Advances in Parallel Computing Technologies and Applications
D.J. Hemanth et al. (Eds.)

doi:10.3233/APC210119

A Noninvasive Model to Detect Malaria Based on Symptoms Using Machine Learning

Ruban S [a,1], Naresh A [b] and Sanjeev Rai [c]
[a] *Asso prof, PG Dept. of IT, St Aloysius college, Mangalore*
[b] *Student, PG Dept. of IT, St Aloysius college, Mangalore*
[c] *Chief Research officer, Father Muller Medical College, Mangalore*

Abstract. The impact of Artificial Intelligence in the domain of Healthcare has been growing, day by day. These applications bring a drastic change in the healthcare system and affects our lives based in the change it brings to the Patientcare system, transforming the traditional way of handling sicknesses and diseases. Machine Learning algorithms that use data, have a big role in the AI based applications that are used in the Healthcare. Hence the Data source and the nature of Data holds an important role in developing effective AI based solutions for many health issues in the society. Data is available in all the hospitals and medical care facilities for many years now. However, without transforming them into a format where Machine Learning algorithms work, it is impossible to use them to develop an AI based application. In this research paper, we briefly discuss the process of developing an AI based application to predict Malaria, which is one of the most common vector borne diseases in the coastal districts of Karnataka. This pioneer work was done over the data collected from the clinical notes of a 1500 bed hospital situated in Mangalore. Few machine learning algorithms like Logistic regression, Support vector machine XGB Booster classifier, CAT Booster Classifier and Random forest classifier were used over the dataset. Our experimental study revealed that, Random Forest classifier works efficiently for this data set, compared with the other algorithms that we used. It gave the best accuracy of 90.92.

Keywords. Artificial Intelligence, Support vector machine, machine learning, random forest, Logistic regression, Malaria, vector borne disease.

1. Introduction

The transforming power of Artificial Intelligence in the health care sector is very evident, in modern times [1]. As it is defined conventionally, AI is about developing machines with intelligence in contrast to the intelligence of human beings [2]. With more and more advances happening in the collection of data, processing and computing, intelligent systems are now aiding in these various tasks that once depended on human arbitration. From Finance to Medical care [2] scenarios are transforming drastically, in

[1] Ruban S, Department of IT, AIMIT, St Aloysius College (Autonomous), Mangalore, India.
E-mail: ruban@staloysius.ac.in.

a way people never imagined before. One could say that machine learning is applied in variety of application such as [17] cost control, [18] soil environment preservation etc,. but, all these benefits do come with various limitations. The limitations ranges from the algorithms, hardware implementation, development of application etc. AI involves developing systems that exhibit cognitive aptitude that uses technologies such as Machine Learning [3]. Every instance of the role of AI that we hear about, and its applications [4] in Health care takes advantage of the Data. Despite the digital revolution, most of the medical data are still handwritten [5]. Problems arise when other stake holders are involved either for interpretation or study. Poor handwritten clinical notes poses a serious threat for researchers who are involved in data analysis. Dakshina kannada is one among the coastal districts of Karnataka, and reports many Malaria cases in a year. It is also named as one of the malaria endemic district [6] in the state and the country. Malaria is one of the important vector borne diseases globally [7] . Few studies have been done to analyze the trends of vector borne diseases [8]. Artificial Intelligence (AI) has been used as a surveillance and prediction tool to predict vector borne diseases [9]. The researchers of the above study came out with a system, that could predict the outbreak of dengue much earlier taking advantage of various data and parameters that were stored in different silos. Similar studies have also been done in other places as well [10]. Few such works are carried out in our country. [11–13]. However, in Indian scenario, there is hardly any study that is done in a deeper level involving clinical notes digitization. Many works take the demographic details and analyze. So an attempt was made to study the trends, symptoms, treatments of Malaria patients from the hospital records who were admitted in the span of four years (2015-2018) in a Medical College Hospital in Mangalore, India. The study was conducted after taking Ethics committee permission of the medical college. The data are maintained by the Hospital Medical Records Department (MRD). Section 2 elaborates the Framework that was used and the following sections narrates the results that we obtained from this study and is then followed by conclusion.

This experimental work results have helped to understand the dynamics of the Malaria fever in this region, and can serve as a tool to assist the doctors, for managing patients quickly and effectively.

