Towards an efficient Prediction Model of Malaria Cases in Senegal

Ousseynou Mbaye, Mouhamadou Lamine Ba
Gaoussou Camara, and Alassane Sy

Université Alioune Diop de Bambey, Bambey, Senegal
firstmidlle.last@uadb.edu.sn

Abstract. Among the most deadly disease in the world, Malaria remains a real flail in Sub-saharan Africa in particular. In countries like Senegal, such a situation is acute due to the lack of high quality healthcare services and well-formed staffs able to perform accurate diagnosis of diseases that patients suffer from. This calls for the need of finding automated tools to help medical actors in their decision making process. In this paper, we present first steps towards an efficient way to automatically diagnosis Malaria occurence or not based on patient signs and symptoms, and the outcome from the quick diagnosis test. Our prediction approach is built on the logistic regression function. First expermients on a real world patient dataset, as well as a semi-synthetic dataset, show promising performance results regarding the effectiveness of the proposed approach.

Keywords: Malaria · Diagnosis · Data imputation · Prediction Model.

1 Introduction

Malaria is one amongst the most deadly disease in the world, especially in sub-saharan Africa countries such as Senegal. Malaria is caused by parasitic single-celled microorganisms belonging to the Plasmodium group; it is an infectious disease which is transmitted to human being through bites from infected female Anopheles mosquitoes. Someone who suffers from Malaria may present symptoms that typically include fever, tiredness, vomiting, and headaches. In its severe form, the disease can cause yellow skin, seizures, coma or death.

Studied problem and motivations. According to the last report [25] about the propagation of Malaria disease around the world, published in November 2017 by the World international Health Organization (WHO in short), 216 millions of cases have been reported in 2016. As a result, the number of cases has significantly increased when compared to the 211 millions of reported Malaria patients in 2016. As for the number of death due to Malaria, it does not decrease between 2016 and 2017 (446.000 vs. 445.000) despite the huge effort made by governements and non-governmental organization to improve healthcare services and the awareness strategies, especially in critical areas. When analyzing the statistics above in details, one can easily notice that the burden of the Africa region of the World international Health Organization is colossal.

Indeed, 90% of Malaria cases and 90% of deaths due to the disease were located in this area in 2016. More specifically, 80% of the burden in terms of morbidity is distributed in fifteen countries, all located in Sub-saharan Africa except India. This demonstrates that Malaria is a real flail in Sub-saharan Africa states and Senegal is not spared at all. We investigate in this study an efficient approach to predict, using machine learning, the occurence or not of Malaria when a patient has to be diagnosed. Given the patient signs and symptoms, as well as the result from the quick diagnosis test, our solution should be to automatically tell if she suffers from Malaria or not with a high accuracy.

Malaria is an acute problem in Senegal due mainly to the lack of high quality health-care services and well-formed staffs able to perform accurate diagnosis of diseases that patients suffer from. Over the past years, the government with the help of international organizations have tried to eradicate Malaria by implementing various proactive and reactive solutions to fill the gap in terms of services and human resources. However, the mortality rate is still very high, e.g. in underserved areas, areas without required health-care needs, uneducated people, population with low income, etc. Most of these deaths cases are reported to be caused by inaccurate diagnosis, sometimes incomplete leading to a bad prediction of the exact type of Malaria. On the other hand, Malaria occurence or complication can often occur during popular events (for instance religious events such as the Grand Magal of Touba [19]) which gather thousands of persons from everywhere in the country during a short time period. During those popular events, non-permanent medical points are set in order to assist and treat ill persons; the staff in a given health point might be formed sometimes by only volunteers without advanced medical skills. Each of these medical points might receive and treat hundreds of patients each day with some of them potentially suffering from Malaria. This appeals for the need of finding automated tools to help medical actors in their decision making process, and thereby to improve provided healthcare services.

Proposed diagnosis approach. In this paper we present first steps towards an efficient manner to automatically diagnosis Malaria occurence or not based on patient signs and symptoms, and the outcome from the quick diagnosis test. We define our diagnosis task as a classical binary classification problem by considering two classes: “Malaria” and “not-Malaria”. Given a patient data, our main goal is to properly find to which class the patient belongs. To solve this classification problem we rely on machine learning and use the logistic regression function as the basis of our prediction approach. Machine learning has been largely used in several domains (e.g. Health Informatics [8]) for various purposes whereas logistic regression has demonstrated its efficiency when dealing with a binary classification problem.

