S Ruban, Sanjeev Rai

Enabling Data to develop an AI based application for detecting Malaria and Dengue

Abstract: Data as we know, is no more an option for medical care reforms. It is the central theme over which the entire process of health reforms is on. Any application of AI, and its benefits in health care takes advantage of Data. These data which have been acquired and stored over the span of several years, are at various levels starting from patient admission, diagnostics, treatment summary, lab reports, scans and x-rays, patient discharge summary etc. Though it is easy to state "we need Data". In reality, we did have data stored in multiple repositories such as hospital medical records, doctors' office, clinical records and insurance records. However, the challenge remains in enabling this data which is mostly a hand-written or scanned image of hand-written records into a format which is computable. This is a foremost challenge to develop AI based solutions on Indian health data stored in multiple health repositories of our country. The second important challenge being the need to integrate the data that are available in different repositories. This case study discusses a pioneering work undertaken in a 1500 bed hospital in the coastal district of Karnataka, to develop an AI based application based on the hospital Data for detecting Malaria and Dengue. This lists out the various challenges faced at every phase of developing this health care application.

Results:

This study elaborates the use of various data available in various repositories of hospital and medical college to develop an AI based application. The step by step method to transform clinical notes into a data format that could be in turn useful for diagnosis and prognosis has been described. The practical issues and challenges faced in working with Indian hospital data has been dealt with.

Conclusion:

The task of transforming clinical notes into a format that can provide value and insight is elaborated. This will inspire the medical researchers and clinicians to adopt machine learning in their studies. The principles that we have adopted here can be applied to any other similar task.

S Ruban, PG Department of Software Technology, St Aloysius College (Autonomous), Mangalore - 575022

Sanjeev Rai, Chief Research Officer, Father Muller Medical College, Mangalore – 575002.

Keywords: Artificial Intelligence, Machine Learning, Health Care, Medical Care, Dengue, Malaria.

1 Background:

Data is a four letter word that forms the key of the AI revolution that is transforming the industry like never before. The impact of Artificial Intelligence in health care sector is very evident, in recent times [1]. As it is defined traditionally 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 assisting in these various tasks that once depended on human intervention. From Finance to Medical care [2] scenarios are changing drastically, in a way people never imagined before. However, all these advantages do come with various challenges. The challenges ranges from the algorithms, hardware implementation, development of application etc. One such challenge is the nature of data on which the AI based solutions are built. Where do we find Data? It is all around. It is true. However, one hindrance to adopt Artificial Intelligence in Healthcare is the n on availability of computable Data. Most of the time the data is either available in silos, not consistent or of poor quality [3]. AI involves developing systems that exhibit cognitive aptitude that uses technologies such as Machine Learning [4]. Every instance of the role of AI that we hear about, and its applications [5] in Health care takes advantage of the Data. Data is our asset. Every organization has data and especially in health sector, all hospital and medical institution have it. It could have been acquired at various levels starting from patient admission, diagnostics, treatment summary, lab reports, scans and x-rays, patient discharge summary etc. Though, we always state “we need Data”. We did have data stored in multiple repositories such as hospital medical records, doctors’ office, clinical records and insurance records. However, the challenge remains in transforming this data which is mostly a hand-written or scanned image of hand-written records into a format which is computable. Despite the digital revolution, most of the medical data are still handwritten [6]. 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. This is one of the foremost challenges to develop AI based solutions on Indian Medical Data that are available in multiple health repositories in our country. The second important challenge is the need to integrate the data that are available in different repositories. This research study lists out various strategies adopted to overcome these above two challenges that hinders the development of AI-based solutions based on the valuable Indian medical records available in different forms and scattered across different repositories.

