Machine Learning based Malaria Prediction using Clinical Findings

Conference Paper · April 2021

DOI: 10.1109/ESCI50559.2021.9396850

5 authors, including:

Samir Yadav
Dr. Babasaheb Ambedkar Technological University
17 PUBLICATIONS **612** CITATIONS
SEE PROFILE

Vinod Kadam
Dr. Babasaheb Ambedkar Technological University
27 PUBLICATIONS **287** CITATIONS
SEE PROFILE

Shivajirao Jadhav
Dr. Babasaheb Ambedkar Technological University
15 PUBLICATIONS **392** CITATIONS
SEE PROFILE

2021 International Conference on Emerging Smart Computing and Informatics (ESCI)
AISSMS Institute of Information Technology, Pune, India. Mar 5-7, 2021

Machine Learning based Malaria Prediction using Clinical Findings

1st Samir S. Yadav
Department of Information Technology
Dr. Babasaheb Ambedkar Technological University, Lonere
Raigad, India
ssyadav@dbatu.ac.in

2nd Vinod J Kadam
Department of Information Technology
Dr. Babasaheb Ambedkar Technological University, Lonere
Raigad, India
vjkadam@dbatu.ac.in

3rd Shivajirao M. Jadhav
Department of Information Technology
Dr. Babasaheb Ambedkar Technological University, Lonere
Raigad, India
smjadhav@dbatu.ac.in

4th Sagar Jagtap
Electronics and Telecommunication Engineering
Dr. Babasaheb Ambedkar Technological University, Lonere
, India
sagar.jagtap@aissmsioit.org

5th Prasad R. Pathak
Department of Information Technology
Dr. Babasaheb Ambedkar Technological University, Lonere
Raigad, India
prasadpathak4444@gmail.com

***Abstract*—Even today, Malaria is the most deadly disease in Asia and sub-Saharan Africa and particularly in Senegal. This is mainly due to inadequate medical care support with frequent late and error-diagnoses by medical professionals. Besides, mostly used diagnostic standards such as the rapid diagnostic test is not fully reliable. With the development and widespread acceptance of automated systems in the healthcare system, machine learning algorithms can support medical professionals in their decision-making procedure. An experimental analysis of different machine learning techniques to predict Malaria is proposed in this work. These techniques attempt to determine whether or not a patient suffersfrom Malaria using various clinical findings like signs and symptoms. The algorithms' efficiency has been thoroughly validated and analysed over two actual data sets of malaria patients' taken from Senegal. 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.**

***Index Terms*—Malaria, Machine learning, ANN, SVM, Random Forest**

I. Introduction

Malaria, also known as *"fivre des marais"* in French, is an infectious disease caused by a mosquito of the type *Plasmodium*. The illness can lead to *yellow skin*, *seizures*, *coma* or *death* in its extreme nature. Therefore, the World Health Organisation (WHO) is now identifying and addressing malaria as a major health issue globally, with a pandemic in Sub-Saharan Africa in particular. Approximately 228 million malaria cases were reported globally in 2018, and most of the malaria patients were in the WHO African Region(about 93%), according to the Global Malaria Survey for 2019 [1]. In Senegal's particular case, the problem is acute because of inappropriate care support means and 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 becomes crucial to avoid its evolution towards a critical state.

Over the past years, many efforts have been made 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 [2]–[5]. The Rapid Diagnostic Test (RDT) [4] is one of the most successful and prominent introduced tools to automatically predict whether a given patient suffers from Malaria. It relies on the detection of specific Plasmodium proteins, PfHRP2, pLDH and aldolase. The RDT is mostly 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, the RDT is not entirely accurate as illustrated in [4]: we demonstrate in the III section that the accuracy of the RDT for the real datasets used in this work is around 90%