As an application scenario, we focus on predicting Malaria cases in Senegal. At this end, we use a large volume of patient record dataset collected during the most popular religious event in Senegal from the different installed health points, namely more than twenty points that deadly receive hundreds of patients. As an immediate result of this work, we introduce a data preparation pipeline in order to (i) explore the dataset for profiling purpose; (ii) only retain records related to Malaria; (iii) clean and transform attributes, as well data values, into the extracted Malaria dataset; and (iv) impute missing values (there were lot of missing values in the collected health

dataset as reported in Section3). Such a data preparation pipeline has been realized using *OpenRefine* (formely Google Refine) to perform various cleaning and profiling tasks on our raw atient dataset and *missForest*, a robust algorithm for imputing missing data of diverses types; see Section 3 for more details. Expermients on the real world patient dataset, augmenting with a semi-synthetic dataset, show promising performance results regarding the effectiveness of the proposed approach.

Paper organization. The remaining of the paper is organized as follows. We summarize the related work on data imputation and binary classification methods in Section 2. In Section 3 we introduce a data preparation pipeline on the raw collected patient records for the prediction phase. We then present our prediction model for Malaria cases in Section 4. Experiments and performance analysis on the collected real-world dataset, as well as a semi-synthetic dataset, are detailled in Section 5 before we conclude in Section 6.

2 Related work

In this section, we summarize the sate-of-the-art research on Malaria in general, and in particular the use of machine learning techniques to tackle the various aspects related to one of the major healthcare problems worldwide which is Malaria.

As it is well-known, Malaria is caused by the bite of the female Anopheles, the most dangerous of which is Plasmodium falciparum. Many early works have been consequently focused on the study of the evolution and the distribution of the responsible mosquito, mainly with the goal to detect or diagnosis the severity of the disease given an infected patient [10,5]. Recent research on Malaria have largely adopted machine learning and showed its ability to solve various aspects of the disease. Most of these machine learning based techniques are based on the analysis of blood data obtained from high-definition microscopic screenshots as in [12]. The authors in [12] propose an unsupervised learning algorithm that detects and determines the types of infected blood cells. Used prediction approach consists of quantifying the amount of plasmodium parasites in a blood smear. In the same research intuition of harnessing blood, the Jordan-Elman neural network classifier introduced in [7], on the other hand, to quickly determine the occurrence of Malaria and its severity level as well: the neural network analyzes the features of the blood data of the patients. Still using ML, DIAZ et al. have proposed in [9] a semi-supervised algorithm enable to quantify and classify the erythrocytes infected by Malaria parasistes through microscopic images. The orginiality of this work comes from its usability even in the presence of thin blood drandruff infected by falciparum Plasmodium for the quantification and the classificationi tasks. Besides blood data, sign and symptom records were also used to study Malaria with ML methods. Indeed, decision trees based approach has been proposed in Nigeria [23] to predict the occurrence of Malaria given diagnostic data. However a decision tree suffers from various limitations as a classifier. Indeed it can easily overfit or can be extremely sensitive to small pertubations in data for instance. Even though we both rely on signs and symptoms, the prediction model in [23] differs from ours on numerous facets: our model is built upon logistic regression and is trained using also inputs from the quick

diagnosis test. In addition, we apply our method in the context of patients living in Senegal. An example of previous work that has used logistic regression is that of Farida et al. in [3]. The logistic regression is exploited there for the selection of features in order to construct stable decision trees. The decision trees are then used to predict the severity criteria of Malaria in the context of Afghanistan.

In the same line of works applying machine learning, in [16], Pranav et al. propose Malaria likelihood prediction model built on a deep reinforcement learning (RL) agent. Such a RL predicts the probability of a patient testing positive for Malaria using answers from questions about their household. In the presented approach the authors have also dealt with the problem of determining the right question to ask next as well as the length of the survey, dynamically. Moreover, statistically enhanced rule-based classification model to diagnose malaria has been proposed in [6]. A corresponding prototype which incorporates the rules and statistical models have been implemented; the main goal of the study was to develop a statistical prototype to perform clinical diagnosis of malaria given its adverse effects on the overall healthcare, yet its treatment remains very expensive for the majority of the patients to afford.

To the best of our knowledge this is the first work in Senegal that attempts to provide a prediction model for identifying the occurrence of Malaria given patient data.

3 Data Preparation

In this section, we detail the data preparation pipeline followed to obtain a proper Malaria dataset for the prediction phase. We start by presenting the used data cleaning and normalization techniques.

3.1 Data cleaning and normalization

In order to set up an efficient prediction model for Malaria cases in Senegal, we have relied on a real-world patient dataset for validation purposes. The dataset was extracted in 2016 during the Grand Magal of Touba [19]. An estimated 4-5 million individuals gather each year in the holy city of Touba, Senegal during the Grand Magal religious pilgrimage. Several health points are set during this religious event; every point dairly receives hundreds of patients, some of them suffering from Malaria. The patient data we are using here have been manually recordered from these points in registers as no electronic health management system does exist.

