2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N) | 978-1-6654-3811-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICAC3N53548.2021.9725539

A Machine Learning Approach for Predicting Dengue Outbreak

Lovjot Kaur
Department of Computer Scienceand Engineering
PEC University of Technology
Chandigarh, India
lovi.hari@yahoo.com

Rajesh Bhatia
Department of Computer Scienceand Engineering
PEC University of Technology
Chandigarh, India
rbhatiapatiala@gmail.com

Shweta Nagpal
Department of Computer Scienceand Engineering
PEC University of Technology
Chandigarh, India
nagpal.shweta6@gmail.com

***Abstract*—Dengue, well known as a deadly mosquito disease, is speedily spreading around the globe. The transmission cycle of Dengue virus is highly supported by the Temperature, Relative Humidity and Rainfall conditions as these factors promote its growth and habitat. Till date,there is no vaccination or medication available for Dengue fever. In lack of medical support, the only way to prevent Dengue is by controlling the spread of its vector. In this study, a Hidden Markov Model (HMM) is proposed for Dengue outbreak prediction taking into account the data set of previous 10 years. It elaborates the firm relationship of the Dengue outbreak and meteorological factors. The results obtained from the model predicts the outbreak with accuracy of approx- imately 93.38%.This model can be successfully implemented for control and prevention of dengue outbreak.**

Keywords—HMM; Dengue; Vector; Meteorological; Epidemio logical; Outbreak

I. INTRODUCTION

As known for a deadly arthropod-borne disease, Dengue, is spreading swiftly. Beginning from the 18th century, major epidemics started emerging in Africa and North America after the Second World War due to rapid increase in economic and urban growth. This epidemic has remained in Asia since the 20th century [27]. Another contributing factor that added to Dengue incidence is the international travel as it promotes the spread between population which evolves in making it an endemic disease in a large number of countries of the world. The incidence of its vector is also affected by the climatic conditions (i.e. Temperature, Relative Humidity, Rainfall etc.). Dengue has increased several folds in the last few years. According to the studies, it is estimated that nearly 50 to 100 million cases of Dengue fever (DF) and approximately 250000-500000 cases of Dengue Hemorrhagic Fever (DHF) arise every year [14]. Most fatal manifestations of DF are Dengue Shock Syndrome (DSS) and Dengue Hemorrhagic Fever (DHF) which arise with flu like symptoms as fever, rash, joint pain, and weakness, headache, nausea leading to shock or even circulatory failure [25]. The Dengue fever also arises severely with distinctive and expanded bounds of symptoms[11]and this flu like appearances are generally observed in adults and older children. It has an extended recovery and is pronounced with weakness and depression [3]. Presently there is no medical aid to cure dengue fever. However, if left untreated it can lead to death. Without medical support, the only way to prevent this disease is by controlling the spread of

vector. This is possible by predicting the outbreak beforehand; so that the government and individuals are left with sufficient time to take the adequate measures.

II. DENGUE EPIDEMIOLOGY

The Dengue epidemiology depends upon the following factors.

- **Agent factors-** The four virus serotypes responsible for Dengue fever are DENV-1, DENV-2, DENV-3 and DENV-4. They are antigenically similar hence, closely related virus serotypes. They are members of the Flavivirus genus in the family Flaviviridae. All the four viruses have divergent interactions with antibodies. The genome of Dengue virus consists of a single strand of RNA and it is directly translated into proteins.
- **Vector factors-** Aedes aegypti and Aedes albopictus have a high susceptibility to be infected by DENV and have elevated ability to replicate and transmit it. Density, behavior and vectorial capacity of vector influence the disease incidence.
- **Host factors-** Infants and elderly, pregnancy, obesity, asthma, peptic ulcers, menstruation, diabetes mellitus, high blood pressure are the risk factors.
- **Environmental factors**
  1) Temperature- between 16-30 degrees. With even 2 degree rise in temperature, the mosquito will bite more often due to dehy-dration and more blood thirst.
  2) Relative humidity- 60- 80 percent, rainfall increases breeding.
  3) Endophily- indoor habitats, grow around fresh water containers.