The Liverpool Model on Malaria (LMM) extended in [3] is an example of a mathematical model that tries to model the parasite life cycle. It promotes malaria propagation at regular analysis using average daily temperatures and cumulative precipitation for ten days. The ultimate objective is to create a climate-driven model of Malaria that will provide a deeper understanding of the complexities of malaria transmission. LMM is not a diagnostic system. We defer the reader to Section II for an exhaustive review of the literature. Despite existing works, Malaria prediction accuracy is still a concern: used mechanisms, e.g., domain knowledge and RDT, in Senegal, are error-prone. Machine Learning (ML) [6], [7] technologies can assist healthcare professionals in their decision-making phase with the advancement and the implementation of automated software in the healthcare system. There are already some attempts to apply ML techniques to predict or better understand various diseases, e.g., [8], [9]. For example, logistic regression has been tested in [8] for the prediction of Malaria and provides promising results. This paper proposes a comprehensive comparative analysis of six machine learning algorithms, one of the most popular for malaria prediction in Senegal. The evaluated and compared ML algorithms are **Naive Bayes** (NB) [10], **Logistic Regression** (LR) [11], **Decision Tree** (DT) [12], **Support Vector Machine** (SVM) [13], **Random Forest** (RF) [14], and **Artificial Neural Network** (ANN) [15]. 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. They contain clinical data such as a sign, symptom, and final diagnostic of patients living in distinct Senegal locations (for the first dataset) or within the same area (for the second dataset). Doctors have examined those patients 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 is *Precision*, *Recall*, *F1-score*, *True Positive Rate*, and *False Positive Rate* on both datasets augmented with semi-synthetic datasets obtained after imputation 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 least equal to 92%, 85% and 89% respectively on both datasets. More specifically, those three learning approaches outperform the RDT, representing the baseline automatic diagnostic tool primarily adopted as a standard within the Health system in Senegal.

The remainder of the article is organized as follows. Initially, the literature study on existing investigation works dealing with Malaria is reviewed in section II. In section III, a detailed description of two real-world datasets contains medical records about Senegal patients are given. More precisely, each dataset's characteristics, imputation to deal with missing values and the precision values of RDT are presented. Experimental design, result and discussions are given in section IV. At last, conclusions are drawn in section V.

II. Related Work

A brief overview of ML and Modern ML called deep learning models for healthcare applications is mainly provided. Machine Learning applications in healthcare can better understand each patient's treatment experience, medication options, or new drugs' effects. Researchers are currently using ML methods to detect various diseases such as diabetes, stroke, cancer, malaria, and heart disease. Some of these methods are discussed as follows. These algorithms are chosen among the most used ones in the health field, according to studies [16]–[25].

Decision tree (DT) [12] is a supervised classifier which is obtained by recursively partitioning the labelled set of observations. It is among the most accepted classifiers, due to its simple design and easy explanation. Hyperparameters for CART algorithms are the impurity requirements (entropy and Gini), the large number, the required tests to be separated, and the minimum leaf samples. The decision tree algorithm has been extended to several clinical cares, such as improving dermatological treatment efficiency, estimating critical hypertension, detecting and classifying cardiac arrhythmias [20]–[22], [25]. **Random Forest (RF)** [14] is an ensemble approach built upon many decision tree classifiers. This is Supervised learning classifier consists of same hyperparameters that of DT. [14].

Naive Bayes classifier (NB) [10] 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). This classification method is based on the well-known *Bayes' theorem* using strong and naive assumptions. It simplifies learning by assuming that features are independent of a given class. This classifier is also used in many medical applications such as heart disease diagnosis, psychiatric emergencies etc [3], [10].

Logistic regression (LR) [11] is supervised learning classifer used for binary classificaions [26]. The raw version of this variable is based on a logistic representation of a binary dependent variable by considering contextual or/and ordinary explanatory variables as feedback to calculate the probability of a particular class label [16], [27]. The logistic regression is applied to predict malaria and identify at-risk populations in public health research and outreach, and the results are very relevant citeashton2020risk.

Support Vector Machine (SVM) [13] Classification is a supervised technique, the intuition of which is to describe the data input with a planeas well as to decide the optimum decision boundary dividing the aircraftby the specific value into two regions. SVM is used to study the diagnosis of coronary artery disease [28], [29].

An Artificial Neural Network (ANN) [15] is often known as the connectionist method included in computer vision. ANNs are roughly based on the biological neural network to simulate how we develop as humans. Assume it as a computer device, organised as a sequence of layers, each layer composed of one or more neurons. The types of the layers include *input*, *output* and *hidden* layers [30], [31].

III. Datasets

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.: The first dataset that we refer to as DT1 contains medical records about patients living in distinct places in Senegal. It was collected in 2016 during the "Grand Magal of Touba", one of the most famous religious events in Senegal. Such an event gathers several millions of persons from various areas around the country [32]. 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 [1] where the prevalence of Malaria is very high and collected patient records from its different health structures.