In detail, the dataset consists of thousands of patient records having each 16 attributes. Some of these attributes (also knwon as features) comprise personal data about the patient, but also patient signs and symptoms reported by the doctor who took in charge the patient. The other attributes describe clinical data such as information about the final diagnosis of the doctor (the disease that the patient suffers from), the income of the quick diagnosis test, and the status (i.e. admission, dead or put under observation) of the patient. For privacy concerns and some restrictions in data use, we have disregarded personal data about the patient during this work. Due to the fact that patient records have been collected manually in registers, we have noticed many inconsistencies such as misspellings, same attribute values with different writings (e.g., “DIARRHEE INFECTIEUSE” and “INFECTIEUSE DIARRHE”), and multi-valued attributes (e.g. sign and

symptom reported values). We use OpenRefine [13,1] to first clean and then normalize values in the patient dataset.

OpenRefine (formerly called Google Refine) is a powerful open source tool that allows researchers or scientists to accomplish the data wrangling activity, i.e. working with messy data: cleaning it; transforming it from one format into another; and extending it with Web services and external data. We used the following methods provided by OpenRefine to pre-treat our raw dataset.

- **Text filter function:** text filter enables to expore attribute values, clean them, and to identify those that may have many variants.
- **Transform functions:** OpenRefine provides two different Transform functions: preset transformations for resolving trivial formatting issues like trimming whitespaces and advanced transformations based on the OpenRefine Expression Language (GREL) to normalize data in batch or split them. This second class of transformations is very useful, especially when the number of piece of data values to normalize is very important (doing the same task manually would be time-consuming and prone to errors). For instance, GREL allows to use a simple regular expression all variants of a symptom in the Symptom column.
- **Cluster and edit function:** Clustering option in OpenRefine also provides users with methods to merge and normalize variations across the dataset. The power of clustering is that it is able to automatically detect small data variations which follow a certain pattern.

As for the special case of multi-valued attributes such as symptom and diagnosis columns in our raw dataset, we splitted them into multiple values in distinct columns. Indeed, in the raw dataset information like the symptoms a given patient suffer from were stored in a single column, separated with the special character '+', e.g. "DOULEUR ARTICULAIRE" + "DOULEUR PELVIENNE" + "VOMISSEMENTS".

After this step of data cleaning and normalization, we have proceeded to the extraction of Malaria features.

3.2 Extraction of Malaria features

To properly study Malaria, one needs to have a patient dataset with the main features of the disease. Unfortunately, some of these Malaria features were not explicitly specified in our raw dataset. As a result, we have inferred twelve new attributes that better describe the signs and symptoms of Malaria according to experts in the health domain. Those new attributes are : *lack of appetite*, *tiredness*, *fever*, *cephalalgia*, *nausea*, *arthralgia*, *digestive disorders*, *dizziness*, *chill*, *myalgia*, *diarrhea*, and *abdominal pain*. We have then added the new attributes in our dataset and transformed this latter accordingly by filling the value of each new attribute based on the list of reported signs and symptoms for every patient.

A medical diagnostic is the results the analysis of the reported signs and symptoms; in general such a diagnostic is further confirmed by a medical test. Since our raw dataset does not contain only information about patients from Malaria, yielding to records with various diagnostics. For the purposes of our study, we replaced any diagnostic that is

not Malaria by the class "Not-Malaria". At this step of our data preparaton pipeline, we came up with a patient dataset that contains required Malaria features. However, our Malaria was not yet complete and ready because of values missingness. As a last step, we have completed our dataset by using a robust data imputation approach.

3.3 Missing Data Imputation

As shown in Table 1, we observed many missing values in our dataset, affecting the majority of the data attributes. Such missing values should not be ignored as data completeness and quality are very important when dealing with a prediction problem; this could negatively impact the accuracy of our prediction and should be treat appropriately. One has to note that machine learning relies on complete dataset. The sources and types of missing values can be various [22]. In our context, missingness is *not completely random* and can be due to an incomplete knowledge of the patient data, the fact that the medical staff do not specify a attribute value when it is not observed, or a difficulty for the patients to properly describe some piece of information (e.g. related to the signs or symptoms of their diseases) at the diagnostic time. Since they might have a certain relationship between attribute values for the same patient, or even a correlation between patient records, we decide to solve our problem of missing values by using imputation algorithms instead of choosing arbitrary values or removing records with missing values.

Data imputation is often used in the machine learning field when dealing with missing information. Many algorithms have been proposed in the literature [18,22], depending on the nature of the missingness or the type of data. MissForest [21] has been proved to be efficient at the presence of various types (e.g. numerical data, string, categorical data, ect.) of data simultaneously as in our case. The algorithm missForest relies on Random Forest, a non-parametric prediction method that is able to to deal with mixed-type data and allows for intreactive and non-linear regression effects. Such a imputation algorithm aims at handling any type of input dataset by minimizing (when possible) assumption about the structural apsects of the data. Given an input dataset, missForest solves the missing data problem using an iterative imputation scheme by training a Random Forest on observed values in a first step, followed by predicting the missing values and then proceeding iteratively until convergence.

