Detecting Dengue Disease Using Ensemble Classification Algorithms

S. Ruban, Naresha, and Sanjeev Rai

Abstract Health care has grown beyond imagination in the last few years with the impact of artificial intelligence. Artificial intelligence applications are used to solve many health issues in the society. However, developing these applications involves transforming the data from the original format to a format that is understandable by the system. It also involves using suitable algorithms appropriate for the problem to be solved. This work discusses an approach used to detect dengue. Performance evaluation was done with the real-time dataset. Few classification algorithms have been used over the dataset. To have a better accuracy, ensemble learning methods were used. Out of the three ensemble machine learning algorithms, like light gradient boost classifier, logistic regression, and support vector machine classifier that were used, the experimental study reveals that light gradient boost classifier gives a better accuracy of 94.47% compared with the other algorithms that are used.

Keywords Machine learning · Ensemble · Dengue · Vector-borne disease · Light gradient boost classifier · Logistic regression · Support vector machine classifier

1 Introduction

Artificial intelligence is transforming the health care like never before. From the collection of healthcare data, to processing and understanding the data, AI applications are playing a tremendous role. AI is all about developing machines or applications that can assist the human beings by emulating human intelligence at various roles and levels. A recent report published by the World Health Organization (WHO) [1] considers artificial intelligence as a technology that holds a great promise for transforming the health care globally. However, they insist on putting the sound ethical principles to guard and formulate its usage in the design, development, and

S. Ruban (✉) · Naresha
St Aloysius College (Autonomous), Mangalore, India
e-mail: rub2kin@gmail.com

S. Rai
Father Muller Medical College, Mangalore, India

S. Shukla et al. (eds.), *Data Science and Security*, Lecture Notes in Networks and Systems 462, https://doi.org/10.1007/978-981-19-2211-4_4

deployment of AI-based solutions. Another work that was published recently by Shneiderman [2] lists out the various conflicts that happen during the AI-enabled development, to address healthcare problems. This work emphasizes two important issues of AI, human emulation, and developing useful applications that could contribute to the development of the healthcare solutions. Though the objectives of AI research were set many years back, when Alan Turing asked this question, "can the machines think?" [3], the usage of AI in health care is very widely discussed and adapted in the recent times. Recent advances of using AI in health care have led to solutions that can predict the outcome of a procedure, assist in predicting emergencies like respiratory arrests and lung cancers so that the healthcare institutions can take better measures in providing the healthcare facilities to the patients. Artificial intelligence is a broader domain. Researchers have also studied the possibility of using AI to detect the epidemiological patterns that are the reasons for causing epidemics. Machine learning models can be built over the medical data and can be used to analyze vast amount of data to understand and detect the possible reason behind the epidemic. Similarly, the clinical notes that are available in the medical institutions can provide the base for developing noninvasive methods of diagnosing diseases earlier before the diagnostic tests could find out them.

Based on the data that are available with the World Health Organization, occurrence and spread of dengue are seen rising in many parts of the world. Accordingly, studies suggest around 50 million infections each year [4–6]. Early diagnosis of dengue fever can reduce this burden to a larger extent. Seeing the success of machine learning in different domains, researchers began to use machine learning techniques to develop tools that can assist clinicians diagnose illnesses at an early stage. Other benefits include saving the costs and the time taken in the diagnostic tests [7, 8]. Another work that was done [9] in this domain gives us a better result.

The next section deals with the existing work in this area; Sect. 3 describes the methodology that is adopted for this research study; Sect. 4 describes the results and discussion and finally the conclusion.

2 Literature Survey

Few of the earlier works in detecting dengue, by applying machine learning algorithms, points to the usage of artificial neural network (ANN) algorithm [10]. A recent work that was done by the researchers in Paraguay, also, points out the usage of ANN and SVM for finding dengue [11]. ML has been used to determine the most effective treatment [12], identify the patients at risk for a particular disease, suggesting treatment plans, and also used to predict a disease. Traditionally, predictive analytics was using just conventional logistic regression modeling. However, with machine learning models that are better at prediction [13], they help to uncover diseases and symptoms which are hiding in plain sight.