2. Materials and Methods

2.1. Data Sources

The Real Time Data Collection was done primarily in two locations - the DHO office in Mangalore, and the Father Muller Medical College. The Data from the DHO Office were gathered from different records, files and also by visiting different primary Health centers (PHC) and National Urban Health Mission centers (NUMC) in and around Mangalore. The data that were gathered from the PHC and NUMC did not have detailed clinical notes but has only basic demographic details. Hence the Data related to Malaria from Father Muller Medical College was accessed after getting the approval from the scientific and Ethics committee of the Father Muller Medical College, Mangalore. We followed the CRISP-DM model, for our experimental study, understanding the data, transforming it into a format where the Model could be built and finally the evaluation.

Figure 1. CRISP-DM Model.

2.2. Malaria Data

There are various symptoms for a patient to be identified as having Malaria. If a patient has fever with migraine, back pain, chills, rigors, sweating, nausea and puking [14]. A confirmed complicated/severe malaria is defined as a confirmed case with symptoms/signs of complicated/severe malaria (prostration, impaired consciousness, respiratory distress (acidotic breathing), multiple convulsions, circulatory collapse, abnormal bleeding, jaundice, hemoglobinuria, severe anemia, etc.) Confirmed Malaria cases that were treated in Father Muller Medical college hospital from the year 2014 to 2018 were considered for the study.

Fields	**Sample Data 1**	**Sample Data 2**	**Sample Data 3**
IP Number	54xxxx45	87xxxx41	46xxxx32
Patient Name	Xxxxxxx	Xxxxxxx	Xxxxxxx
Age	45	54	19
Sex	Male	Male	Female
City	Mangalore	Chickmangalur	Kasargod
DOA	20-11-2014 /09:15	10-11-2015 /10:15	02-11-2017/ 09:15
Discharge Date	25-11-2014 / 01:31	17-11-2015 / 02:31	10-11-2017/ 01:31
Primary Code	B54	B50.9	B50.9
Primary Code Description	Unspecified Malaria	Plasmodium vivax Malaria without complication.	Plasmodium vivax Malaria without complication

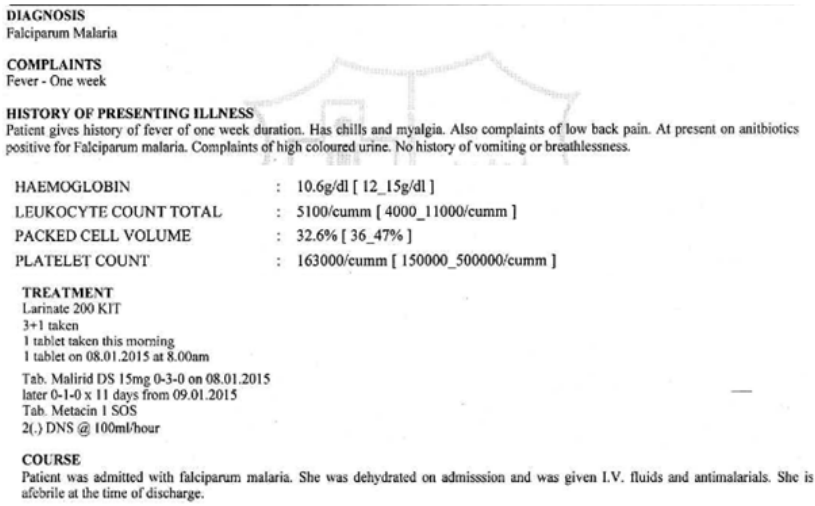
Figure 2. Sample of Malaria Data maintained in the Registration Department.

Few case sheets were reported as unspecified malaria. The individual patient medical records were accessed. The Data were available in two departments. The Registra-

tion department maintains the details of the in-patients regarding their Inpatient number, Name, Sex, Age, City, Date of admission and Date of discharge.

2.3. Data Gathering from Medical Records Department

The Data related to Malaria from Father Muller Medical College was accessed after getting the required permission from scientific and Ethical committee of the Father Muller Medical College, Mangalore. The Data related to Malaria were stored as Electronic Medical Records (EMR). The case sheets were scanned and stored in the MRD repository. The corresponding patient history was accessed through the In-Patient number, is stored in the Medical Records Department and the data that is stored in the Registration department.



