Patient Signs and Symptoms based Malaria Prediction Using Machine Learning Algorithms: an experimental study

Ousseynou Mbaye, Mouhamadou Lamine Ba, and Alassane Sy

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

Abstract. Still today, Malaria remains one of the most feared diseases in Sub-Saharan Africa and especially in Senegal. This is mainly due to inappropriate medical care support coupled with an often late and error-prone diagnosis from the medical staff. In addition, largely used diagnostic standards such as the Rapid Diagnosis Test is not fully reliable. With the development and increasing adoption of automated tools in the health field, machine learning applications might help medical actors in their decision-making process. In this paper, we propose an experimental study of six machine learning algorithms for the prediction of Malaria in Senegal. These algorithms aim at predicting whether or not a given patient suffers from Malaria based on his signs and symptoms. The performance of the algorithms have been extensively tested and evaluated over two real data sets about patients in Senegal that suffer or not from Malaria. The results obtained show that Random Forest, Support Vector Machine with Gaussian Kernel and Artificial Neural Networks 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 on which they outperform the Rapid Diagnostic Test.

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 4 we show that the precision of the RDT 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 3 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 3. In Section 4, 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 2 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 5. Finally, we conclude this paper in Section 6.

2 Review of evaluated ML algorithms

We mainly provide a brief overview of Machine Learning and Deep Learning models for healthcare applications.

In healthcare, Machine Learning apps can help better understand each patient's care journey, medical decisions, or the impact of new drugs. Today researchers are using machine learning algorithms in the diagnosis of several diseases such as diabetes, stroke, cancer, malaria and heart disease. In the following we discuss about some of these methods. Those algorithms are chosen among the most used ones in the health field according to studies[**?**,**?**].

Decision tree (DT)[**?**] is a supervised classifier which is obtained by recursively partitioning the labelled set of observations. It is one of the most adopted classifiers, thanks to its simplicity and its straightforward interpretation. For CART algorithms, hyperparameters are the impurity criteria (entropy and gini), the maximum depth, the minimum samples to split and the minimum samples at a leaf Decision tree algorithm has been applied in many medical tasks, for examples, in increasing quality of dermatologic diagnosis[11], predicting essential hypertension [12], and predicting cardiovascular disease [13], predict and diagnose of heart disease [14].Decision tree is one of the most popular tools for classification and prediction.

Random Forest (RF)[**?**] is an ensemble approach built upon many decision tree classifiers. It is a supervised classifier which requires the same hyper parameters as DT, plus the number of trees to create and the random number of features to look at when splitting the labelled data during the training step [**?**].

Naive Bayes classifier (NB) [**?**] is a*supervised* machine learning algorithm, i.e. requires to be trained, used for classifying observations to given distinct classes based on *input explanatory variables* (a.k.a feature or attribute). It is a classification technique based on the well-known *Bayes' theorem*[1] with strong and naive assumptions. It simplifies learning by assuming that features are independent of given class. The Bayesian classifier has been applied in many medical issues, for examples, in measuring quality of care in psychiatric emergencies [21], predicting and diagnosis heart disease [14]. and assisting diagnosis of breast cancer [22].

[1] https://en.wikipedia.org/wiki/Bayes%27_theorem

Logistic regression (LR) [**?**] is a statistical model used in the machine learning domain as a supervised classifier for binary classification [**?**]. It is based, in its basic form, on a logistic function to describe a binary dependent variable[**?**,**?**] by considering as input qualitative or/and ordinal explanatory variables in order to measure the probability of a given class label. The greatest advantage of the logistic regression classifier is the fact that you can use continuous explanatory variables and it is easier to handle more than two explanatory variables simultaneously and 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. One the other hand, the logistic regression is one of the most used multi-valued models in epidemiology[23]. 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 logistic regression is applied to predict malaria in [26] and in identification of at-risk populations in public health research and outreach [27] and the results are are very relevant.
Support Vector Machine (SVM) [**?**] is a supervised classification approach whose intuition is to represent input data in a space and to determine the optimal hyper-plane that divides that space in two regions depending on the targeted value. SVM is used to studied the diagnosis of coronary artery disease[28].
An Artificial Neural Network (ANN) [**?**] is a computational approach also referred to as a Connectionist System used in Machine Learning. ANNs are loosely modeled after the biological neural network in an attempt to replicate the way in which we learn as humans. Think of it as a computing system, structured as a series of layers, each layer consisting of one or several neurons. The types of the layers comprise *input*, *output* and *hidden* layers [**?**,**?**].