Vector-borne disease is an important public health problem in India, resulting in about 7 lakhs deaths annually. They are infections transmitted by the bite of infected

mosquitoes or other flies. Karnataka state, with good coastal region and irrigation, facilitates growth of mosquitoes that yields to high transmission of dengue [14]. AI and big data technology can help to understand the disease outbreaks [15]. It can also be used for building predictive model [16] and also help to correlate with other factors like climate and rainfall. One of the work that was done on dengue was done in the Thiruvananthapuram district, Kerala [17].

Another experimental study involving Bayesian network was done in Malaysia [18]. This dengue surveillance tool was implemented in the state of Penang. It incorporates features to enter data related to the dengue outbreaks and uses the geographical location to track and predict the outbreak earlier. The authors of this work claim an accuracy of 81.08% and also have predicted 37 outbreaks that have happened in the region, a month in advance. This study has helped to validate the claim of using machine learning as a tool for real-time surveillance.

Another experimental study involving support vector regression algorithm [19] was done by collecting data from the province of Guangdong in the country of china. The authors collected the meteorological data from 2011 to 2014 and developed a model to predict the occurrence of data in the locality. The authors also studied the usage of different machine learning algorithms for their study, however decided to use SVR algorithm, since it gives the optimal performance. The authors claim this experimental study finding could help the government and other people relevant to public health to respond early to the dengue outbreak.

Similar experimental study was done in Manila, Philippines [20]. The authors captured various data variables related to the meteorological factors such as direction and the speed of the breeze, the temperature, and humidity. Four years' data were gathered and used for the study. The authors used various modeling techniques such as random forest, gradient boosting, and general additive modeling. Though they concluded stating that every modeling technique is able to predict the outcome, random forest performed well for the given dataset.

This study is aimed at deriving insights from the existing hospital data of the Father Muller Hospital in dealing with dengue. Performance evaluation was done with the real-time dataset available for time period from 2015 to 2018. Few classification algorithms, have been used over the dataset. To have a better accuracy, ensemble learning methods were used. Out of the three ensemble machine learning algorithms, like light gradient boost classifier, logistic regression, and support vector machine classifier that were used, this experimental study reveals that light gradient boost classifier performs better with an accuracy of 94.47% compared with the other algorithms that were used.

3 Methodology

A medical institution holds a huge repository of health data (Textual and Image) that have remained underutilized over the period of time. Machine learning and data analytics provide a path of transforming this data into a huge wealth by discovering

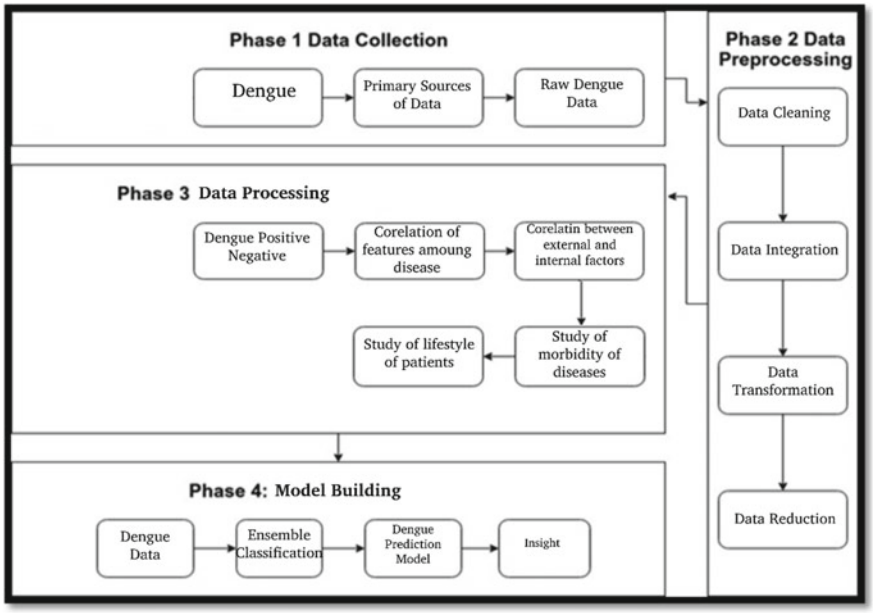

Fig. 1 Methodology of dengue fever classification using ensemble methods

patterns. The workflow and the methodology followed in this experimental study is elaborated and is represented in Fig. 1. Different data sources from where the real-time data were collected, the format of dengue data, data gathering process, data preprocessing, data processing, and ensemble classifiers are elaborated below.