DIAGNOSIS
Falciparum Malaria

COMPLAINTS
Fever - One week

HISTORY OF PRESENTING ILLNESS
Patient gives history of fever of one week duration. Has chills and myalgia. Also complaints of low back pain. At present on anitbiotics positive for Falciparum malaria. Complaints of high coloured urine. No history of vomiting or breathlessness.

HAEMOGLOBIN	:	10.6g/dl [12_15g/dl]
LEUKOCYTE COUNT TOTAL	:	5100/cumm [4000_11000/cumm]
PACKED CELL VOLUME	:	32.6% [36_47%]
PLATELET COUNT	:	163000/cumm [150000_500000/cumm]

TREATMENT
Larinate 200 KIT
3+1 taken
1 tablet taken this morning
1 tablet on 08.01.2015 at 8.00am
Tab. Malirid DS 15mg 0-3-0 on 08.01.2015
later 0-1-0 x 11 days from 09.01.2015
Tab. Metacin 1 SOS
2(.) DNS @ 100ml/hour

COURSE
Patient was admitted with falciparum malaria. She was dehydrated on admisssion and was given I.V. fluids and antimalarials. She is afebrile at the time of discharge.

Figure 3. Sample Discharge summary in clinical notes.

2.4. Data Preprocessing

Major portion of the time in this research study was spend in this Data pre-processing step. All the health data thus collected go through Data pre-processing i.e., cleaning process where unnecessary information was removed. This phase consists of four primary sub steps: Data Cleaning, Data Integration, Data Transformation and Data Reduction.

- Data Cleaning: Data cleaning was done to sort our issues related to missing values, non-readable handwritings, and redundant data within the data that were available in the clinical notes.
- Data Integration: The Data were extracted from the Registration department which gave us the demographic information about the patient and other data from the medical records department which gives information about the treatment that was given. These data sources must be integrated, to a single data point which is uniform that can be analyzed.

- Data Transformation: The data we collected was in formats that are not optimal for processing. For example, if dates are involved, the data must be formatted from text to date format. In this state, we convert raw data into a useful format that can be processed with mathematical libraries. In this project the date of admission and date of discharge fields are used to compute the number of days the patient was admitted.
- Data Reduction: Redundant data is identified and removed. Any unnecessary data is removed. This ensures that only valid data is used for processing.

2.5. Data Processing

The preliminary idea was to take screenshots of the patient discharge sheets and to extract text from those images. Each image which contained patient information from the day of his/her arrival to the day of discharge was recorded. In order to extract data from the images we used a python tool called Python-Tesseract. All the images were run through the modified python program and the image files were transformed into text files which contained all the textual information got from the images. Still there was a challenge with respect to the extracted files. Some of them were so distorted and blur, so the python program couldn't recognize the words in them and some of the extracted data was wrong. So, the only alternative was to store the data in the database by manually entering the patient information. For this we created a python script or program which takes the user input by using the input () function of python. Then the data was stored in Mongo DB using the python's pymongo package which allows us to set the database and the collections to store the data.

The next step is to create a prediction model. For this the entire data is split into train', 'test' and 'validation' set. Since the task involved in this model is classification, we created a supervised machine learning model [15, 16]. The model was created using the classification algorithm. In our case we used Logistic Regression, Support Vector Machine, XGB Booster classifier, CAT Booster Classifier and Random Forest. The model was then fed with input data i.e., the training data which is the set of input prepared from the clinical notes of the patients who were attended in the Medical College Hospital for Malaria. For the model to be trained or fit, few of the parameters were adjusted at regular intervals to get the best accuracy. Further, the model tunes its attributes based on the frequent evaluation results on the validation set. The working accuracy of the model is derived from the test set. The program was written using Python 3.7 using the following libraries and packages such as Numpy, Pandas, MatPlotlib, SKlearn, Dash, Spell Checker and nGram.

3. Results and Discussion

The Exploratory Data Analysis was performed over the data that were collected. Few of the results are presented below. From the Data that was collected from the Father Muller Medical College, Mangalore, the following insights were derived. The data has now been enabled to perform any machine learning tasks such as classification or prediction or regression based on the need. There are various algorithms for each of the tasks that are mentioned above. Since the task that we have been trying to solve is a classification

related task, we tried to find out the different algorithms that can be used for building a model. We tried using Logistic Regression, Support Vector Machine, XGB Booster classifier, CAT Booster Classifier and Random Forest for training the model. The various values that were generated for different metrics such as Accuracy, Precision, Recall and F- Measure are displayed below.

