January 2021

On the efficiency of machine learning models in Malaria prediction

Ousseynou MBAYE, Mouhamadou Lamine BA [1] and Alassane SY
LIMA, Université Alioune Diop, Bambey, Senegal

Abstract.
Malaria is still a real public health concern in Sub-saharan African countries such as Senegal where it represents approximatively 35% of the consultation activities in the hospitals. This is mainly due to the lack of appropriate medical care support and often late and error-prone diagnosis of the disease. In addition, largely used diagnostic tools such as the Rapid Diagnosis Test are not fully reliable. This study proposes an extensive study of the efficiency of the most popular machine learning models for the task of Malaria occurrence prediction. We have considered patients from Senegal and have evaluated the overall precision of each considered algorithm based on sign and symptom information from various datasets. Our main result is that Random Forest, Logistic Regression, Support Vector Machine with Gaussian kernel and Artificial Neural Network exhibit very promising performance for the studied prediction problem.

Keywords. Malaria, prediction, ML, performance, evaluation, Sign, Symptom

1. Introduction

Malaria is a transmissible disease through the bites of infected female Anopheles mosquitoes. It comes with symptoms such as fever, headache, and chills in its early stage and can evolve to more severe health problems (severe anaemia, respiratory distress, etc.) often leading to death. In 2019, the number of Malaria cases worldwide has been estimated to 229 millions. The number of deaths caused by Malaria has been approximatively estimated to 409,000 in 2019; the African area represents around 94% of the reported Malaria cases and deaths in 2019, thanks to the annual world Malaria report [1].

Over the past years, many efforts have been made by governmental and non governmental organizations (e.g. WHO) to eradicate Malaria in the world. In the research field, many studies, aiming at understanding the disease from the Plasmodium mosquito point of view or proposing automated detection tools, have been conducted [2,3,4,5]. The Rapid Diagnostic Test (RDT) [5] is one of the most successful and prominent introduced tool to automatically predict whether or not a given patient suffers from Malaria. It relies on the detection of the presence of specific Plasmodium proteins, PfHRP2, pLDH and aldolase in human blood. The RDT is largely used and adopted as a standard in many Sub-saharan African countries such as Senegal. However, as proved in [5], RDT is not

[1]Corresponding Author: Mouhamadou Lamine BA, LIMA, Université Alioune Diop, BP.3400 Bambey, Senegal; E-mail: mouhamadoulamine.ba@uadb.edu.sn.

fully reliable: Section 2 shows that the precision of RDT is about 90% for datasets used in this study. Despite those advanced tools, Malaria is still a real public health issue in Africa because of the lack of appropriate care support or late and error-prone detection of the disease. Artificial intelligence is now recognized as a domain that may help medical actors in their decision-making process [6,7] and to overcome the lack of enough health resources. This paper proposes an extensive comparative study of the efficiency of the most popular machine learning models for the task of Malaria prediction. The evaluated and compared ML algorithms are Naive Bayes (NB), Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), and Artificial Neural Network (ANN). We conducted experiments on five datasets about patients living in Senegal. The raw datasets have been collected in different settings and contain clinical data such as signs, symptoms, the final diagnostic of the doctor, as well as the outcome of the RDT. As a main contribution, our tests show that Random Forest, Logistic Regression, Support Vector Machine with Gaussian kernel, and Artificial Neural Network outperform all RDT and present very high precision in the Senegalese patient datasets. The rest of the paper is organized as follows. We start by presenting the methods used in this work in section 2. Then, section 3 details the results of the intensive experimentations conducted over various datasets. Finally, we conclude this paper in section 4.

2. Methods

This work studies the problem of Malaria occurrence prediction and proposes to comparatively evaluate the efficiency of the most popular machine algorithms to tackle this issue. To this end, we rely on real data and evaluation metrics, as we detail them next

Data collection and preparation. In order to carry out our experiments in a real setting, we have collected two real world datasets about patients living in Senegal. Our first dataset, called DT1, contains medical records about patients living in distinct places in Senegal. and has been collected in 2016 during the *"Grand Magal of Touba"*, a religious event in Senegal gathering several million of persons every year [21]. The second dataset, denoted by DT2, contains clinical records about patients living in Diourbel, Thies, and Fatick regions where the prevalence of Malaria is very high. After the collection step, we have conducted data *cleaning*, *transformation*, and *imputation* tasks on the raw datasets in order to deal with noisy information and missing values. We have then proceeded to *feature selection* which enables to only retain the data attributes (or variables) such as *lack of appetite, tiredness, fever, cephalalgia, nausea, arthralgia, digestive disorders, dizziness, chill, myalgia, diarrhea, and abdominal pain* pertaining for our study. For privacy reasons and certain restrictions in the use of the data, we have ignored patient personal data. Table 1 summarizes the main statistics of each dataset after preparation and the precision of RDT as well. We synthetically generated from DT1 and DT2 three addi-

Dataset	Variables	Observations	Variables types		Classes		Precision of RDT
			Numeric	Boolean	Malaria	not Malaria	
DT1	16	21083	2	14	614	20469	90.23%
DT2	16	5809	2	14	5108	701	90.49%

Table 1. Main statistics of real datasets DT1 and DT2

tional datasets DT3, DT4 and DT5 respectively obtained by (i) concatenating DT1 and DT2 : (ii) selecting 2354 patients who have been tested negative for Malaria from DT1 and add them to DT2 in order to obtain a balanced dataset; and (iii) oversampling DT1 using *SMOTE algorithm* in order to have the same number of individuals in both classes.