3.1 Data Collection

Dengue is one of the transcendent vector-borne maladies universally, caused by the mosquitos. A person enduring this illness has mostly fever with other symptoms [21]. Indications ordinarily last about two to seven days. Normally, individuals start recouping after seven days. The personal clinical notes of each patient treated within the stipulated time interval were taken from the Department of Patient Registration and Clinical Records section (MRD). The clinical records are arranged based on ICD [22].

3.2 Data Preprocessing

The collected data were raw, and most of the portion were handwritten. Most of these handwritten portion of the clinical notes were written by various doctors and nurses who were attending to the patient who was undergoing treatment. Only, the discharge summary which was part of the clinical notes is the typed one. Since the quality of data is important, missing entries, inconsistencies, typographical, and semantic errors that were there in the raw data were clarified and rectified based on the discussion with the healthcare professionals who were assigned for that. This step does not give any meaningful insight but helps to find out the right assumption, that has to be made for the analysis and the features that has to be extracted. Tesseract optical character recognition engine [23] (OCR) was used for extracting the raw data from the scanned images that were stored in the clinical records. However, the accuracy of data that were extracted was moderate depending upon the clarity of the images, blur, and noise that were affecting the quality of the images. So, the data that were extracted had to be manually checked by the healthcare professionals. It was followed by pattern identification.

3.3 Data Processing

Each image which contained raw data of the dengue patient was captured. For extricating information, Python-tesseract was used. It has the capability of recognizing and reading the text embedded in images. Textual content was generated from images. To identify the features from the clinical notes, a dictionary was prepared, which was finalized after consulting the physician in the hospital. The list of the symptoms was looked in the clinical records. The entire process of the data processing is elaborated in the research work that has been published earlier [24] by the authors. Similar kind of work was also done for another common infectious disease called malaria [25]. The data dictionary is listed in Fig. 2.

3.4 Model Building

Machine learning techniques are basically algorithms that try to find out the relationship between different features that are found in the dataset. A machine learning model that produces discrete categories [26] are called as classification algorithm. Few of the case studies to understand classification algorithm include, those which can predict whether a patient has malaria or not, whether a tumor is benign or malign etc. In medicine, such kind of classification problems do exist, and classification algorithms are used in those areas. In this research work, we have used few algorithms such as light gradient boost classifier, logistic regression, and SVM

Diagnosis Discharged Improved	Joint Pain	Diagnosis Dengue
Burning Micturition	Fever, Cough	Vomiting
Cold, Chills	Breath Per Minute	Loose Stools
Headache, Nausea	Pallor, Clubbing	Abdominal Pain
Decreased Appetite, Diet	Diabetes, Hypertension	Sleep, Asthma
Bowel, Bladder	Icterus, Cyanosis	Lymphadenopathy
Temperature, IHD	TB, Malaria	Heart Beat Per Minute

Fig. 2 Snapshot of the data dictionary created for dengue data processing

classifier. One of the supervised machine learning algorithm is LR. The outcome obtained either be Yes or No, 0 or 1, true or False, etc. [27]. Another popular classification algorithm is support vector machine [28]. The intention of the SVM is to set rules to create the satisfactory line or selection boundary which can segregate. This best decision boundary is known as a hyperplane. Similarly, LGB machine learning algorithm [29] is also one of the popular ensemble-based classification algorithms.

4 Results and Discussion

The real-time data were subjected to analysis after the preprocessing was completed. The analysis was done based on the data dictionary that was developed. Few of the insights that were generated are presented below from Figs. 3, 4, 5, 6, and 7.

The model was developed using the three classification algorithms such as Light gradient boost classifier, Logistic regression, and Support vector machine. Different metrics for evaluating the machine learning models were generated and are listed below. The light gradient boost classifier (LGBM) gives the best accuracy of 94.47%, followed by the support vector machine (SVM) at 91.77% and the logistic regression at 86.15%.

