Instructions for the Preparation of an Electronic Camera-Ready Manuscript in LaTeX

Book Production MANAGER[a,1], Second AUTHOR[b] and Third AUTHOR[b]

[a] *Book Department, IOS Press, The Netherlands*

[b] *Short Affiliation of Second Author and Third Author*

Abstract. .

Keywords.

1. Introduction

Malaria, also known as *fiévre des marais* in French, is an infectious disease caused by a mosquito of the type *Plasmodium*. In its severe form, the disease can lead to *yellow skin*, *seizures*, *coma* or *death*. As a result, Malaria is now recognized and treated as a serious health problem worldwide by the World Health Organization (WHO), and particularly pandemic in Sub-Saharan Africa. In 2018, an estimated 228 million cases of Malaria occurred worldwide, thanks to the 2019 World Malaria Report [**?**]. Most Malaria cases in 2018 were in the WHO African Region (213 million or 93%). In the special case of Senegal the problem is acute because of the use of inappropriate care support means coupled with an often late and error-prone diagnostic from the local medical staff. Setting up a reliable way to predict the disease when a patient visits a doctor become then crucial in order to avoid its evolution towards a critical state.

Over the past years, many efforts have been done by governmental and non governmental organizations to eradicate Malaria: actions continuously conducted by the WHO are real examples of those. 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 [**?**,**?**,**?**,**?**]. The Rapid Diagnostic Test (RDT) [**?**] 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 specific Plasmodium proteins, PfHRP2, pLDH and aldolase. The RDT is largely used and adopted as a standard in many health structures in Sub-African countries because of its simplicity to utilize and does not require any specific domain knowledge. However as highlighted in [**?**] the RDT is not fully reliable: in Section **??** we show that the precision of the RDT

[1] Corresponding Author: Book Production Manager, IOS Press, Nieuwe Hemweg 6B, 1013 BG Amsterdam, The Netherlands; E-mail: bookproduction@iospress.nl.

is about 90% for the real datasets used in this study. The Liverpool Model on Malaria (LMM) extended in [**?**] is an example of mathematical model that tries to model the parasite life cycle. It simulates the spread of Malaria at a daily resolution using the average daily temperature and the accumulated precipitation over 10 days. the final goal is to build a climate- or weather-driven Malaria model, allowing for a better understanding of Malaria transmission dynamics. Clearly LMM is not a diagnostic system. We defer the reader to Section **??** for an exhaustive review of the literature. Despite existing works, the accuracy of Malaria prediction is still a concern: used mechanisms, e.g. domain knowledge and RDT, in Senegal are error-prone.

With the development and increasing adoption of automated tools in the health field, machine learning (ML) [**?**,**?**] applications might help medical actors in their decision-making process. There are already some attempts to apply ML techniques for the prediction or a better understanding of various diseases, e.g. [**?**,**?**]. For example, machine learning is used to analyze blood data obtained from high definition microscopic screenshots in [**?**]. On the other hand, logistic regression has been tested in [**?**] for the prediction of Malaria and provides promising results.

In this paper, we propose an extensive comparative study of six machine learning algorithms, among the most popular for the prediction of Maria in Senegal. 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) [**?**]. Whereas the four first algorithms are simple models, the two last ones are built on more complex learning strategies. RF is an ensembling model and ANN performs Deep Learning. We conducted experiments on five datasets based on the two real world datasets about Senegalese citizens that suffer or not from Malaria. These two datasets have been collected in two different contexts and contain clinical data such as sign, symptom and final diagnostic of patients living in distinct locations in Senegal (for the first dataset) or within the same area (for the second dataset). Those patients have been examined by doctors in given health services and their clinical data recorded: for each patient the final diagnostic is provided with the corresponding signs and symptoms. The outcome of the RDT is also provided. To evaluate the performance of every considered algorithm we have considered common measures of the accuracy of a prediction system that are *Precision*, *Recall*, *F1-score*, *True Positive Rate*, and *False Positive Rate* on both datasets augmented with semi-synthetic datasets which are obtained after imputation in order to deal with missing values.

Our main result is that RF, SVM with Gaussian Kernel and ANN are promising and offer the best overall accuracy to predict the appearance or not of the disease with precision, recall and F1-score at least equal to 92%, 85% and 89% respectively on both datasets. More specifically, those three learning approaches outperform the RDT which represents the baseline automatic diagnostic tool largely adopted as a standard within the Health system in Senegal.

The rest of the paper is structured as follows. We first review the literature of existing research works dealing with Malaria in Section **??**. In Section **??**, we then provide a detailed description of our two real world datasets which contain medical records about patients living in Senegal. More precisely, we present the characteristics of each dataset, their imputation to deal with missing values and the precision values of RDT. In the sequel, we briefly describe in Section **??** the six ML algorithms that are evaluated and compared in this study. We present our experimentation setting, considered performance measures and discuss about the results of the experiments in Section **??**. Finally, we conclude this paper in Section **??**.

References

[1] Petitti DB, Crooks VC, Buckwalter JG, Chiu V. Blood pressure levels before dementia. Arch Neurol. 2005 Jan;62(1):112-6.

[2] Rice AS, Farquhar-Smith WP, Bridges D, Brooks JW. Canabinoids and pain. In: Dostorovsky JO, Carr DB, Koltzenburg M, editors. Proceedings of the 10th World Congress on Pain; 2002 Aug 17-22; San Diego, CA. Seattle (WA): IASP Press; c2003. p. 437-68.