We have applied missForest on our Malaria patient dataset by using its open-source Python implementation [2]. We shall study and prove in Section 5 the accuracy of the prediction made by the algorithm with the hep of the normalized root mean squared error metric.

4 Prediction Model

To learn from labelled patient dataset and be able to properly predict the occurence or not of Malaria given a new patient, we harness the logistic regression function as our *classifier*. In this section, we briefy recall the basic of the logistic regression function and how it can act as a binary classifier. We start by introducing the binary classification problem we have to solve in the study.

Attribute name	#missing_values
Lack of appetite	21068
digestive disorders	21062
Loss of weight	21017
Arthragia	20940
Chill	20925
Nausea	20874
Myalgia	20870
Tiredness	20713
Diarrhea	20481
Vomit	20051
Abdominal pain	19770
Dizziness	19628
Fever	18245
Temperature	17636
Arterial pressure	16924
Cephalalgia	15370
Diagnostic	2875
Quick diagnostic test	76

Table 1. The number of missing values per attribute

4.1 Binary classification problem

Let us assume two given classes of Malaria diagnostic: *Malaria* and *Not-Malaria*. We also consider P and C as respectively the set of patients and a prediction model. A patient p in P is defined by a set of pairs $(a_1, v_1), (a_2, v_2), \ldots, (a_n, v_n)$ where a_i and v_i, for each $1 \leq i \leq n$, respectively corresponds to a given Malaria feature and its associated value defined as follows.

$$v_i = \begin{cases} 1 \text{ if } a_i \text{ is observed} \\ 0 \text{ otherwise} \end{cases} \quad (1)$$

Definition 1. *(Our prediction problem) We define our binary classification problem for the prediction of the occurrence or not of Malaria on a given patient dataset as a mapping* C *of every patient p in* P *to one and only one class in* {*Malaria, Not-Malaria*}. *Formally, we present such a mapping as* C*:* P $\mapsto$ {*Malaria, Not-Malaria*}.

We define and use C with the help of the logistic regression for the specific purpose of our study.

4.2 Logistic regression

The logistic regression (a.k.a the logit function) is a statistical model used in the machine learning domain for binary classification [17]. In its basic form, it is based on a logistic function to model a binary dependent variable [11,20]. As an input, the logistic regression takes qualitative or/and ordinal predictive variables (e.g. the presence or not

of fever given a patient) in order to measure the probability of the outcome (e.g. the occurrence or not of the Malaria) by using the *Sigmoid function*. Figure 1 shows the shape of the curve of the Sigmoid function.

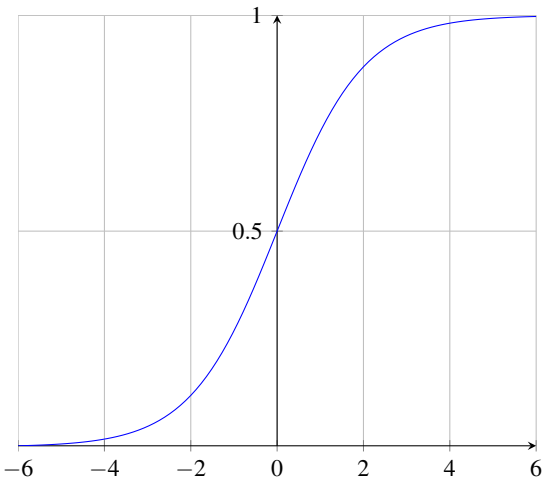


Fig. 1. The curve of the Sigmoid function

The logistic regression is one of the most used multi-valued models in epidemiology [4,15] (i.e. the study of the incidence, distribution, and possible control of diseases and other factors relating to health issues that can affect groups of population). In such a context, the variable to explain is often the occurrence or not of an event like a disease and the explanatory variables, i.e. the features, are those that highly impact the occurrence of this event, i.e. variables assessing the exposure to a risk factor or a protective factor, or a variable representing the confusion factor. The main interest of using logistic regression is its ability to quantify the strength of the relationship between each explicative variable and the variable to explain, given the other variables integrated to the model [4].

Formal definition of our regression model. Let us assume that Y represents the variable we are trying to explain in this study, i.e. the variable which models the occurrence or not of Malaria and whose two classes *Malaria* and *Not-Malaria* are respectively denoted by M+ and M-.