5 Conclusion

This research work explores the possibility of a noninvasive method of identifying dengue from the symptoms. Before the lab results could confirm the illness, preliminary treatment can be started to avoid the adverse effects of dengue. More data from

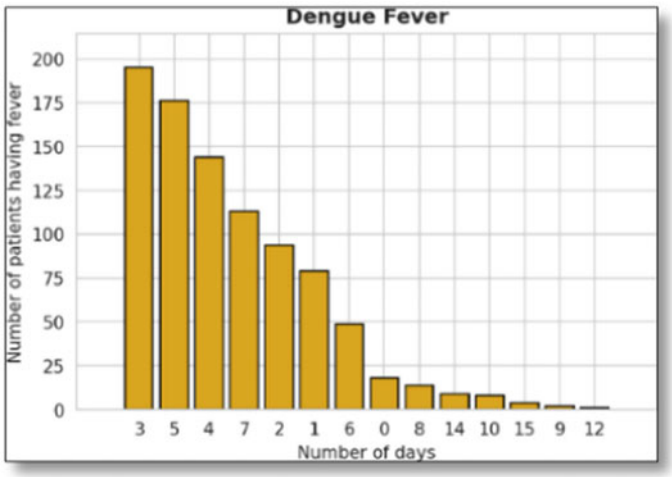

Fig. 3 Dengue cases recorded from 2015 to 2018, who showed the symptoms of fever

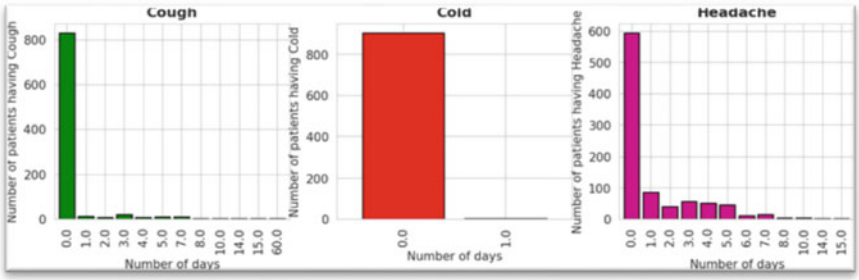

Fig. 4 Patients who displayed the symptoms of cough, cold, and headache

Fig. 5 Patients who displayed the symptoms of body ache, joint pain, and burning micturition

Fig. 6 Patients who displayed the symptoms of vomiting, chills, and loose stools

Model Name	Train Accuracy	Test Accuracy	Precision	Recall	F1-Score
Support Vector Machine(SVM)	92.42	91.77	1.00	0.85	0.92
Logistic Regression	88.40	86.15	1.00	0.76	0.86
Light Gradient Boost Classifier(LGBM)	94.11	94.47	1.00	0.90	0.94

Fig. 7 Performance comparison of the classification algorithms

various other hospitals could facilitate better results. Though traditional classification algorithms could be used to solve this problem, ensemble methods prove to be much effective than the other methods.

Acknowledgements Authors acknowledge that this work was carried out in the Big Data Analytics Lab funded by VGST, Govt. of Karnataka, under K-FIST(L2)-545, and the data were collected from Father Muller Medical College, protocol no: 126/19 (FMMCIEC/CCM/149/2019).

References

1. WHO report on AI. https://www.who.int/news/item/28-06-2021-who-issues-first-global-report-on-ai-in-health-and-six-guiding-principles-for-its-design-and-use. Accessed on 28 Oct 2021
2. Shneiderman B (2020) Design lessons from AI's two grand goals: human emulation and useful applications. IEEE Trans. Technol. Soc. 1(2):73–82
3. Turing AM (1950) Computing machinery and intelligence. Mind 49:433–460
4. WHO (1999) Strengthening implementation of the global strategy for dengue fever/dengue haemorrhagic fever prevention and control. Report of the informal consultation. Geneva, Switzerland
5. San Martın JL, Solórzano JO et al (2010) Epidemiology of dengue in the Americas over the last three decades: a worrisome reality. Am J Trop Med Hyg 82(1):128–135