Data features. : Table I contains:

- Each dataset's main characteristic in terms of several recorded variables (mainly clinical features).
- Several observations.
- Variable types.
- Several observations per class (Malaria or Not Malaria).
- 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. These factors (also known as features or attributes or clinical findings) provide the patient's records and the signs and symptoms documented by the physician who assessed the patient afterwards. Other features characterise health details like details about the doctor's clinical outcome (patient disease), the results of the Rapid Diagnosis Test, and the condition of the patient (i.e. admission, death or observation). During this analysis, we overlooked patient personal information for purposes of confidentiality and some limitations on data use. We can also 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 not equal. 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 unbalanced datasets and data messes, we followed a data preparation pipeline to fit our datasets into the excellent format for our experimentation; we discuss such a data preparation step next.

Data preparation.: We have followed the same process as in [8] to cleanse, normalize, impute, and balance information in our real datasets. Firstly, the raw datasets come with many inconsistencies because the data collected initially within the health structures. Indeed information about patients is manually recorded in the majority of health structures in Senegal. Second, when we did a descriptive analysis of real-world data sets, it is found that the datasets were imbalanced and came with missing values as mentioned above. We then used *OpenRefine*[2] 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 [8]. Figure II summarizes the new characteristics of DT1 and DT2 after the data 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 oversampling of DT1 by the SMOTE method of python. This method consists of dividing DT1 into two parts: 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.

IV. Methods

In the section, we detail and analyze the results of the experimentation we performed using the six ML algorithms presented in Section **??** over the two real datasets described in Section III.

A. Experimental Design

All the performed tests have been done in the same machine and the same operating system. To test our six chosen ML algorithms' performance, we relied on their Python implementations available through the scikit-learn library[3]. Scikit-learn is an open-source efficient and straightforward 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.

[1] https://en.Wikipedia.org/wiki/Diourbel_Region

[2] https://openrefine.org/

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

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 I
RAW DATA CHARACTERISTICS

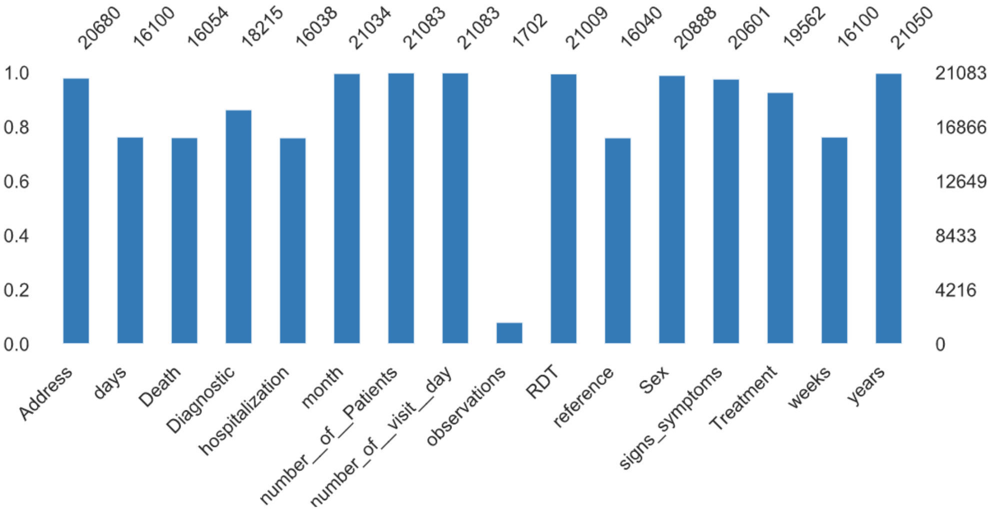

Fig. 1. Proportion of missing values per variable

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 II
DATA CHARACTERISTICS AFTER PREPARATION STEP

- 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[4]. 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 in table III.

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

Dataset	Training	Testing
DT1	70%	30%
DT2	80%	20%

TABLE III
SPLITTING OF DATASETS

B. Performance measures

We have used five different performance measures to evaluate the performance of six different ML classifiers such as precision, recall, F-measure, Specificity and Receiver operating characteristic also called ROC. The value of these measures calculated by using confusion matrix parameters. This matrix consists of four different values: True Positive(TP)value means correctly predicted positive values, False Positive(FP) means wrongly predicted positive value, True Negative(TN)means correctly predicted negative values, and finally, False Negative(FN)value represents the incorrectly predicted negative value. By using these parameters, we can define a given performance as follows:

$$Precision = TP/TP + FP \tag{1}$$

$$Recall = TP/TP + FN \tag{2}$$

$$Specificity = TN/TN + FP \tag{3}$$

$$F1Score = 2 * (Recall * Precision)/(Recall + Precision) \tag{4}$$

ROC curve is a graph used to evaluate the diagnostic ability of the ML classifier. It is created by using a plot of recall against false positive rate(1-specificity) at different threshold values. The Area under curve(AUC) is another measure useful in performance measure. AUC of ROC 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.