The Dengue epidemiology is illustrated in figure 1.

ISBN: 978-1-6654-3811-7/21/$31.00 ©2021 IEEE

An HMM model is proposed in this study to predict Dengue outbreak and to calculate its predictive power. It was used to classify and predict quite a few diseases in the past. Wu et al in [31] used HMM to detect the heart diseases using record of 325 heartbeats from ten patients. The HMM developed in this study comprised of four states and the model had an accuracy of 95%. In another study [21], HMM was used

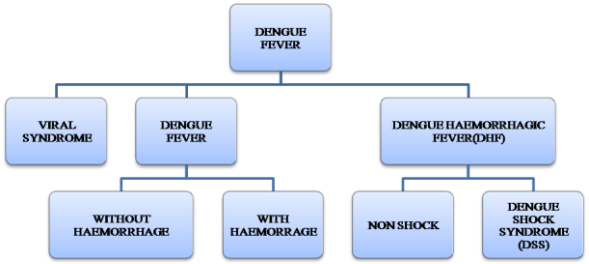

Fig. 1 Dengue Epidemiology

by Li et al to predict the occurrence of lung cancer. The data of 508 patients was collected from the hospitals and model proved to be a powerful prediction system. Further, Vimala et al in [29] also used HMM to classify ECG signals of patients suffering from stress and heart diseases. These studies confirmed that HMM as a statistical model uses learning and identification of relationships among various states, leading to classification and prediction.

III. RELATED WORK

The prediction of Dengue outbreak is the only way to take timely precautions to avoid the Dengue virus from infecting the human population. Various experiments have already been conducted which are primarily based on the factors that give better predictions, such as Dengue cases, weather factors, geographical locations and data of previous years which are taken into account to carry the research forward. The data is processed under the concept of data mining that selects models and inspects very large data with the motive to identify the flow and pattern which gives fair and useful analysis [7][12].

Previously various researchers have proposed the predictive models describing different frameworks, using different inputs and reflected relationship among those inputs [9][6][30][28][23][16]. In [15], Hii Y. L. et al have revealed the importance of weather variables for the progress of a less complex and precise Dengue early warning and also have elaborated the strong impact and synergy of the environment, the ecology, human and virus factors.

The different algorithms distinguish Dengue from various illnesses and predict the severe disease. Lukas Tanner et al in [26] have used C4.5 decision tree classifier for analyzing clinical, hematological data differentiating Dengue from other diseases. In [17], the author uses environmental factors as inputs for predicting the Dengue fever outbreak with artificial neural network. The system proposed serves the task for health department by intimating the Dengue outbreak along with the areas at risk beforehand. Yohanis Yusof et al in [32] presents a model for the Dengue outbreak prediction that associates least square support vector machine using Dengue cases and measure of rainfall level. In another research, ALV Gomes et al in [13] uses quantitative real-time PCR (The amplified DNA product) to measure the level of genes that are responsible for DENV innate immune responses and classify DHF and DF patients. The support vector machine is used in the research to examine the pattern of 12 genes in PBMCs (Peripheral Blood Mononuclear Cells) including 28 patients of Dengue. In [1] ,

the author performed the analysis using the neural network and compared its performance with the Support Vector Machine.

The appropriate combination of classifiers increases the pre- dictive power of the model. In [2], a model presents the use of multiple classifiers for a predictive model for the outbreak of Dengue. The various classifiers used are Decision Tree, Rough Set Classifiers, Associative Classifiers and Nave Bayes. The combination of all the classifiers gives more accuracy as compared to the accuracy of a single classifier.