3 relate dwork

We mainly provide a brief overview of Machine Learning and Deep Learning models for healthcare applications.

In healthcare, Machine Learning apps can help better understand each patient's care journey, medical decisions, or the impact of new drugs. The survey of [**?**] explores the usefulness of various data mining techniques such as classification, grouping, association, regression in the health field. This survey also highlights the applications, challenges and future issues of data mining in healthcare. The recommendation regarding the appropriate choice of available data mining technique is also discussed in this article. The authors of [2] explain the fundamental principles of logistic regression and the stages of its application. Using two examples (the quality of follow-up care for diabetics and hospital mortality after acute myocardial infarction), they demonstrate the value that this statistical tool can have in studies carried out by the medical service of the national health fund, especially in studies aimed at evaluating professional practice. Another example of previous work that has used logistic regression is that of Farida et al. [**?**]. 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 [18] Uddin et al, provide a broad overview of the relative performance of different variants of supervised machine learning algorithms for disease prediction. Thus, their results showed that the Support Vector Machine (SVM) algorithm is applied most frequently (in 29 studies) followed by the Naïve Bayes algorithm (in 23 studies). However, the Random Forest (RF) algorithm has shown comparatively higher accuracy. Of the 17 studies where it has been applied, RF has shown the highest accuracy in 9 of them, or 53%. This was followed by SVM which exceeded 41% of the studies it considered.[**?**] Presents a reference of 7 machine learning algorithms used on binary classification tasks and applied to hospital data. In the study [**?**] Data mining acts as a solution to many health problems and it is useful for predicting early stage heart disease. The Naive Bayes algorithm is one such data mining technique that helps predict heart disease in patients. Gharehchopogh et al. [**?**] explain the use of medical data mining in determining methods of medical operation. They show that the decision tree algorithm designed for this case study generates a correct prediction for more than 86% of test cases. Indeed, decision trees based approach has been proposed in Nigeria [**?**] to predict the occurrence of Malaria given diagnostic data. In the same line of works applying machine learning, in [**?**], 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.

3.1 Description des données

4 REAL PATIENT HEALTH DATASETS

In order to carry out our experiments in a real setting we have collected two real world datasets about patients living in Senegal. We describe each of them in the sequel.

Data collection. Our first dataset, that we to refer to it as DT1, contains medical records about patients living in distinct places in Senegal. It has been collected in 2016 during the "Grand Magal of Touba" which is one of the most popular religious event in Senegal. Such an event gathers every year several millions of persons that come from various areas around the country [**?**]. During the event several fixed and mobile health points are set up to enable the examination and treatment of ill persons. The second dataset, denoted by DT2, has been collected by drawing our attention on medical records about patients living in the same area. We focused on the district of Diourbel,Thies and Fatick [2] where the prevalence of Malaria is very high and collected patient records from its different health structures.

Data features. Table 1 contains the main characteristic of each dataset in terms of number of recorded variables (mainly clinical features), number of observations, vari-

[2] https://en.wikipedia.org/wiki/Diourbel_Region

able types, number of observations per class (Malaria or Not Malaria), and the precision of the Rapid Diagnosis Test. In details, values for seventeen variables have been extracted for each observation in both datasets; two variables are basically of numerical types while the remaining are Boolean. Some of these variables (also called features or attributes) include personal data about the patient, but also signs and symptoms of the patient reported by the doctor who treated this later. The other attributes describe clinical data such as information about the doctor's final diagnosis (the patient's disease), the outcome of the Rapid Diagnosis Test and the patient's status (i.e. admission, death or observation). For privacy reasons and certain restrictions in the use of the data, we have ignored patient personal data during this study. In addition, we can observe that the first dataset is larger than the second one (21083 observations versus 5809 observations). Moreover, both datasets are unbalanced because the proportion of observations per class is largely unequal. As an example for dataset DT1 we have 614 observations in the first class and 5108 observation in the second class. Finally, we remarked that the precision of the Rapid Diagnosis Test is around 90% for both datasets, meaning that the systematically performed RDT in Senegal is not fully reliable.

On the other hand, Figure 1 shows that the raw datasets come with missing values for some variables on given observations. To resolve the problem of unbalanced datasets and data messness, we followed a data preparation pipeline in order to fit our datasets into the good format for our experimentation; we discuss about such a data preparation step next.