C. Results and Discussion

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

Experimental design for DT1 Dataset.: Table IV and Figure 2 show the Precision, Recall, F-measure and the ROC curves of the results of our six classifiers after experimentation on the dataset DT1 respectively. Observation shows that all classifiers have the same precision 99% but present different recall and F-measure (see Table IV). 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 different with values between 0.50 and 0.87.

Experimental design for DT2 Dataset.: Table V 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 V). Likewise ROC curves follow the same trends with AUC values between 0.50 and 0.70

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 IV
PRECISION, RECALL AND F1-SCORE MEASURES OVER DT1

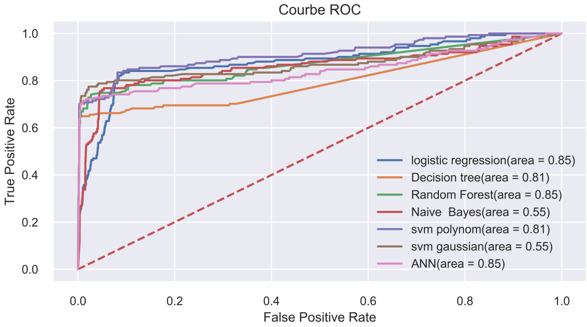

Fig. 2. True Positive Rate over False Positive Rate on DT1

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 V
PRECISION, RECALL AND F1-SCORE MEASURES ON DT2

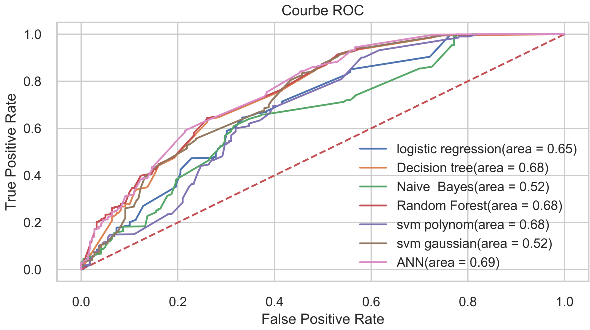

Fig. 3. True Positive Rate over False Positive Rate on DT2

D. 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 Gaussian 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

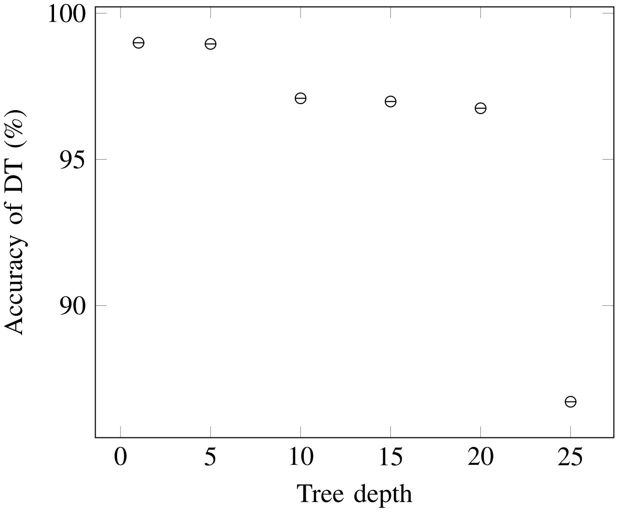

Fig. 4. Accuracy of DT with respect to the tree depth

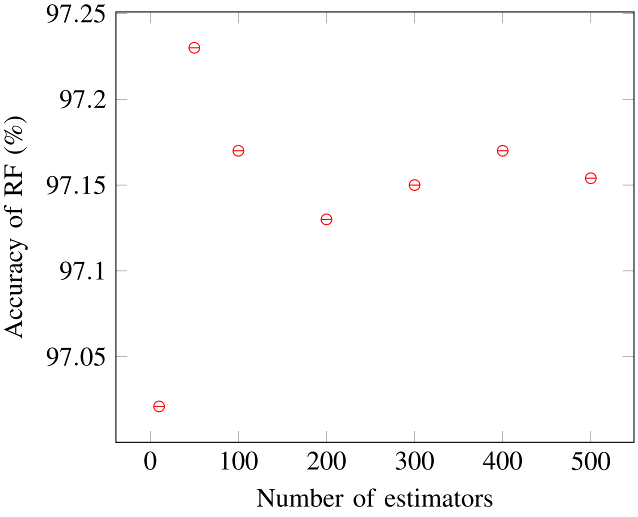

Fig. 5. Accuracy of RF with respect to the number of used decision trees

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.

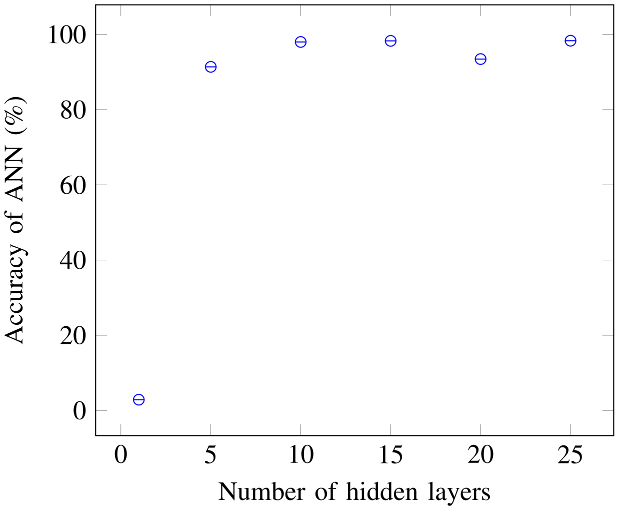

Fig. 6. Accuracy of ANN with respect to the number of hidden layers

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.

V. Conclusion

This paper discussed the comparison of six different ML classification techniques for detecting malaria disease using various clinical findings features. Malaria is a life-threatening disease for peoples in Asia and Subsaharan Africa, particularly in Senegals. Therefore in this paper, two actual data sets of malaria patients' taken from Senegal are used to evaluate the performance