The climatic factors have a vital impact on human health and the outbreak of disease. In case of Dengue outbreak a relation- ship is always observed between climate and Dengue vector

i.e. mosquito . The meteorological conditions as Temperature, Relative Humidity, Precipitation and Rainfall influence the development of vector with Dengue virus [24] . A study was carried to model Dengue Fever risk on mosquito count and clinically confirmed cases of Dengue fever by Khormi HM et al [19]. Kawinga H.W.B. et al in [18] described the model for prediction of Dengue outbreak using vector correction method occupying Relative Humidity and Temperature only.

The work done by Ramadona et al in [24] elaborates the effect of weather variables for developing an Early Warning System for Dengue incidence which is simple and cost effective. The model rely on the fact that Temperature and Rainfall have a vital effect on vectors and viruses of Dengue. In [5] ,Young Jo Choi et al have proposed a Negative Binomial model taking Mean, Average, Maximum Temperature and monthly Cumulative Rainfall values as the dataset and observes that the Temperature and the Rainfall specifically are linked with incidence of Dengue fever.

IV. MATERIALS AND METHODS

A. Study Site

The study is conducted in four districts of Punjab. Punjab lies in the subtropical region due to which vast variation of climate is observed throughout the year. The northeast location of the state results in extreme summers and winters with abundant rainfall during monsoons. The temperature during summer season goes up to 45◦ C (beginning from April) and drops to 0◦ C during winters (December to February). The rainy season begins in the month of July and lasts till September with high Relative Humidity. The four districts of Punjab, namely, Amritsar, Bathinda, Ludhiana and Patiala were selected as study area.

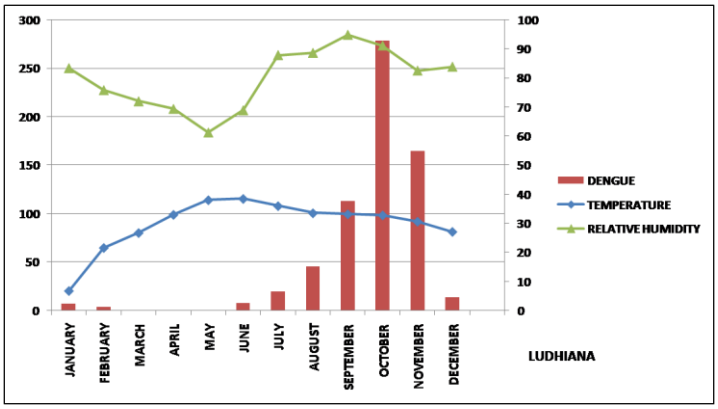

Fig. 2 Dengue Fever incidence corresponding to Temperature and Relative Humidity in Ludhiana

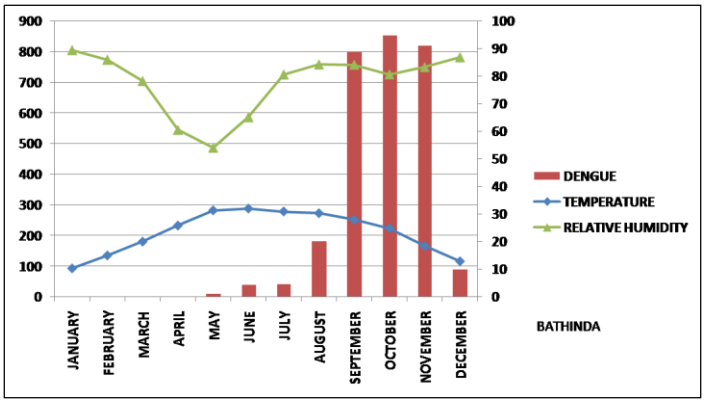

Fig. 3 Dengue Fever incidence corresponding to Temperature and Relative Humidity in Bathinda

B. *Data Sources*

This study includes the realistic data from the authenticated sources. It covers the meteorological data and Dengue incidence data from 2006 to 2015. The meteorological data was obtained from The Indian Meteorological Department, Pune (Maximum Temperature, Minimum Temperature, Mean Temperature, Maximum Relative Humidity, Minimum Relative Humidity, Mean Relative Humidity and Rainfall). The count of clinically confirmed Dengue cases was collected from The Department of Health and Family Welfare, Punjab.