In the special case of only one explicative variable a (which case corresponds to a simple regression), formally the model is written as follows.

$$\Pr(\text{M+} \mid a) = \frac{e^{\alpha+\beta\times a}}{1+e^{\alpha+\beta\times a}} \tag{2}$$

where the coefficients α abd β are the parameters of the model.

$\Pr(\text{M+} \mid a)$ measures the probability of the occurrence of Malaria if the variable a is observed. Figure 1 represents the corresponding logistic function *f(a)*. Again, the main interest of this function lies in the simplicity of reaching an estimation of an odds ratio (OR) which measures the strength of the association between the disease M and an exposure variable in a regression analysis. Indeed if the value exposure variable is either 0 (the variable is not observed) or 1 (the variable is observed) as in our setting, the model enables to obtain after some simplifications $\text{OR} = e^{\beta}$. The coefficient β of the exposure variable in the logistic model is then the logarithmic of the odds ratio which measures the relationship between the explanatory variable (sign or symptom) and the disease (Malaria); this eases the analysis of the results of the logistic regression.

An extension of the simple regression to a model with multiple variables (a.k.a multiple regression) is straightforward as we show with the formula below.

$$\Pr(\text{M+} \mid a_1, a_2, \ldots, a_n) = \frac{e^{\alpha+\sum_{i=1}^{n}\beta_i \times a_i}}{1+e^{\alpha+\sum_{i=1}^{n}\beta_i \times a_i}} \tag{3}$$

where to every variable a_i is associated a coefficient β_i. The corresponding odds ratio OR_i, quantifying the relationship between a_i and M+ is equal to e^{β_i}.

Model optimization. The question that generally raises when using a multiple regression approach is how to select the minimum set of variables amongst the a_i's that better explain the variable Y. Several optimization strategies are possible to obtain the best final prediction model which takes into account the maximum of information while restricting as much as possible the number of explanotory variables in order to ease the analysis of the results: *stepwise descendant* and *stepwise ascendant* are the most used approaches. Both approaches apply an iterative regression by first including in the model the variable that presents the best determination coefficient and then by adding the variable which improves this coefficient and so on for stepwise ascendant. For stepwise decendant, the entire set of variables are considered at the beginning and variables are gradually excluded from the model, depending on those which do not significantly improve the determination coefficient.

We next present the results of our prediction of Malaria cases by using our logistic regression model above on real-world patient datasets.

5 Experimentation and results

In this section, we present the performance of our prediction model for Malaria occurrecne through an analysis of the results of the experimentations we have conducted on real-world datasets and a semi-synthetic dataset. We start by presenting our experimentation setting.

5.1 Experimentation setting

We ran tests on three different datasets using the Python Implementation of the logistic regression function. To impute missing data, we have used the R package of the algorithm missForest.

Our datasets. We collected and used a real-world patient dataset from the different health points which were set during the Grand Magal of Touba in 2016. We also generated and used two variants of this real-world dataset. The description of the characteristics of our raw real-world dataset, as well as the data preparation pipeline that we have proposed in order to clean, normalize and impute information, are given in Section 3; we refer to this cleaned and complete real-world patient dataset by DT1.

We generated the first variant, denoted by DT2, of the raw real-world patient dataset by removing records with missing attributes, instead of using an imputation algorithm that will predicts values for missing information. Such a variant will help to study the impact of removing records with missing values in the prediction accuracy.

The second variant, called DT3, is a semi-synthetic dataset which has been set up by using a sampling strategy over our raw real-world dataset. Indeed when we have performed some explanatory analysis on the real-world dataset have been revealed that the dataset was not balanced, i.e. it shows strong between class imbalance; the amount of records about patients suffering from Malaria was largely less than the number of patients that do not suffer from Malaria as shown in Figure 5.1. Harnessing sampling approaches may enbale to obtain a balanced semi-dataset regarding the two classes to predict. To solve our problem of imbalanced dataset, we used the algorithm SMOTE [24], which is a synthetic minority oversampling technique, through its Python implementation in the package *imbalanced-learn* [14]. SMOTE consists of predicting a sample of synthetic dataset based on the value of the minority class of the targeted class (here the attribute Diagnostic). It randomly chooses the k-nearest neighbours of a given record in order to randomly create new observations. We have applied an oversampling of the minority class into our patient dataset for generating a semi-synthetic dataset DT3 containing the same number of records for both classes.