of ML techniques. The results showed that RF, Gaussian Kernel SVM and ANN are useful and give the highest overall accuracy to predict Malaria disease with precision, recall and F1-score at least equal to 92%, 85% and 89% on both datasets on which the Rapid Diagnostic Test is conducted. And, hence, these techniques can be helpful for medical professionals and healthcare researchers. The data used in this research is small, and from only one region, this may affect the performance of these ML classifiers. Hence, it will be interesting to use extensive data and data from different areas of future work.

References

[1] W. H. Organization, "2019 world malaria report," December 2019, https://www.who.int/malaria/publications/world-malaria-report-2019/en/.

[2] T. Lepes, "Review of research on malaria*," *Bulletin of the World Health Organization*, vol. 50, no. 3-4, pp. 151 – 157, 1974.

[3] V. Ermert, A. Fink, A. Jones, and A. Morse, "Development of a new version of the liverpool malaria model," *Malaria journal*, vol. 10, p. 35, 02 2011.

[4] S. Houz, "Rapid diagnostic test for malaria," *Bull. Soc. Pathol. Exot.*

[5] J. Garrido-Crdenas, J. Cebrian-Carmona, L. Gonzalez-Ceron, F. Manzano-Agugliaro, and C. Mesa-Valle, "Analysis of global research on malaria and plasmodium vivax," *International Journal of Environmental Research and Public Health*, vol. 16, 05 2019.

[6] T. M. Mitchell *et al.*, *Machine learning*. McGraw-Hill,In, 1997.

[7] A. M. S. M. K. A. e. a. Yadav, Abhishek, "Better healthcare using machine learning," *International Journal of Advanced Research in Computer Science*, vol. 9, no. 1, 2010.

[8] O. Mbaye, M. L. Ba, G. Camara, A. Sy, B. M. Mboup, and A. Diallo, "Towards an efficient prediction model of malaria cases in senegal," in *International Conference on Innovations and Interdisciplinary Solutions for Underserved Areas*. Springer, 2019, pp. 173–188.

[9] R. Gholami and N. Fakhari, "Support vector machine: Principles, parameters, and applications," in *Handbook of Neural Computation*. Elsevier, 2017, pp. 515–535.