C. *Data Analysis*

In this study, we used HMM to perform time series analysis on the collected meteorological and epidemiological data. Since Punjab experiences hot and cold season in extreme, therefore, choice of HMM was made as it has the ability to take account of the data in time sequential order.

D. *Data Preprocessing*

The data preprocessing enhances the performances of the

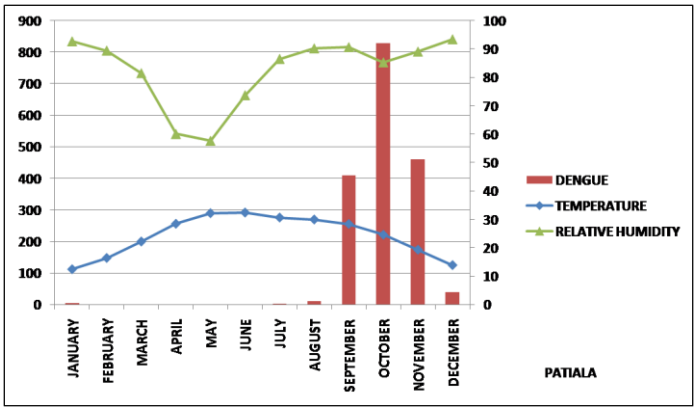

Fig. 4 Dengue Fever incidence corresponding to Temperature and Relative Humidity in Patiala

model as it reduces the time complexity and also the cost of the model. It also improves the accuracy and consistency of results generated by the model. The need for data preprocessing arises as the real data is generally incomplete with few missing values

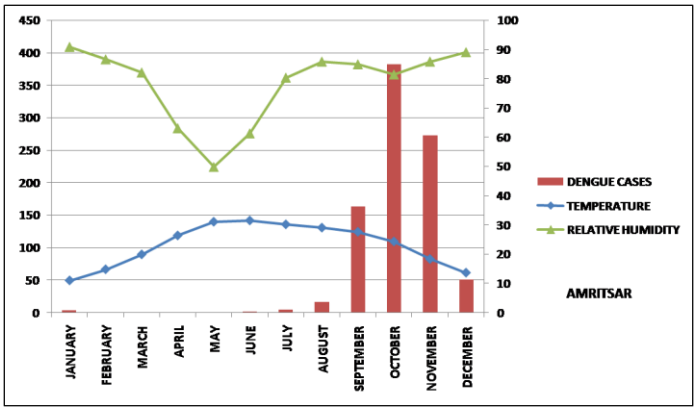

Fig. 5 Dengue Fever incidence corresponding to Temperature and Relative Humidity in Amritsar

and is inconsistent. The data was cleaned by replacing the missing values using the mean of the instances.

V. HIDDEN MARKOV MODEL FOR PREDICTION OF DENGUE OUTBREAK

Hidden Markov Model is the analytical method used for the modeling of the time series data. This model represents the probability distribution of the observed sequences. It consists of hidden states (invisible process) and observable symbols (visible process). The hidden states combine to form a set of Markov Chain. The basic structure of Hidden Markov Model is seen in figure 6. The Hidden Markov Model have a finite number of states. Depending upon the previous state, a new state is generated according to the transition probability distribution. An observation output symbol is made after each

A. EXPERIMENTAL SETUP

The data used in this study for the prediction of Dengue outbreak includes the meteorological and epidemiological data.