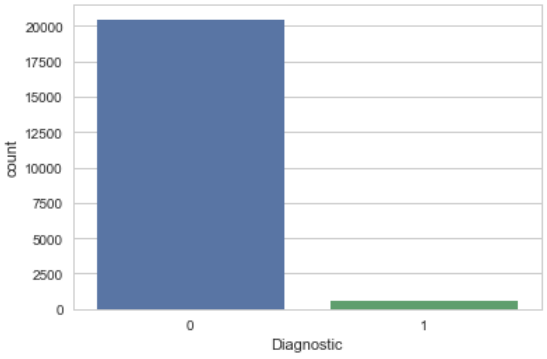


Fig. 2. The number of records by class

Prediction model setting. In order to set up our logistic regression-based classification model, we rely on the Python implementation of the logistic regression in the *sklearn*

library[1]. This python package defines the logistic regression with the required input parameters, as well as optimization strategies, to properly perform binary classification using the best final model. For the purposes of our tests, we have used the following input parameters of the logistic regression:

- **random_state**: it models the seed of the pseudo random number generator to use when shuffling the data. Its value is set to 0 as we do not need to shuffle data in our experimentations.
- **class_weight**: weights associated with classes. We set it to *None*, i.e. all classes are supposed to have weight one.
- **dual**: dual or primal formulation. Dual formulation is only implemented for l2 penalty with liblinear solver. This parameter is set to *False* as the number of samples is greater than the number of features.
- **fit_intercept**: useful if a constant (a.k.a. bias or intercept) should be added to the decision function. Consequently, we fixed the fit_intercept to *True*.
- **intercept_scaling**: this parameter, set to 1, is useful only when the solver "liblinear" is used and fit_intercept is set to True.
- **max_iter**: maximum number of iterations taken for the solvers to converge.
- **multi_class**: if the option chosen is "ovr", then a binary problem is fit.
- **n_jobs**: number of cpu cores used when parallelizing over classes if multi_class="ovr". This parameter is ignored when the solver is set to "liblinear" regardless whether "multiclass" is specified or not.
- **penalty**: this parameter is used to specify the norm in the pernalization. We fixed the penality to its default value *l*2.
- **solver**: it enables to specify the strategy used to solve the optimization underlying our model. For the solver we fix it to *liblinear*.
- **tol:** tolerance for stopping criteria which is set to 0.0001.
- **verbose**: for the liblinear solver set verbose to any positive number for verbosity.
- **warm_start**: when set to *True*, reuse the solution of the previous call to fit as initialization, otherwise, just erase the previous solution. Useless for liblinear solver.

As the logistic regression performs a supervised learning we have used 60% for the training set and 30% for the test set.

5.2 Performance measures

To evaluate the performance of our prediction approach over the different used datasets, we have computed the precision, recall (or sensitivity), F-measure, and specificity of the predicted classes. We have also drawn the graph of the Receiver operating characteristic of the logistic regression to study its shape. The sensitivity, specificity and ROC plot are often used in the medecine domain as performance measures for prediction tasks.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Precision. The precision p, or positive value rate, for a class is the number of true positives (i.e. the number of cases correctly labeled as belonging to the class M+) divided by the total number of cases labeled as belonging to the class M+ (i.e. the sum of true positives and false positives, which are cases incorrectly labeled as belonging to the class).

$$p = \frac{\sum_{i=1}^{|R|} Entity(i)}{R}. \tag{4}$$

Entity(i) is a binary function that returns true if the predicted class for the i-th case is correct (w.r.t test set) and false otherwise. R is the sum of the number of predicted true positive and false positive cases.

Recall. The recall r (also known as sensitivity) is defined as the number of true positives divided by the total number of cases that actually belong to the clss M+ (i.e. the sum of true positives and false negatives, which are cases which were not labeled as belonging to the class M+ but should have been).

$$r = \frac{\sum_{i=1}^{|R|} Entity(i)}{G} \tag{5}$$

G is the sum of the number of predicted true positive and false negative cases.

F-measure. The F-measure, denoted by F_1, is a metric that measures the accuracy of a test in statistical analysis of a binary classification. It is computed using both the precision p and the recall r of the test as the ratio of the number of correct positive results and the number of all positive results returned by the classifier.

$$F_1 = 2 \times \frac{p \times r}{p + r} \tag{6}$$

Specificity. The specificity, also known as the true negative rate, measures the proportion of actual negatives that are correctly identified as such (e.g., the percentage of people not suffering from Malaria who are correctly identified as not having the condition).

Receiver operating characteristic. A receiver operating characteristic, or ROC in short, is a graph that shows the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the *true positive rate* (i.e. sensitivity or recall in machine learning) against the false positive rate (1 - specifity) at various threshold settings.

5.3 Experiments and analysis of the results

For each given dataset, we have performed two kinds of tests with our prediction model: a test without including the outcome of the quick diagnostic test or another with the

result of the quick diagnostic test (or QDT for short) among the input features. We first describe below the obtained results for each dataset and then present a comparative analysis.