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. Raw Data characteristics

Data preparation. We have followed the same process as in [**?**] in order to cleanse, normalize, impute, and balance information in our real datasets. Firstly, the raw datasets come with lot of inconsistencies due to the way the information were originally collected within the health structures. Indeed information about patients are manually recorded in the majority of health structures in Senegal. Second, when we did an explanatory analysis of the set of real-world data, they have revealed that the datasets were not balanced and come with missing values as mentioned above. We then used *OpenRefine*[3] to first clean and normalize information in our datasets. After that, we resolved our problem of missing values and unbalanced datasets by respectively using a K-Nearest Neighbours based imputation algorithm and an oversampling of the minority class: for more details we defer the reader to [**?**]. Figure 2 summarizes the new characteristics of DT1 and DT2 after the data preparation step.

[3] https://openrefine.org/

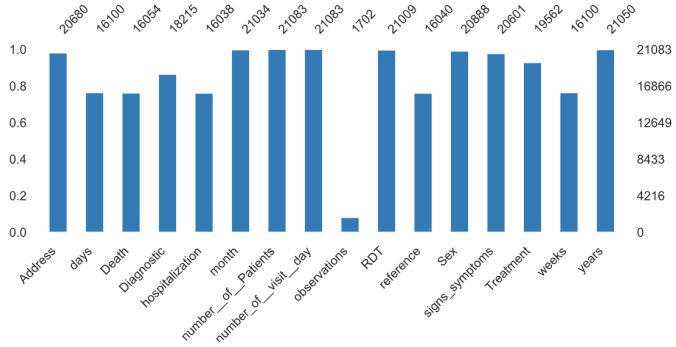

Fig. 1. Proportion of missing values per variable

Dataset	Variables	Observations	Variables types		Classes	
			Numeric	Boolean	Malaria	not Malaria
DT1	16	61396	2	14	30698	30698
DT2	16	14336	2	14	7168	7168

Table 2. Data characteristics after preparation step

From DT1 and DT2 we built three news datasets DT3, DT4 and DT5 data sets as below.
DT3: It is obtained by concatenating the DT1 and DT2 datasets. Thus it concerns 37,175 patients of which 9,837 are diagnosed positive for malaria.
DT4: It is obtained by considering the 16,092 patients in the DT2 data set (including 9,223 patients with malaria). Since this DT2 is unbalanced, we randomly selected 2354 patients who tested negative for malaria from the DT1 data set at the end of the rebalance. Thus it concerns 18,446 patients, 9,223 of whom are suffering from malaria.
DT5: is obtained by the over sampling of DT1 by the SMOTE method of python. This method consists first of dividing DT1 into two parts, one for training (train set) and the other for testing (test set). The train set being unbalanced, then we apply the SMOTE method to remedy it. Thus we obtain a new train set comprising 30,369 patients, half of whom tested positive for malaria.

5 Experimentation and results

We detail and analyze in the section the results of the experimentation we performed using the six ML algorithms presented in Section 2 over the two real datasets described in Section 4. We start by presenting our experimentation setting.

5.1 Experimentation Setting

All the performed tests have been done in the same machine and the same operating system. To test the performance of our six chosen ML algorithms, we relied on their

Python implementations available through the scikit-learn library[4]. Scikit-learn is an open source simple and efficient tool for predictive data analysis that implements most of the existing ML algorithms. We set the following values for the various parameters of each algorithm.

- NB
  - priors=None, var smoothing=1e-09.
- LR
  - C=1.0, class weight=None, dual=False, multi class='warn', fit intercept=True, intercept scaling=1, l1 ratio=None, max iter=1000, penalty='l2', random state=0, solver='lbfgs', verbose=0, warm start=False, n jobs=None, tol=0.0001.
- DT
  - class weight=None, criterion='gini', max depth=None, max features=None, max leaf nodes=None, min samples split=2, min weight fraction leaf=0.0, presort=False, min impurity decrease=0.0, min impurity split=None, min samples leaf=1.
- RF
  - class weight=None1 criterion='gini', max depth=None, max features=None, max leaf nodes=None, min impurity decrease=0.0, min impurity split=None min samples leaf=1, in samples split=2, min weight fraction leaf=0.0, presort=False, random state=0, splitter='best'.
- SVM
  - C=1.0, cache size=200, class weight=None, coef0=0.0, decision function shape='ovr', degree=3, gamma='auto', kernel='rbf', max iter=-1, probability=True, random state=None, shrinking=True, tol=0.001, verbose=False.
- ANN
  - activation='relu' alpha=0.0001, batch size='auto', beta1=0.9, beta2=0.999, early stopping=False, epsilon=1e-08, hidden layer sizes=(15, 15, 15), learning rate='constant', learning rate init=0.001, max iter=10000, momentum=0.9, numb iter no change=10, nesterovs momentum=True, power t=0.5, random state=None, shuffle=True, solver='adam', tol=0.0001, validation fraction=0.1, verbose=False, warm start=False.