Machine learning models. We have considered and compared the six most popular machine learning approaches [9,10] which are Decision Tree (DT) [11], Random Forest (RF) [12], Naive Bayes (NB) [13], Logistic regression (LR) [14], Support Vector Machine (SVM) [17], Artificial Neural Network (ANN) [18]. All of them are supervised learning algorithms, i.e., require a *training phase* with labelled data.

Experimentation setting. We have trained and validated each algorithm over every dataset using *stratified-5-fold cross-validation*[2] and the same experimentation environment. We have relied on ML implementation of the algorithms available with the *Scikit-Learn Python library*. To evaluate the efficiency of each algorithm, we finally measured the *precision*, *recall*, *f1-score*, *AUC*[3], and *specificity* of each algorithm on each dataset.

3. Results and discussion

Table 2 details the performance of the tested ML algorithms on our various datasets. More specifically, Figures 1 and 2 respectively compare the precision values and the AUC scores. While an one-all-fits algorithm cannot be deduced from our tests, by closely

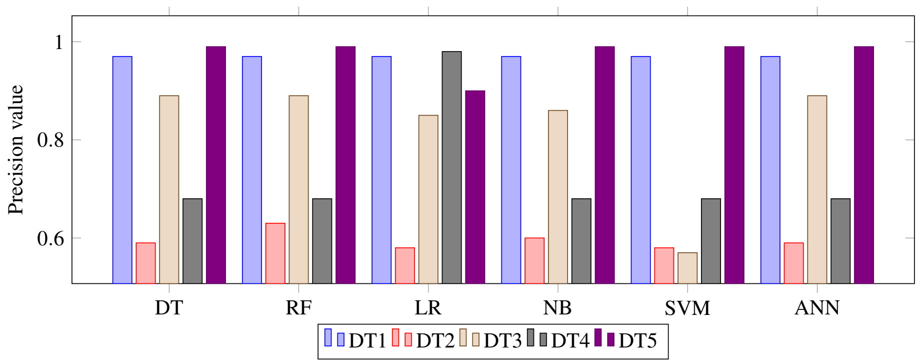

Figure 1. Precision values of compared classifiers on different datasets

analyzing the overall performance measures one can observe that RF, LR, SVM, and ANN generally outperform the others for each dataset. Another important result is that ML algorithms have better precision than RDT at least for DT1.

4. Conclusion

In this study, six ML models have been extensively tested and compared over various datasets in order to evaluate their performance for the task of predicting Malaria occur-

[2] `https://scikit-learn.org/stable/modules/cross_validation.html`

[3] `https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=fr`

ML Algorithms	Datasets	Precision	Recall	F1-score	AUC	Specificity
DT	DT1	0.97	1	0.98	0.78	0.05
	DT2	0.59	0.48	0.48	0.64	0.80
	DT3	0.89	0.85	0.87	0.86	0.69
	DT4	0.68	0.57	0.62	0.70	0.74
	DT5	0.99	0.84	0.91	0.76	0.58
RF	DT1	0.97	1	0.99	0.81	0.07
	DT2	0.63	0.34	0.44	0.64	0.85
	DT3	0.89	0.85	0.87	087	0.70
	DT4	0.68	0.56	0.62	0.70	0.74
	DT5	0.99	0.84	0.91	0.76	0.60
LR	DT1	0.97	1	0.99	0.79	0.05
	DT2	0.58	0.36	0.44	0.63	0.81
	DT3	0.85	0.88	0.86	0.86	0.55
	DT4	0.98	0.56	0.92	0.70	0.72
	DT5	0.90	0.78	0.88	0.84	0.75
NB	DT1	0.97	1	0.99	0.81	0.00
	DT2	0.60	0.34	0.43	0.63	0.83
	DT3	0.86	0.87	0.86	0.85	0.60
	DT4	0.68	0.59	0.63	0.70	0.73
	0.99	0.82	0.90	0.84	0.71	0.71
SVM	DT1	0.97	1	0.99	0.84	0.00
	DT2	0.58	0.05	0.09	0.62	0.97
	DT3	0.57	0.86	0.86	0.85	0.64
	DT4	0.68	0.58	0.62	0.70	0.73
	DT5	0.99	0.86	0.92	0.80	0.62
ANN	DT1	0.97	1	0.99	0.84	0.04
	DT2	0.59	0.40	0.48	0.65	0.80
	DT3	0.89	0.85	0.87	0.87	0.69
	DT4	0.68	0.58	0.62	0.70	0.75
	DT5	0.99	0.84	0.91	0.79	0.65

Table 2. All performances measures of our ML models over our five datasets

rence in a patient knowing his signs and symptoms. The results obtained show that some of them are promising and overcome RDT in particular settings. As future work, we plan to study the implementation of an ensemble method for predicting Malaria occurrence built on the algorithms offering the best performances in our present study.

