1311	0.2308	0.1923	(FN_R)
0.6444	0.556	1.333	(MSE)
78	78	78	Total Positive
102	102	102	Total Negative
180	180	180	Total

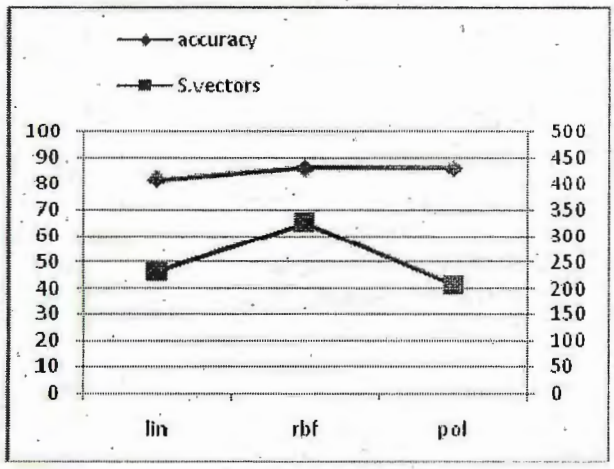

Figure 1(a) Graph of Accuracy Vs Support Vectors for 'SVM_0' Malaria cases

Table 3(b): SVM_1 Feature Selection Results

SVM_0	SVM_1	SVM_2	Performance
86.11	86.11	81.67	Accuracy (%)
208 x 8	324 x 8	232 x 8	Support Vector
15	11	0	True
140	144	147	True
7	3	0	False
18	22	33	False
0.4545	0.3333	0	Sensitivity (TP_R)
0.9524	0.9797	1	Specificity (TN_R)
0.0400	0.0204	0	(FR_R)
0.5455	0.6667	1	(FN_R)
0.5556	0.5556	0.7333	(MSE)
33	33	33	Total Positive
147	147	147	Total Negative
180	180	180	Total

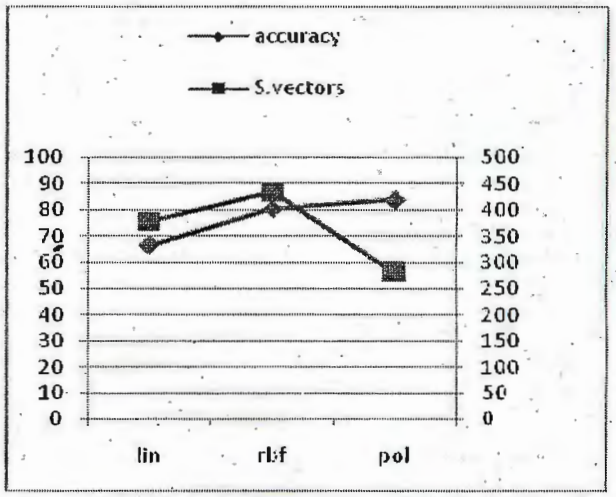

Figure 1(b) Graph of Accuracy Vs Support Vectors for 'SVM_1' Malaria cases

Table 3(c): SVM_2 Feature Selection Results

SVM_2 (pol)	SVM_2 (rbf)	SVM_2 (lin)	Performance
88.89	85.60	79.44	Accuracy (%)
147 x 8	308 x 8	195 x 8	Support Vectors
63	58	52	True Positive
97	96	9	True Negative
14	15	20	False Positive
06	11	17	False Negative
0.9130	0.8406	0.7536	Sensitivity (TP_R)
0.8736	0.8649	0.8198	Specificity (TN_R)
0.1262	0.1351	0.1802	(FR_R)
0.0870	0.1594	0.2464	(FN_R)
0.4444	0.5778	0.8222	(MSE)
698	69	69	Total Positive
111	111	111	Total Negative
180	180	180	Total