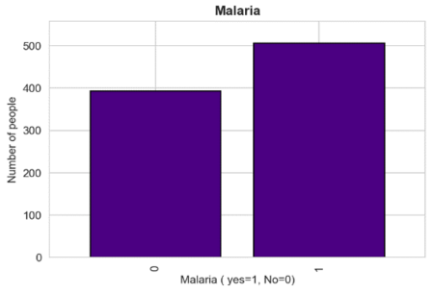


Figure 4. Malaria Cases

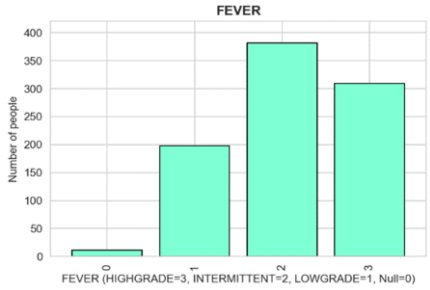


Figure 5. Malaria Cases with Symptoms of Fever

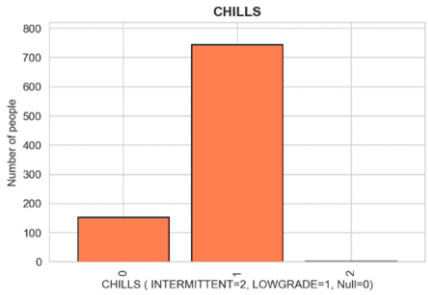


Figure 6. Malaria Cases with Symptoms of Chills

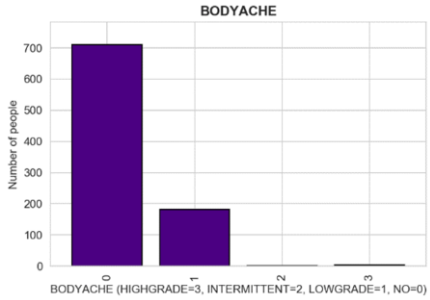


Figure 7. Malaria Cases with Symptoms of Body Ache

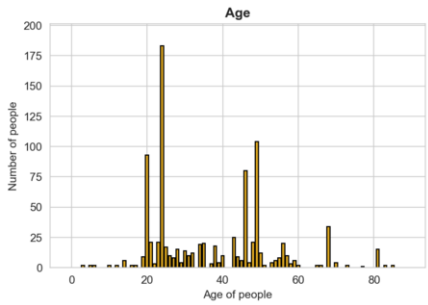


Figure 8. Malaria cases with Age classification

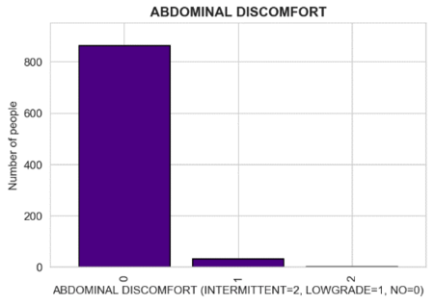


Figure 9. Malaria Cases with Abdominal Discomfort

Table 1. Performance Comparison of Different Machine Learning Algorithms

Algorithm	Precision	Recall	F1-Score	Accuracy
Logistic Regression	0.59	0.72	0.65	62.63
Support Vector Machine	0.55	0.46	0.50	55
Random Forest Classifier	0.75	0.76	0.76	90.97
XGB Booster Classifier	0.74	0.79	0.77	86.25
CAT Booster Classifier	0.75	0.70	0.72	88.19

This Figure shows the results that we obtained from each classifiers. From this, we can understand that the model was effective as it gives the accuracy highest of 90.97 for the Random Forest classifier algorithm. From the results that is obtained from the test we can conclude that the Random Forest Classifier decision tree classifier gives the best result out of all the classifiers we used in the study. It gives higher accuracy than the other classifiers. Random Forest classifier algorithm is one of the best machine learning algorithms to examine the data categorically and continuously. Thus in conclusion we can say that if the dataset is larger in size the model will give more accurate result.

4. Conclusion

This research study based on clinical notes of the patient, treated for Malaria, provides an insight into the types of symptoms prior to hospital admission. It also explores the efficiency of diagnostic treatment for Malaria. The quicker a physician assesses based on the symptoms, more effective the treatment tends to be. This study was done with data collected from one specific location. More data from different hospital setting and different places would increase the efficiency of the System. We intend to add more real time data from different hospital settings. However, the same steps that were performed in the preprocessing stages can be repeated for any hospital setting to gather data and transform the raw clinical data into a meaningful data over which effective AI based model can be built.