This research study to develop AI-based application for detecting Malaria and Dengue is based on the data collected from a 1500-beded hospital, located in Dakshina Kannada district of Karnataka, India. This is one among the coastal districts of Karnataka, and reports many Dengue and Malaria cases in a year. It is also named as one of the malaria endemic district [7] in the state and the country. Recent data shows a

declining trend of these vector borne diseases in the region. However, few indices also display the need of a multidisciplinary approach to solve the problem. Similarly, Dengue which is one of the most important vector borne diseases globally [8] also poses a serious threat to this coastal district. Few studies have been done to analyze the trends of Dengue [9]. These studies have indicated that the surge of these diseases happen in monsoon and the early winter. The emerging trends, helps to plan certain preventive strategies to control the spread of these diseases. Artificial Intelligence (AI) has been used as a surveillance and prediction tool to predict vector borne diseases such as Dengue in different parts of the world [10]. 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 [11]. Few such works are carried out in our country [12–14]. 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 Dengue and Malaria patients from the hospital records who were admitted in the span of four years (2015-2018). The study was conducted after the permission from the Research committees of the medical college. The medical data are maintained by the Hospital Medical Records Department (MRD). The hospital is yet to introduce the Electronic Health Records (EHR). The case sheets we accessed for this study are the Electronic Medical Records (EMR). These are scanned images of the patient case sheets. The patient medical record has clinical notes that are hand written and the discharge summary, which is the only scanned document that is legible and easy to read. The process of transforming this data that is handwritten to a data set where an AI based application is built is elaborated below. Section 2 elaborates the methodologies that were used and the following section discusses the results that we obtained from this study followed by a conclusion.

This study results have helped to understand both the Malaria and Dengue fever dynamics in this region, and can be used to predict the type of fever based on symptoms and hence probably can be used to assist the doctors for treating their patients quickly and effectively.

2 Materials and Methods:

2.1 Study Area:

Dakshina Kannada is the southern coastal district of Karnataka. The population here is around 21 lakhs, based on the previous census. The district is further divided into five talukas (portion of a revenue district). Mangalore being the headquarters of the district, houses several health care and other educational institutions. The city is well connected with the outside world through the International Airport, Port, Train stations and Bus stations. With Migrant workers pouring inside the city from various other parts of the country, this city also houses a good portion of student community, who step into the city for higher education in Medicine, Engineering, Para Medical courses and Arts and Science. This study was conducted at Father Muller Medical College Hospital, Mangalore, Karnataka state, India. The Following Fig.6.1 elaborates the detailed work flow process followed in this study.

2.2 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 case sheets rather only basic demographic details. Hence the Data related to Dengue and 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.

2.2.1 Data related to Dengue:

A patient suffering from Dengue presents few Mild symptoms such as fever, aches and pains. However most common symptom of dengue is fever with any of the following: eye pain, headache, muscle pain, rash, bone pain, nausea/vomiting, joint pain [16]. Symptoms of dengue typically last 2–7 days. Most people will recover after about a week. 4590 positive confirmed dengue cases that were treated in Father Muller Medical college hospital during the year 2015-2018 were taken for the study. The individual patient medical records were accessed. The Data were available in two departments. The Registration department had patient details such as their IP number, name, sex, age, city, date of admission and date of discharge. The MRD department maintains a huge repository of case sheets where they are organized based on the ICD code [17].

Tab. 6.1. Format of Dengue Data maintained in the Registration Department

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	A90	A90	A90
Code Description	Dengue Fever	Dengue Fever	Dengue Fever

2.2.2 Data related to Malaria:

Similarly, a case of Malaria is considered when a patient presents with fever accompanied with headache, backache, chills, rigors, sweating, myalgia, nausea and vomiting [18]. 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.). 4554 confirmed Malaria cases that were treated in Father Muller Medical college hospital from the year 2014 to 2018 were considered for the study. Few case sheets were reported as unspecified malaria. The individual patient medical records were accessed. The Data were available in two departments. The Registration department maintains the details of the in-patients regarding their Inpatient number, Name, Sex, Age, City, Date of admission and Date of discharge.