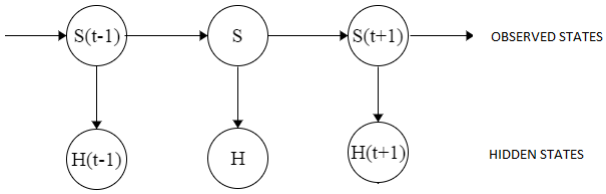

Fig. 6 Hidden Markov Model

The attributes given as meteorological data are Maximum Temperature, Minimum Temperature, Mean Temperature, Maximum Relative Humidity, Minimum Relative Humidity, Mean Relative Humidity, Rainfall and the count of Dengue cases as epidemiological data. The output of the model gave the extent of the outbreak as No Outbreak, Low, Mid, High respectively. For the proposed model the number of states (N=4) are No Outbreak, Low, Mid and High respectively. These states were connected with one another in such a way that every state could be outreached from any of the states. Further, the number of observations in each state were

Where $1 \leq$ and $j \leq N$

Succeeding a transition probability matrix was computed as T=

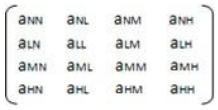

Fig. 7 Matrix of Transition Probability

In addition, the probability distribution of observations was determined for each state and a matrix was constructed for the same. In this model Viterbi Algorithm [8] was used to computethe most appropriate state sequence. Figure 8 shows the flowof Research Methodology.

VI. RESULTS

On providing the model with the relevant inputs the accuracy obtained for the four districts is 97% for Bathinda, 92% for Patiala, 97% for Ludhiana and 86% for Amritsar. The results reflect a proficient accuracy by the model and also an interesting relationship was found between the Dengue incidence and the weather factors.The accuracy depicts the correctly classified instances from the entire input of instances given to the model as shown in

The data was preprocessed to replace the missing values and for normalization. The attributes given as

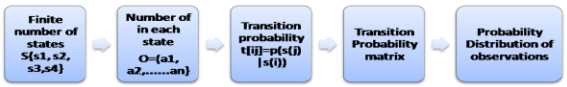

Fig. 8 Flow Diagram of Research Methodology

computed and the observations considered in this study were the meteorological factors (Temperature, Relative Humidity and Rainfall). The monthly mean computed from daily data was assigned to each observation. Subsequently, the transition probability depicted the probability of change of one state to another corresponding to the number of Dengue cases i.e. the transition from state Low to state Mid. The equation for Transition Probability is shown in equation 1.

$$t(ij) = p(s(j)|s(i)$$

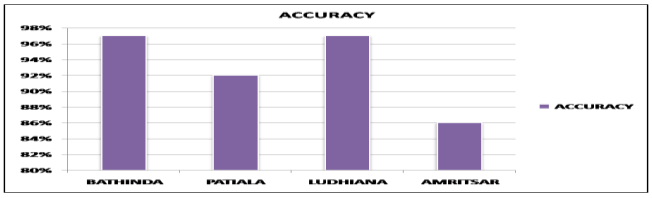

Fig. 9 Accuracy Distribution

equation 2.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN)} \quad (2)$$

The accuracy calculated for the four districts is as in figure 9. The results also include the True Positive Rate (TPR) and False Positive Rate (FPR) respectively as these compute the performance of the model. The TPR is represented as equation 3.

$$TPR = \frac{TP}{(TP + FP)} \quad (3)$$

It is defined as the percentage of instances that are actually positive and are correctly identified to be positive. The FPR is represented as equation 4.

$$FPR = \frac{FP}{(FP + TN)} \quad (4)$$

It is defined as the percentage of instances that are actually negative and are incorrectly identified to be positive. The performance of Hidden Markov Model for the four Districts is shown in Table I. The performance of Hidden Markov Model is compared with the algorithms namely Nave Bayes [22]Multilayer

TABLE I. THE PERFORMANCE PARAMETERS

	TRUE POSITIVE RATE	FALSE POSITIVE RATE
BATHINDA	0.9575	0.0255
LUDHIANA	0.956	0.021
AMRITSAR	0.736	0.07
PATIALA	0.754	0.04