5. Acknowlegement

The authors would like to acknowledge, that this work was done in the lab funded by Vision Group of science and Technology (VGST), Government of Karnataka, under the Grant scheme K- FIST(L2)-545 and the data was collected from Father Muller Medical College Hospital, based on the Ethics committee approval via protocol no: 126/19(FMMCIEC/CCM/149/2019) on 12.06.2019.

References

[1] Rong G, Mendez A, Assi EB, Zhao B, Sawan M. Artificial intelligence in healthcare: review and prediction case studies. Engineering. 2020 Mar 1;6(3):291-301.

[2] Weng J, McClelland J, Pentland A, Sporns O, Stockman I, Sur M, Thelen E. Autonomous mental development by robots and animals. Science. 2001 Jan 26;291(5504):599-600.

[3] Priya, D. Top 5 Limitations of Artificial Intelligence. [Internet]. [updated 2019 March 17; cited 2020 September 27]. Available from: https://www.analyticsinsight.net/top-5-limitations-artificial-intelligence.
[4] Huang G, Huang GB, Song S, You K. Trends in extreme learning machines: A review. Neural Networks. 2015 Jan 1;61:32-48.
[5] Guo Y, Liu Y, Oerlemans A, Lao S, Wu S, Lew MS. Deep learning for visual understanding: A review. Neurocomputing. 2016 Apr 26;187:27-48.
[6] Rodríguez-Vera FJ, Marin Y, Sanchez A, Borrachero C, Pujol E. Illegible handwriting in medical records. Journal of the Royal Society of Medicine. 2002 Nov;95(11):545-6.
[7] Rajesh BV, Kumar A, Achari M, Deepa S, Vyas N. Malarial trend in Dakshina Kannada, Karnataka: an epidemiological assessment from 2004 to 2013. Indian Journal of Health Sciences and Biomedical Research (KLEU). 2015 Jul 1;8(2):91-4.
[8] Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, Drake JM, Brownstein JS, Hoen AG, Sankoh O, Myers MF. The global distribution and burden of dengue. Nature. 2013 Apr;496(7446):504-7.
[9] George T, Jakribettu RP, Yesudas S, Thaliath A, Pais ML, Abraham S, Baliga MS. Trend analysis of dengue in greater Mangalore region of Karnataka India: Observations from a tertiary care hospital. International Journal of Advanced Research (IJAR). 2018;4(6):92-6.
[10] Sharma R. Epidemiological investigation of malaria outbreak in village Santej, district Gandhi Nagar (Gujarat). Indian Journal Preventive Social Medicine. 2006;37(3):125-32.
[11] Zacarias OP, Boström H. Predicting the incidence of malaria cases in mozambique using regression trees and forests. International Journal of Computer Science and Electronics Engineering (IJCSEE). 2013;1(1):50-4.
[12] Linder N, Turkki R, Walliander M, Mårtensson A, Diwan V, Rahtu E, Pietikäinen M, Lundin M, Lundin J. A malaria diagnostic tool based on computer vision screening and visualization of Plasmodium falciparum candidate areas in digitized blood smears. PLoS One. 2014 Aug 21;9(8):e104855.
[13] Devi SS, Sheikh SA, Talukdar A, Laskar RH. Malaria infected erythrocyte classification based on the histogram features using microscopic images of thin blood smear. Indian Journal of Science and Technology. 2016 Dec 20;9:1-10.
[14] Darla, B. Malaria: Causes, Symptoms, and Diagnosis. [Internet]. [updated 2019 March 7; cited 2020 September 24]. Available from: https://www.healthline.com/health/malaria#diagnosis.
[15] Sidey-Gibbons JA, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC medical research methodology. 2019 Dec;19(1):1-8.
[16] Beam AL, Kohane IS. Big data and machine learning in health care. Journal of the American Medical Association. 2018 Apr 3;319(13):1317-8.
[17] Xu X. Machine learning-based prediction of urban soil environment and corpus translation teaching. Arabian Journal of Geosciences. 2021 Jun;14(11):1-5.
[18] Luo Y. Environmental cost control of coal industry based on cloud computing and machine learning. Arabian Journal of Geosciences. 2021 Jun;14(12):1-6.