6. Shepard DS, Undurraga EA, Betancourt-Cravioto M et al (2014) Approaches to refining estimates of global burden and economics of dengue. PLoS Neglected Tropical Diseases 8(11)
7. Jain A (2015) Machine learning techniques for medical diagnosis: a review. In: Conference center, New Delhi, India
8. Kononenko I (2001) Machine learning for medical diagnosis: history, state of the art and perspective. Artif Intell Med 23(1):89–109
9. Raval D, Bhatt D, Kumhar MK, Parikh V, Vyas D (2016) Medical diagnosis system using machine learning. Int J Comput Sci Commun 7(1):177–182
10. Ibrahim F, Taib MN, Abas WABW, Guan CC, Sulaiman S (2005) A novel dengue fever (DF) and dengue haemorrhagic fever (DHF) analysis using artificial neural network (ANN). Comput Methods Programs Biomed 79(3):273–281
11. Mello-Roman JD et al (2019) Predictive models for the medical diagnosis of dengue: a case study in Paraguay. Comput Math Methods Med 1–9
12. Obermeyer Z, Emanuel EJ (2016) Predicting the future-big data, machine learning and clinical medicine. N Engl J Medicine 375:1216–1219
13. Cuddeback J (2017) Using big data to find hypertension patients hiding in plain sight. AMGA Analytics
14. Arali PK et al (2019) Assessment of national vector borne disease control programme in state of Karnataka. Int J Community Med Public Health 6(2):525–532
15. Wong ZSY et al (2019) Artificial Intelligence for infectious disease big data analytics. Inf Disease Health 24:44–48
16. Guo J, Li B (2018) The application of medical artificial intelligence technology in rural areas of developing countries. Health Equity 2(1)
17. Valson JS, Soman B (2017) Spatiotemporal clustering of dengue cases in Thiruvananthapuram district, Kerala. Indian J Public Health 61:74–80
18. Sundram BM, Raja DB, Mydin F, Yee TC, Raj K (2019) Utilizing artificial intelligence as a dengue surveillance and prediction tool. J Appl Bioinformatics Comput Biol 8
19. Guo P, Liu T, Zhang Q et al (2017) Developing a dengue forecast model using machine learning: a case study in China. PLoS Neglected Tropical Diseases 11(10)
20. Carvajal TM, Viacrusis KM, Hernandez LFT, Ho HT, Amalin DM, Watanabe K (2018) Machine learning methods reveal the temporal pattern of dengue incidence using meteorological factors in Metropolitan Manila, Philippines. BMC Inf Diseases 18(1):183
21. Symptoms of Dengue. https://www.cdc.gov/dengue/symptoms. Accessed on 3 Oct 2020
22. ICD code for Dengue. https://icd.codes/icd10cm/A90. Accessed on 21 Sept2020
23. Smith R (2007) An overview of the Tesseract OCR engine. In: Proceedings of ninth international conference on document analysis and recognition (ICDAR), IEEE computer society, pp 629–633
24. Ruban S, Rai S (2021) Enabling data to develop an AI-based application for detecting malaria and dengue. In: Tanwar P, Kumar P, Rawat S, Mohammadian M, Ahmad S (eds) Computational intelligence and predictive analysis for medical science: a pragmatic approach, De Gruyter, Berlin, Boston, pp 115–138
25. Ruban S, Naresh A, Rai S (2021) A noninvasive model to detect malaria based on symptoms using machine learning. In: Advances in parallel computing technologies and applications, IOS Press, pp 23–30
26. Gibbons S, Gibbons S (2019) Machine learning in medicine: a practical introduction. BMC Med Res Methodol 19–64
27. Panagiotis Pintelas IEL (2020) Special issue on ensemble learning and applications. Editorial MDPI 4
28. Harimoorthy K, Thangavelu M (2020) Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system. J Amb Intell Human Comput 1
29. Ogunleye A, Wang QG (2019) XGBoost model for chronic kidney disease diagnosis. IEEE/ACM Trans Comput Biol Bioinf 17(6):2131–2140