Experiments with DT1. Table 2 and Figure 3 respectively show the performance measures and the ROC curve of the results of the our classification approach tested on dataset DT1 that is our cleaned real-world dpatient with imputed missing values. The results on Table 2a) and Figure 3a) are obtained without considering the quick diagnostic test output in contrast of measures in Table 2b) and Figure 3b). One can easily see that the accuracy of our prediction model is sensibly the same when considering or not the QDT; this accuracy is quite good as proven by the precision which is greater than 90%.

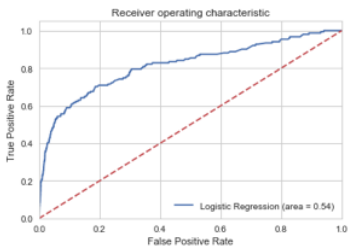
(a) Prediction without the QDT outcome

Precision	Recall	F-measure
0.97	1.0	0.99

(b) Prediction with the QDT outcome

Precision	Recall	F-measure
0.98	1.0	0.99

Table 2. Performance measures of the prediction on DT1



(a) Prediction without the QDT outcome

(b) Prediction with the QDT outcome

Fig. 3. The curve of the Receiver Operating Characteristic for prediction on DT1

Experiments with DT2. Table 3 and Figure 4 respectively show the performance measures and the ROC curve of the results (with or without taking into account the QDT) of the our classification approach tested on dataset DT2 that is our cleaned real-world patient with the deletion of records have missing data. The accuracy measures on Table 3a) et Figure 4) respectively compare to those in Table 3b) and Figure 4b) show that

our classifier does well when considering the QDT as a feature; without the QDT the precision of the prediction decreases a lot.

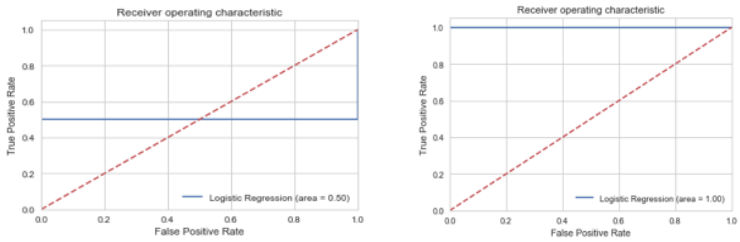
(a) Prediction without the QDT outcome

Precision	Recall	F-measure
0.75	1.0	0.86

(b) Prediction with the QDT outcome

Precision	Recall	F-measure
1.0	1.0	1.0

Table 3. Performance measures of the prediction on DT2



(a) Prediction without the QDT outcome (b) Prediction with the QDT outcome

Fig. 4. The curve of the Receiver Operating Characteristic for prediction on DT2

Experiments with DT3. similarly to results on DT2, performance measures on DT3 show a better prediction accuracy (See Table 4 and Figure 5) when the QDT is considered as a feature.

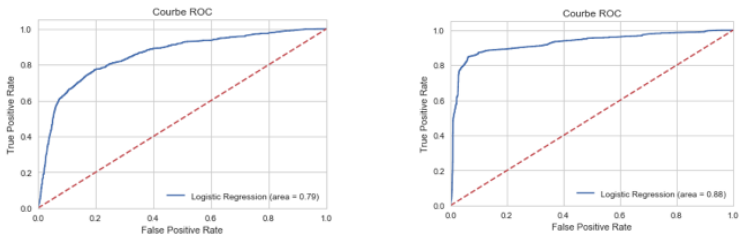
(a) Prediction without the QDT outcome

Precision	Recall	F-measure
0.77	0.82	0.79

(b) Prediction with the QDT outcome

Precision	Recall	F-measure
0.87	0.90	0.89

Table 4. Performance measures of the prediction on DT3



(a) Prediction without the QDT outcome (b) Prediction with the QDT outcome

Fig. 5. The curve of the Receiver Operating Characteristic for prediction on DT3

Comparative analysis of the results. In sum, the first experimentations detailed above prove that our logistic regression based prediction model does well in general and in particular when the outcome of the quick diagnostic test is considered as a feature for the learning process. More specifically, the precision of our prediction is greater than 90% for datasets DT1 and DT2, reaching 100% for DT1. For the specific case of the real-world patient dataset with missing data filled using an imputation algorithm this accuracy does not decrease even though the QDT is not considered during the prediction process, and only the Malaria features, i.e. signs and symptoms are taken into account. As a result, we can conlcude that there is a hope to construct an efficient prediction model for Malaria without the need to perform the quick diagnostic test in order to declare a given patent he is affected by the disease.

6 Conclusion

In this paper we have studied the problem of predicting the occurrence or not of Malaria given ill-patient dataset in the context of Senegal and by using machine learning techniques. To tackle this problem we have first presented a data preparation pipeline that enables to clean, normalize and impute missing values given a real-world dataset using efficient tools and algorithms. We also introduced a manner to extract the features that characterize the Malaria disease. We have then proposed a prediction model based on the logistic regression to determine the occurrence of Malaria. The performance of such a model has been demonstrated through extensive experimentations on real-world and semi-synthetic datasets. As a research perspective we plan to first include a prevalence factor into our prediction function in order to improve its accuracy. Second, we will use other binary classification models such as Support Vector Machine (or SVM in short) and compare their results to those obtained with the logistic regression based model.