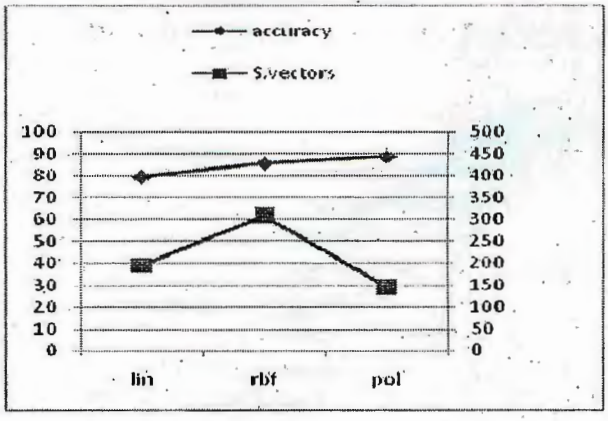

Figure 1(c): Graph of Accuracy Vs Support Vectors for 'SVM_2' Malaria cases

Dynamics for the Analysis of Malaria Transmission *Malaria Journal*, *3*(1), 2004, 29.
2. S.E. Randolph, Tick-borne disease systems. *Rev sci tech OffintEpiz*, *27*(2), 2008, 1-15.
3. P.L. Alonso, G. Brown, M. Arevalo-Herrera, F. Binka, C. Chitnis, F. Collins, F. and K. Mendis. A Research Agenda to Underpin Malaria Eradication. *PLoS Medicine*, *8*(1), 2011, e1000406.
4. T. Bousema, L. Okell, I. Felger, and C. Drakeley. Asymptomatic Malaria Infections: Detectability, Transmissibility and Public Health Relevance. *Nature Reviews. Microbiology*, *12*(12), 2014,833.
5. O. Bonuwa. Fuzzy Expert System for Malaria Diagnosis, 2014.
6. M.J. Keeling and P. Rohani. *Modeling Infectious Diseases in Humans and Animals* (Princeton University Press, 2008).
7. N. Namdev, S. Agrawal and Silkari. Recent Advancement in Machine Learning Based Internet Traffic Classification. *Procedia Computer Science*, *60*, 2015, 784-791.
8. E.M. Maina, R.O. Oboko and P.W. Waiganjo. Using Machine Learning Techniques to Support Group Formation in an Online Collaborative Learning Environment, *International Journal of Intelligent Systems and Applications*, *9*(3), 2017.
9. T. Zia, Q. Abbas and M.P. Akhtar. Evaluation of Feature Selection Approaches for Urdu Text Categorization. *International Journal of Intelligent Systems and Applications*, *7*(6), 2015, 33.
10. S. Goswami, A. Chakrabarti and B. Chakraborty. An Efficient Feature Selection Technique for Clustering Based on a New Measure of Feature Importance. *Journal of Intelligent and Fuzzy Systems*, (Preprint), 2017, 1-12.
11. M.S. Mohamad. Feature Selection Method using Genetic Algorithm for the Classification of Small and High Dimension Data. In: *Proc. Int. Symp. Info. Com. Tech., 2004* , 13-16.
12. V. Sindhwani, P. Bhattacharyya, Subrata Rakshit. Information Theoretic Feature Crediting in Multiclass Support Vector Machines. First SIAM Int. Conf. on Data Mining, 2001.
13. S. Dumais, J. Platt, D. Heckerman, M. Sahami. Inductive Learning Algorithms and Representations for Text Categorization. Proc. 7th Int. Conf. on Information and
14. T. Joachims. Text Categorization with Support Vector Machines Learning with many Relevant Features. Proc. 10th ECML. LNCS, Vol. 1398, 1998, 137–142.
15. O. Chapelle and S.S. Keerthi. Multi-Class Feature Selection with Support Vector Machines, In: *Proceedings of the American Statistical Association,* 2008, August.
16. J. Bi, K. Bennett, M. Embrechts, C. Breneman and M. Song. Dimensionality Reduction via Sparse Support Vector Machines. *Journal of Machine Learning Research*, 2003, *3*(03), 1229-1243.
17. J. Zhu, S. Rosset, R. Tibshirani and T.J. Hastie. 1-Norm Support Vector Machines. In: *Advances in Neural Information Processing Systems*, 2004, 49-56.
18. S. Ch, S.K.Sohani, D. Kumar, A. Malik, B.R. Chahar, A.K. Nema, and R.C. Dhiman. A Support Vector Machine-Firefly Algorithm based Predicting Model to Determine Malaria. *Neuro Computing*, *129*, 2014, 279-288.

BIOGRAPHY OF AUTORS

Jimoh Rasheed Gbenga is currently the Acting Dean of Faculty of Communication and Information Science (FCIS), University of Ilorin, Nigeria. He attended Universiti Utara Malaysia, Malaysia where he got Ph.D. in Information Technology. His research interests are: Information Security, Soft Computing and Machine Learning. He is a member of Computer Professionals [Registration Council of Nigeria] CPN; member of Nigeria Computer Society of Nigeria (NCS) and IEEE, Nigeria Chapter.

Abisoye Opeyemi A. was born in Ogbomoso, Oyo State, Nigeria. She attended University of Ilorin, Ilorin, Nigeria where she obtained her B.Sc., M.Sc., and Ph.D. degrees in Computer Science. She is major in Computational Intelligence, Machine Learning, Data Mining, and Soft Computing. She serves as a Lecturer I, in the Department of Computer Science, SICT, Federal University of Technology, Minna, Niger State, Nigeria from May 23rd 2007 till date. She is a member of Computer Professionals [Registration Council of Nigeria] (MCPN) since 30th June, 2010.

Uthman, Muhammed Mubashir Babatunde is currently a Senior Lecturer, Department of Epidemiology and Community Health, Faculty of Clinical Sciences, College of Health Sciences, University of Ilorin, Nigeria. He attended University of Ilorin where obtained MB, BS., and M.Ph. degrees. He is a Fellow of West African College of Physicians, Faculty of Community Health. His research interests are: Environmental Determinants of Health and Diseases. He is a member of NMA, MDCAN and APHPN.

FUNDING

This research was supported by TETFund Institutional Based Research Intervention (IBRI), University of Ilorin, Ilorin, Nigeria.