For the details about the description of each parameter we refer to the official documentation of the implementation of these algorithms in scikit-learn[5]. Concerning the segmentation of both datasets for the training of our ML algorithms and their testing we have considered the following splitting of the initial data.

- DT1
  - Training set : 70%
  - Test set : 30%
- DT2
  - Training set : 80%
  - Test set : 20%

[4] https://scikit-learn.org/stable/

[5] https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

5.2 Performance measures

To evaluate the performance of every classifier tested during this study, we leverage several different standard measures. We present next the definition of these performance measures.

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 positive class) divided by the total number of cases labeled as belonging to the positive class (i.e. the sum of true positives and false positives, which are cases incorrectly labeled as belonging to the class).

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 positive class (i.e. the sum of true positives and false negatives, which are cases which were not labeled as belonging to the positive class but should have been).

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.

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 - specificity) at various threshold settings. The Area Under the Curve (AUC) of the ROC curve is a discrimination measure which tells us how well our predictor can classify patients in two groups: those with and those without the outcome of interest.

5.3 Results of the experiments

This section presents the results of the experimentation on each real dataset for each of the six classifiers.

Experiments with DT1. Table 3 and Figure 2 respectively show the performance measures (Precision, Recall and F-measure) and the ROC curves of the results of our six classifiers after experimentation on the dataset DT1. Observation shows that all classifiers have the same precision 99% but present different recall and F-measure (see Table 3). At the same time, we note that the surfaces under the ROC curves, i.e. the AUC (Area Under the Surface) values, of the different algorithms are clearly different with values between 0.50 and 0.87.

ML algorithm	Precision	Recall	F1-score
NB	0.99	0.17	0.29
LR	0.99	0.92	0.96
DT	0.99	0.17	0.29
RF	0.99	0.98	0.99
SVM(kernel=gaussian)	0.99	0.98	0.99
SVM(kernel=polynom)	0.99	0.92	0.95
ANN(MLP)	0.99	0.99	0.99

Table 3. Precision, Recall and F1-score measures over DT1

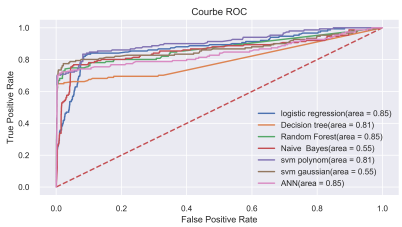

Fig. 2. True Positive Rate over False Positive Rate on DT1

Experiments with DT2. Table 4 and Figure 3 respectively show the performance measures (Precision, Recall and F-measure) and the ROC curves of our six classifiers after experimentation on the dataset DT2. In contrast to the results obtained with DT1, we notice that our classifiers have overall precision which are slightly down and vary between 93% and 96% (see Table 4). Likewise ROC curves follow the same trends with AUC values between 0.50 and 0.70

ML algorithm	Precision	Recall	F1-score
NB	0.96	0.05	0.10
LR	0.93	0.62	0.75
DT	0.92	0.85	0.88
RF	0.92	0.85	0.89
SVM(kernel=gaussian)	0.92	0.86	0.89
SVM(kernel=polynom)	0.93	0.54	0.68
ANN(MLP)	0.93	0.85	0.89

Table 4. Precision, Recall and F1-score measures on DT2

5.4 Analysis of the results and discussion

Analyzing in details the performance of our six classifiers on both datasets, the results of the previous section clearly argue in favor of the classifiers RF, LR, SVM with Gaus-