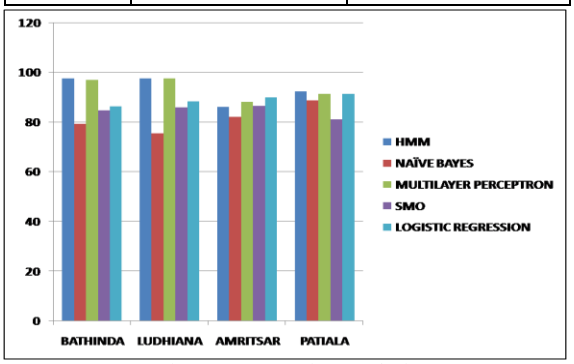

Fig. 10 Comparative Analysis

Perceptron [10], Sequential minimal optimization(SMO) [4] and LogisticRegression [20]. It was observed that the proposed model gave very promising results as displayed in figure 10.

VII CONCLUSION

The expeditious increase in Dengue outbreak is a serious concern for society as well as the Healthcare Department. Use of effective and efficient methodology enables to take preventive measures by predicting the outbreak beforehand. Use of the data of previous years enables to have more precise and perfect results. In this study, the Hidden Markov model uses time series analysis to predict the outbreak of Dengue from the epidemiological and meteorological data which proves to be an efficient approach for Dengue outbreak prediction. The influence of meteorological variables on the incidence of Dengue was determined. Amid various factors which signify the outbreak of Dengue, the study suggests that the Temperature, Relative Humidity and Rainfall play an important role. This approach can be successfully used to predict the outbreak beforehand so that timely measures are taken to control the Dengue vector.

REFERENCES

[1] Mohammadreza Afshin. Application of least squares support vector ma- chines in medium-term load forecasting. *Canada: Ryerson University (Canada)*, page 46, 2007.

[2] Azuraliza Abu Bakar, Zuriyah Kefli, Salwani Abdullah, and Mazrura Sahani. Predictive models for dengue outbreak using multiple rulebase classifiers. In *Electrical Engineering and Informatics (ICEEI), 2011 International Conference on*, pages 1–6. IEEE, 2011.

[3] Anna L Buczak, Phillip T Koshute, Steven M Babin, Brian H Feighner, and Sheryl H Lewis. A data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data. *BMC medical informatics and decision making*, 12(1):124, 2012.

[4] Li Juan Cao, S Sathiya Keerthi, Chong Jin Ong, Jian Qiu Zhang, Uvaraj Periyathamby, Xiu Ju Fu, and HP Lee. Parallel sequential minimal optimization for the training of support vector machines. *IEEE Trans. Neural Networks*, 17(4):1039–1049, 2006.

[5] Youngjo Choi, Choon Siang Tang, Lachlan McIver, Masahiro Hashizume, Vibol Chan, Rabindra Romauld Abeyasinghe, Steven Id- dings, and Rekol Huy. Effects of weather factors on dengue fever incidence and implications for interventions in cambodia. *BMC public health*, 16(1):241, 2016.

[6] Zamil MAH Choudhury, Shahera Banu, and Amirul M Islam. Fore- casting dengue incidence in dhaka, bangladesh: A time series analysis. 2008.

[7] A Shameem Fathima, D Manimegalai, and Nisar Hundewale. A review of data mining classification techniques applied for diagnosis and prognosis of the arbovirus-dengue. *IJCSI International Journal of Computer Science Issues*, 8(6), 2011.

[8] G David Forney. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973.

[9] DO Fuller, A Troyo, and John C Beier. El nino southern oscillationand vegetation dynamics as predictors of dengue fever cases in costa rica. *Environmental Research Letters*, 4(1):014011, 2009.

[10] Matt W Gardner and SR Dorling. Artificial neural networks (the mul- tilayer perceptron)a review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14):2627–2636, 1998.

[11] Robert V Gibbons and David W Vaughn. Dengue: an escalating problem. *BMJ: British Medical Journal*, 324(7353):1563, 2002.