References

[1] Organization WH. 2019 World Malaria Report; 2019. `https://www.who.int/malaria/publications/world-malaria-report-2019/en/`.

[2] Garrido-Cardenas J, Cebrian-Carmona J, Gonzalez-Ceron L, Manzano-Agugliaro F, Mesa-Valle C. Analysis of Global Research on Malaria and Plasmodium vivax. International Journal of Environmental Research and Public Health. 2019 05;16.

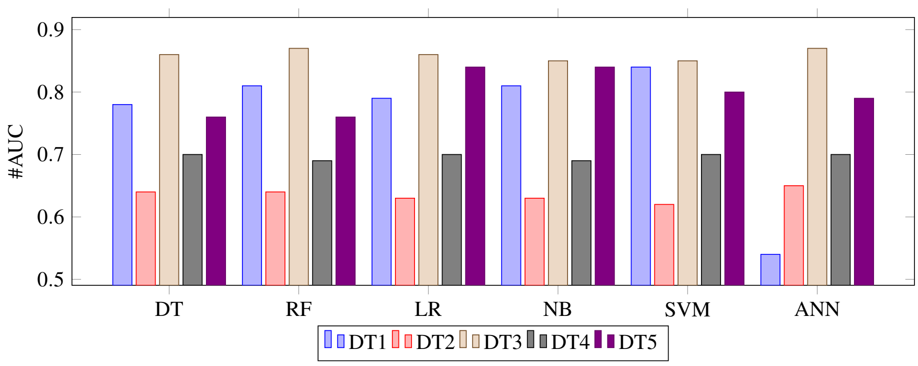

Figure 2. Comparison of the AUC values of the ML algorithms on different datasets

[3] Lepes T. Review of research on malaria. Bulletin of the World Health Organization. 1974;50(3-4):151 – 157.
[4] Ermert V, Fink A, Jones A, Morse A. Development of a new version of the Liverpool Malaria Model. Malaria journal. 2011 02;10:35.
[5] Houzé S. Rapid diagnostic test for malaria. Bull Soc Pathol Exot.
[6] Mitchell TM, et al. Machine learning. McGraw-Hill,In; 1997.
[7] Yadav AMSMKA Abhishek. Better Healthcare using Machine Learning. International Journal of Advanced Research in Computer Science. 2010;9(1).
[8] Mbaye O, Ba ML, Camara G, Sy A, Mboup BM, Diallo A. Towards an Efficient Prediction Model of Malaria Cases in Senegal. In: International Conference on Innovations and Interdisciplinary Solutions for Underserved Areas. Springer; 2019. p. 173–188.
[9] De Oliveira H, Prodel M, Augusto V. Binary Classification on French Hospital Data: Benchmark of 7 Machine Learning Algorithms. In: 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE; 2018. p. 1743–1748.
[10] 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.
[11] Rokach L, Maimon O. In: Decision Trees. vol. 6. Springer; 2005. p. 165–192.
[12] Breiman L. Random Forests. Machine Learning. 2001;45(1):5–32.
[13] Kaviani P, Dhotre S. Short Survey on Naive Bayes Algorithm. International Journal of Advance Research in Computer Science and Management. 2017 11;04.
[14] Morgan SP, Teachman JD. Logistic Regression: Description, Examples, and Comparisons. Journal of Marriage and Family. 1988;50(4):929–936.
[15] Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making. 2019;19(1):1–16.
[16] Wang PW, Lin CJ. Support Vector Machines.; 2014.
[17] Evgeniou T, Pontil M. Support Vector Machines: Theory and Applications. In: Studies in Fuzziness and Soft Computing. vol. 2049; 2001. p. 249–257.
[18] Mehlig B. Artificial Neural Networks. In: arXiv; 2019. 1901.05639.
[19] Anderson JA. A simple neural network generating an interactive memory. Mathematical biosciences. 1972;14(3-4):197–220.
[20] Raschka S. Python machine learning. Packt Publishing Ltd; 2015.
[21] Sokhna C, Mboup BM, Sow PG, Camara G, Dieng M, Sylla M, et al. Communicable and non-communicable disease risks at the Grand Magal of Touba: The largest mass gathering in Senegal. Travel Medicine and Infectious Disease. 2017;19:56 – 60.