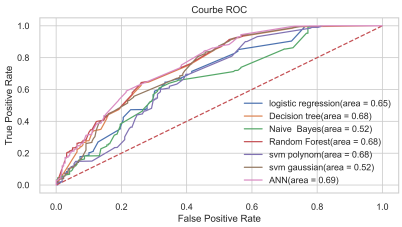

Fig. 3. True Positive Rate over False Positive Rate on DT2

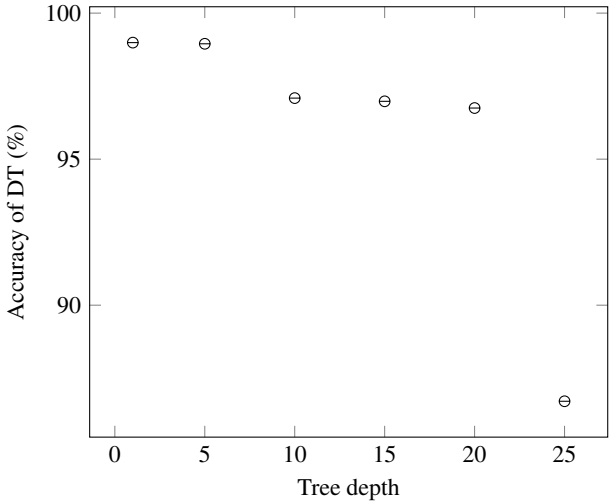

Fig. 4. Accuracy of DT with respect to the tree depth

sian kernel and ANN. Indeed considering the datase DT1, that contains observations about patients living in different areas in Senegal, these four classifiers have a precision of 99%, a recall above 92% and a F-measure above 95%. We note the same trend with the dataset DT2 which contains observations about patients living in the same area in Senegal. So in terms of precision, recall and F-measure those four classifiers outperform the rest. We can also remark that RF, LR, SVM with Gaussian kernel, and ANN present better precision than the systematically performed and used Rapid Diagnostic Test within the majority of health structures in Senegal. The difference of performance of our classifiers on the two datasets can be explained by the fact that climatic factors such as temperature and standing water are very determining in the appearance or not of Malaria in distinct areas in Senegal as they favor the development of mosquitoes responsible for the disease.

When we now restrict ourselves to the ROC curves, we observe that SVM with Gaussian kernel and Naive Bayes present the worst positive prediction rate as they have the lowest AUC values compared to the other classifiers.

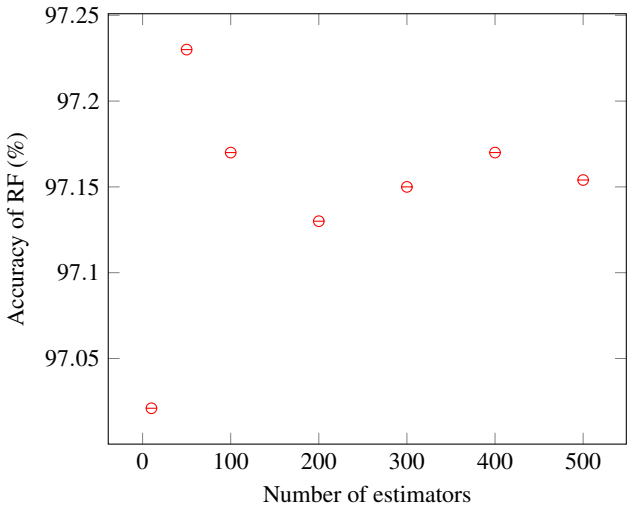

Fig. 5. Accuracy of RF with respect to the number of used decision trees

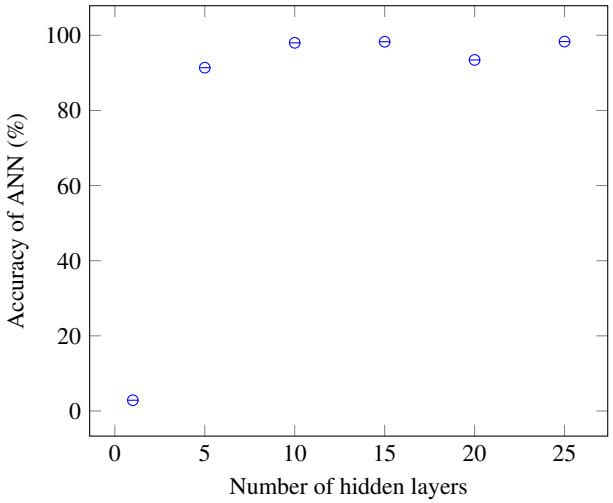

Fig. 6. Accuracy of ANN with respect to the number of hidden layers

In addition we have tried to study the impact of the tree depth, the number of hidden layers and the number of estimators for DT, ANN and RF respectively. While the increase of the tree depth decreases the accuracy of DT (see Figure 4), the highest accuracy of RF corresponds to the use of 50 estimators as shown in Figure 5. Furthermore, Figure 6 shows that when the number of hidden layers increases the accuracy of ANN does so in general. To conclude, we can argue that ANN seems to be the most promising classification approach among the six studied ML algorithms when we are only

interested by the precision, recall and F-measure. If we include the ROC curves in the analysis ANN remains the most efficient approach.

6 Conclusion