Tab. 6.2. Format of Malaria Data maintained in the Registration Department

Fields	Sample Data 1	Sample Data 2	Sample Data 3
IPNo	32xxxx45	87xxxx41	46xxxx32
Patient Name	Xxxxxxx	Xxxxxxx	Xxxxxxx
Age	38	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

2.2.3 Data Gathering from the Medical Records Department (MRD):

The Data related to Dengue and 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 Dengue and

Malaria were stored as Electronic Medical Records (EMR). The case sheets were scanned and stored in the MRD repository. However, the analysis of data was difficult. The corresponding patient history was accessed through the IP number that acts as a unique identifier for the data that is stored in the MRD department and the data that is stored in the Registration department. As illustrated in Fig. 6.2.

2.3 Data Pre-processing:

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. With the pre-processed data, we started finding patterns. We collected N-grams from the data using speechPyspellchecker package, which uses Levenshtein distance algorithm. The following steps were performed over the real time data collected from various Data sources related to Vector Borne diseases. In the initial phase, we dealt with various data quality issues. The initial data gathered is raw and usually not in a format to run the required analysis. It contains missing entries, inconsistencies, and semantic errors. After gathering the data, we clean and transform the data by manually editing it in the spreadsheet or by using Python. This step though does not give us much meaningful patterns or insight, however, consistently helps us to figure it out the right assumptions that should be made. This helps us to apply right models that will assist in the important step of analysis. Data after re-formatting can be converted to JSON, CSV or any other format that makes it easy to load into one of our tools.

Exploratory data analysis forms an integral part at this stage, as the summarization of the clean data can help identify outliers, anomalies, and patterns that can become usable in the subsequent steps. This is the step that answer the question of the purpose for which data was collected. This phase consists of four primary sub steps: Data Cleaning, Data Integration, Data Transformation and Data Reduction.

a. Data Cleaning: Data cleaning helps in pre-processing. This helps to handle missing data, noisy data, detection and removal of outliers, minimizing duplication and computed biases within the data.

b. Data Integration: Different health organizations give us different data sources. These data sources must be integrated, to a single data point which is uniform that can be analyzed by the computer. In this case there were two types of data provided from two departments. One 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.

c. 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.

d. Data Reduction: Redundant data is identified and removed. Any unnecessary data is removed. This ensures that only valid data is used for processing.

2.4 Data Processing:

The initial 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. Python-Tesseract is an optical character recognition (OCR) tool for python. That is, it will recognize and “read” the text embedded in images. Python- tesseract is a wrapper for Google’s Tesseract-OCR Engine. It is also useful as a stand-alone invocation script to tesseract, as it can read all image types supported by the Pillow and Leptonica imaging libraries, including jpeg, png, gif, bmp, tiff, and others. All the images where run through the modified python program and the image files were converted 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 had noise in them 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 or in an Excel sheet by manually entering certain patient information which were not clear or blur. 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 an Excel sheet with all the required features (symptoms) that were required to identify the type of fever that the patients suffer with.

The next step is to create a prediction model. For this the entire data is split into ‘train’, ‘test’ and ‘validation’ set. Creating a supervised machine learning model [19] [20] is all about making a program that is able to generalize to input samples that it has never seen before. This task requires exposing the model during training to a certain number of variations of input examples, which is likely to lead to acceptable accuracy. Training set includes the set of input examples that the model will be fit into or trained on by adjusting the parameters. For the model to be trained, it needs to be evaluated periodically. Subsequently, the model will tune its parameters based on the frequent evaluation results on the validation set. This corresponds to the final evaluation that the model goes through after the training phase (utilizing training and validation sets) has been completed. This step is critical to test the generalizability of the model. By using test set, we can get the working accuracy of our model. The following tools and technologies are used. Python 3.7, Numpy, Pandas, MatPlotlib, SKlearn, Dash, Spell Checker, nGram. Mongo DB.