[10] P. Kaviani and S. Dhotre, "Short survey on naive bayes algorithm," *International Journal of Advance Research in Computer Science and Management*, vol. 04, 11 2017.

[11] S. P. Morgan and J. D. Teachman, "Logistic regression: Description, examples, and comparisons," *Journal of Marriage and Family*, vol. 50, no. 4, pp. 929–936, 1988.

[12] L. Rokach and O. Maimon, *Decision Trees*. Springer, 01 2005, vol. 6, pp. 165–192.

[13] T. Evgeniou and M. Pontil, "Support vector machines: Theory and applications," in *Studies in Fuzziness and Soft Computing*, vol. 2049, 01 2001, pp. 249–257.

[14] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[15] B. Mehlig, "Artificial neural networks," in *arXiv*, 2019, 1901.05639.

[16] H. De Oliveira, M. Prodel, and V. Augusto, "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, pp. 1743–1748.

[17] D. Tomar and S. Agarwal, "A survey on data mining approaches for healthcare," *International Journal of Bio-Science and Bio-Technology*, vol. 5, no. 5, pp. 241–266, 2013.

[18] S. S. Yadav and S. M. Jadhav, "Deep convolutional neural network based medical image classification for disease diagnosis," *Journal of Big Data*, vol. 6, no. 1, p. 113, 2019.

[19] V. J. Kadam, S. S. Yadav, and S. M. Jadhav, "Soft-margin svm incorporating feature selection using improved elitist ga for arrhythmia classification," in *International Conference on Intelligent Systems Design and Applications*. Springer, 2018, pp. 965–976.

[20] S. S. Yadav and S. M. Jadhav, "Machine learning algorithms for disease prediction using iot environment," *International Journal of Engineering and Advanced Technology*, vol. 8, no. 6, pp. 4303–4307, 2019.

[21] V. Kadam, S. Jadhav, and S. Yadav, "Bagging based ensemble of support vector machines with improved elitist ga-svm features selection for cardiac arrhythmia classification," *International Journal of Hybrid Intelligent Systems*, vol. 16, no. 1, pp. 25–33, 2020.

[22] S. S. Yadav, V. J. Kadam, and S. M. Jadhav, "Comparative analysis of ensemble classifier and single base classifier in medical disease diagnosis," in *International Conference on Communication and Intelligent Systems*. Springer, 2019, pp. 475–489.

[23] S. S. Yadav, S. M. Jadhav, R. G. Bonde, and S. T. Chaudhari, "Automated cardiac disease diagnosis using support vector machine," in *2020 3rd International Conference on Communication System, Computing and IT Applications (CSCITA)*. IEEE, 2020, pp. 56–61.

[24] S. S. Yadav, S. M. Jadhav, S. Nagrale, and N. Patil, "Application of machine learning for the detection of heart disease," in *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*. IEEE, 2020, pp. 165–172.

[25] S. S. Yadav and S. M. Jadhav, "Detection of common risk factors for diagnosis of cardiac arrhythmia using machine learning algorithm," *Expert Systems with Applications*, p. 113807, 2020.

[26] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, pp. 1–16, 2019.

[27] P.-W. Wang and C.-J. Lin, "Support vector machines." 2014.

[28] D. Velusamy and K. Ramasamy, "Ensemble of heterogeneous classifiers for diagnosis and prediction of coronary artery disease with reduced feature subset," *Computer Methods and Programs in Biomedicine*, vol. 198, p. 105770, 2020.

[29] M. M. Ghiasi, S. Zendehboudi, and A. A. Mohsenipour, "Decision tree-based diagnosis of coronary artery disease: Cart model," *Computer Methods and Programs in Biomedicine*, vol. 192, p. 105400, 2020.

[30] J. A. Anderson, "A simple neural network generating an interactive memory," *Mathematical biosciences*, vol. 14, no. 3-4, pp. 197–220, 1972.

[31] S. Raschka, *Python machine learning*. Packt Publishing Ltd, 2015.

[32] C. Sokhna, B. M. Mboup, P. G. Sow, G. Camara, M. Dieng, M. Sylla, L. Gueye, D. Sow, A. Diallo, P. Parola, D. Raoult, and P. Gautret, "Communicable and non-communicable disease risks at the grand magal of touba: The largest mass gathering in senegal," *Travel Medicine and Infectious Disease*, vol. 19, pp. 56 – 60, 2017.