References

1. Openrefine. http://openrefine.org/, online; accessed 30 Octobery 2018
2. Python implementation of missforest. https://pypi.org/project/predictive_imputer/, online; accessed 31 October 2018

3. Adimi, F., Soebiyanto, R.P., Safi, N., Kiang, R.: Towards malaria risk prediction in afghanistan using remote sensing. Malaria Journal **9**(1), 125 (May 2010)
4. Aminot I, D.M.: The use of logistic regression in the analysis of data concerning good medical practice. Rev Med Ass Maladie **33**(2), 157–143 (2002)
5. AS, A., AM, V., SH., K.: Malaria parasite development in the mosquito and infection of the mammalian host pp. 195–221 (2009)
6. Bbosa, F., Wesonga, R., Jehopio, P.: Clinical malaria diagnosis: rule-based classification statistical prototype. SpringerPlus **5**(1), 939 (Jun 2016)
7. Chiroma, H., Abdul-kareem, S., Ibrahim, U., Ahmad, I.G., Garba, A., Abubakar, A., Hamza, M.F., Herawan, T.: Malaria severity classification through jordan-elman neural network based on features extracted from thick blood smear. In: Neural Network World (2015)
8. Dua, S., Acharya, U.R., Dua, P.: Machine Learning in Healthcare Informatics. Springer Publishing Company, Incorporated (2013)
9. Daz, G., Gonzlez, F.A., Romero, E.: A semi-automatic method for quantification and classification of erythrocytes infected with malaria parasites in microscopic images. Journal of Biomedical Informatics **42**(2), 296 – 307 (2009)
10. Ferguson, H.M., Mackinnon, M.J., Chan, B.H., Read, A.F.: Mosquito mortality and the evolution of malaria virulence. Evolution **57**(12), 2792–2804 (2003)
11. Hosmer, D.W., Lemeshow, S.: Applied logistic regression. John Wiley and Sons (2000)
12. Kunwar, S.: Malaria Detection Using Image Processing and Machine Learning. ArXiv e-prints (Jan 2018)
13. Kusumasari, T.F., Fitria: Data profiling for data quality improvement with openrefine. In: 2016 International Conference on Information Technology Systems and Innovation (ICITSI). pp. 1–6 (Oct 2016)
14. Lemaître, G., Nogueira, F., Aridas, C.K.: Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research **18**(17), 1–5 (2017)
15. Preux, P., Odermatt, P., Perna, A., Marin, B., Vergnengre, A.: Qu'est-ce qu'une regression logistique ? Revue des Maladies Respiratoires **22**(1, Part 1), 159 – 162 (2005)
16. Rajpurkar, P., Polamreddi, V., Balakrishnan, A.: Malaria likelihood prediction by effectively surveying households using deep reinforcement learning. CoRR **abs/1711.09223** (2017)
17. Robert, C.: Machine learning, a probabilistic perspective. CHANCE **27**(2), 62–63 (2014)
18. Silva, L.O., Zárate, L.E.: A brief review of the main approaches for treatment of missing data. Intell. Data Anal. **18**(6), 1177–1198 (Nov 2014)
19. Sokhna, C., Mboup, B.M., Sow, P.G., Camara, G., Dieng, M., Sylla, M., Gueye, L., Sow, D., Diallo, A., Parola, P., Raoult, D., Gautret, P.: Communicable and non-communicable disease risks at the Grand Magal of Touba: The largest mass gathering in Senegal. Travel Medicine and Infectious Disease **19**, 56–60 (Sep 2017)
20. Sperandei, S.: Understanding logistic regression analysis. Biochem Med **24**, 12–18 (feb 2014)
21. Stekhoven, D.J., Bhlmann, P.: Missforest: non-parametric missing value imputation for mixed-type data. Bioinformatics **28**(1), 112–118 (2012)
22. Swalin, A.: How to handle missing data. https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4, online; accessed 31 October 2018
23. Ugwu, C., Onyejegbu, N.L., Obagbuwa, I.C.: The application of machine learning technique for malaria diagnosis. Int. J. Green Comput. **1**(1), 68–77 (Jan 2010)
24. Wang, J., Xu, M., Wang, H., Zhang, J.: Classification of imbalanced data by using the smote algorithm and locally linear embedding. In: 2006 8th international Conference on Signal Processing. vol. 3 (Nov 2006)
25. WHO: World malaria report in 2017 (2017)