2.5 Data Dictionary:

Data Dictionary is developed as part of the Data processing in this project. This Data Dictionary was decided based on the Domain expertise. As this project was dealing with vector borne diseases such as Dengue and Malaria. The physicians in the medical college hospital were discussed about the Data to be captured from the clinical history that is recorded in the medical records. All the data parameters that has to be captured from the clinical history of the case files were extracted based on the domain experts’ guidance. It acts as a Metadata. The following Data Dictionary was created for extracting information from the clinical files as shown in Fig. 6.3.

3. Results and Discussion:

3.1 Dengue:

Few Experimental Results are presented below. Data were collected from the District Vector Borne Diseases controlling officer's office, Primary Health Center's around Mangalore and Father Muller Medical College, Mangalore. However, since we were accessing the clinical notes from the Father Muller Medical College Hospital, the following insights are derived from them. As a norm for every data science project, good amount of time was invested to clean the data, and later the data were processed. The samples collected ranges from 2015 till 2018. The analysis that was carried out in the dengue data based on various symptoms are illustrated in the Fig. 6.5 to 6.11.

3.2 Malaria:

The insights were derived from the Data that was collected from the District Vector Borne Diseases controlling officer's office, Primary Health Center's around Mangalore and Father Muller Medical College, Mangalore. However since, we were accessing the clinical notes from the Father Muller Medical College Hospital, the following insights are derived from the medical college data. The analysis that was carried out in the malaria data based on various symptoms are illustrated in the Fig. 6.12 to 6.16.

3.3 Performance Evaluation of Classifiers:

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, we tried to find out the different algorithms that can be used for building a model. Before we start deciding the algorithm that should be used, we split the dataset into two parts. Machine Learning algorithms, or any algorithm for that matter, has to be first trained on the data distribution available and then validated and tested before it can be deployed to deal with real-world data. We tried using Random Forest classifier, extreme Gradient Boosting classifier and Cat Boost classifier for training the model. The model was built to classify whether a person has Dengue or not. The various values that were generated for different metrics such as Precision, Recall, Accuracy, and F- Measure are displayed below.

Tab.3.1 Performance Evaluation of Classifiers

Algorithm	Precision	Recall	F1-score	Accuracy
Random Forest Classifier	0.75	0.76	0.76	90.97
eXtreme Gradient Boosting Classifier	0.74	0.79	0.77	86.25
Cat Boost Classifier	0.75	0.70	0.72	88.19

This table shows the results that we obtained from each classifiers. From this, we can understand that the model was effective when we used the Random Forest Classifier as it gives the accuracy highest of 90.97. From the results that is obtained from the test we can conclude that the Random Forest classifier gives the best result out of all the classifiers we used in the study. It gives higher accuracy than the other classifiers. Cat Boost classifier gave an accuracy of 88.19 and eXtreme Gradient Boosting classifier gave the accuracy of 86.25.

4. Conclusion:

This research study based on clinical notes of the patient, treated for Dengue and Malaria, provides an insight into the types of symptoms prior to hospital admission. It also explores the efficiency of diagnostic treatment for both Dengue and 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. 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. This study will hopefully inspire more researchers to take up the task of transforming the clinical data that are available in different parts of this country. The days are not too far for AI systems to be built for different health challenges pertaining to different sections of people in this diverse country.

5. Abbreviations:

EHR – Electronic Health Record, EMR – Electronic Medical Record, HTN – Hypertension
IHD – Ischemic Heart Disease, OCR – Optical Character Recognition, CSV- Comma separated Value, JPEG – Joint photographic Experts Group, ICD – International classification of diseases, AI – Artificial Intelligence, ML – Machine Learning, DL – Deep Learning.

6. Competing Interests:

The author declares that they have no competing interests.

7. Acknowledgement:

The author would like to acknowledge, that this work was done in the lab funded by Vision Group of science and Technology (VGST), Govt of Karnataka, under the Grant scheme K- FIST(L2)-545.

8. Ethical Approval:

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.

9. Availability of Data and Materials.

The data will be made available on request to the author with consent and permission from the concerned authorities.

10. References:

[1] Guogang Rong, Arnaldo Mendez, Elie Bou Assi, Bo Zhao, Mohamad Sawan. Artificial Intelligence in Healthcare: Review and prediction case studies. Engineering (2020), 6(3), 291-301.

[2] J. Weng, J. McClelland, A. Pentland, O. Sporns, I. Stockman, M. Sur, et al. Autonomous mental development by robots and animals. Science (2020), 291(5504), 599-600.

[3] B.X. Tran, G.T. Vu, G.H. Ha, Q.H. Vuong, M.T. Ho, T.T. Vuong, et al. Global evolution of research in artificial intelligence in health and medicine: a bibliometric study. J Clin Med (2019), 8(3), 360.

[4] Limitations of Artificial Intelligence. [Accessed on September 27, 2020 at https://www.analyticsinsight.net/top-5-limitations-artificial-intelligence].

[5] G. Huang, G.G. Huang, S. Song, K. You. Trends in extreme learning machines: a review. Neural Network (2015), 61, 32-48.

[6] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, M.S. Lew. Deep Learning for visual understanding: a review. Neurocomputing (2016), 187, 27-48.

[7] F. Javier Rodriquez – Vera Y Marin, A Sanchez, C Borrachero, E pujal. Illegible handwriting in medical records. Journal of the royal society of medicine (2002), 95, 545-546.

[8] Shiva Kumar, Rajesh BV, Kumar A, Achari M, Deepa S, Vyas N. Malarial trend in Dakshina Kannada, Karnataka: An epidemiological assessment from 2004 to 2013. Indian J health sci (2015),8, 91-94.

[9] Bhatt S, Gething PW, Brady OJ, The global distribution and burden of dengue. Nature (2013).496, 504-507.

[10] George T, Jakribetta RP, Yesudhas S. Thaliath A, Pais MLJ, Abraham S, Baliga MS. Trend analysis of dengue in greater Mangalore region of Karnataka India: observations from a tertiary care hospital. International Journal of Applied Research (2018),4(6),92-96.

[11] Sundram BM, Raja DB, Mydin F, Yee TC, Raj K. Utilizing Artificial Intelligence as a Dengue Surveillance and prediction tool. J Appl Bioinformatics Comput Biol (2019), 8:1.

[12] Laureano-Rosario AE, Duncan AP, Mendez-Lazaro PA, Garcia-Rejon JE, Gomez-carro S, Farfan-Ale J, Savic DA, Muller-Karger FE. Application of Artificial Neural Networks for dengue fever outbreak predictions in the Northwest Coast of Yucatan, Mexico and San Juan, Puerto Rico. Trop. Med. Infect. Dist. (2018)3-5.

[13] Baruah J, Ananda S, Arun Kumar Incidence of dengue in a tertiary care centre-Kasturba Hospital, Manipal. Indian J Pathol Microbiol (2006).49(3),462-3.

[14] Pai Jakribettu R, Boloor R, Thaliath A, Yesudasan George S, George T, Ponadka Rai M et al. Correlation of Cinicohaematological parameters in pediatric Dengue: A retrospective study. J Trop Med (2015).6(47), 162.

[15] Damodar T, Dias M, Mani R, Shipla KA, anand AM, Ravi V et al. clinical and laboratory profile of dengue viral infections in and around Mangalore. Indian J Med Microbiol (2017), 35(2), 256-261.

[16]Symptoms of Dengue [Accessed on October 3rd 2020, https://www.cdc.gov/dengue/symptoms/index.html]

[17] ICD code for Dengue [Accessed on September 21st 2020, https://icd.codes/icd10cm/A90] [18]Symptoms of Malaria [Accessed on September 24th 2020, https://www.healthline.com/health/malaria#diagnosis].

[19] Jenni A. M. Sidey-Gibbons, Chris J. Sidey-Gibbons. Machine Learning in Medicine: a practical introduction. BMC Medical Research Methodology (2019), 19, 64.

[20] Beam A, Kohane I. Big Data and Machine Learning in Health Care. J Am Med Assoc. 2018; 319(13):1317–8.

INDEX