[12] Paolo Giudici. *Applied data mining: statistical methods for business and industry*. John Wiley & Sons, 2005.

[13] Ana Lisa V Gomes, Lawrence JK Wee, Asif M Khan, Laura HVG Gil, Ernesto TA Marques Jr, Carlos E Calzavara-Silva, and Tin Wee Tan. Classification of dengue fever patients based on gene expression data using support vector machines. *PloS one*, 5(6):e11267, 2010.

[14] G Guzman, Gustavo Kouri, et al. Dengue and dengue hemorrhagic feverin the americas: lessons and challenges. *Journal of Clinical Virology*, 27(1):1–13, 2003.

[15] Yien Ling Hii, Huaiping Zhu, Nawi Ng, Lee Ching Ng, and Joacim Rocklöv. Forecast of dengue incidence using temperature and rainfall. *PLoS Negl Trop Dis*, 6(11):e1908, 2012.

[16] Nor Azura Husin, Naomie Salim, et al. Modeling of dengue outbreak prediction in malaysia: a comparison of neural network and nonlinear regression model. In *Information Technology, 2008. ITSim 2008. International Symposium on*, volume 3, pages 1–4. IEEE, 2008.

[17] Seongtae Hwang, Denmar S Clarite, Frank I Elijorde, Bobby D Gerardo, and Yungcheol Byun. A web-based analysis for dengue tracking and prediction using artificial neural network. 2016.

[18] HWB Kavinga, DDM Jayasundara, and Dushantha NK Jayakody. A new dengue outbreak statistical model using the time series analysis. *European International Journal of Science and Technology*, 2(10):35– 52, 2013.

[19] Hassan M Khormi, Lalit Kumar, and Ramze A Elzahrany. Modeling spatio-temporal risk changes in the incidence of dengue fever in saudi arabia: a geographical information system case study. *Geospatial Health*, 6(1):77–84, 2011.

[20] Balaji Krishnapuram, Lawrence Carin, Mario AT Figueiredo, and Alexander J Hartemink. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE transactions on pattern analysis and machine intelligence*, 27(6):957–968, 2005.

[21] Hyo-Ki Lee, Jeon Lee, Hojoong Kim, Jin-Young Ha, and Kyoung-Joung Lee. Snoring detection using a piezo snoring sensor based on hidden markov models. *Physiological measurement*, 34(5):N41, 2013.

[22] Kevin P Murphy. Naive bayes classifiers. *University of British Columbia*, 2006.

[23] KPBP Raju and B Sokhi. Application of gis modeling for dengue fever prone area based on socio-cultural and environmental factors–a case study of delhi city zone. *Int Arch Photogramm Remote Sens Spat Inf Sci*, 37:165–170, 2008.

[24] Aditya Lia Ramadona, Lutfan Lazuardi, Yien Ling Hii, A˚ sa Holmer 11(3):e0152688, 2016.

[25] Vadrevu Sree Hari Rao and Mallenahalli Naresh Kumar. A new intelligence-based approach for computer-aided diagnosis of dengue fever. *IEEE transactions on Information Technology in Biomedicine*, 16(1):112–118,

[26] Lukas Tanner, Mark Schreiber, Jenny GH Low, Adrian Ong, Thomas Tolfvenstam, Yee Ling Lai, Lee Ching Ng, Yee Sin Leo, Le Thi Puong, Subhash G Vasudevan, et al. Decision tree algorithms predict the diagnosis and outcome of dengue fever in the early phase of illness. *PLoS Negl Trop Dis*, 2(3):e196, 2008.

[27] Stephen J Thomas, Daniel Strickman, and David W Vaughn. Dengue epidemiology: virus epidemiology, ecology, and emergence. *Advances in virus research*, 61:235, 2003.

[28] K Ungchusak and DS Burke. Travelling waves in the occurrence of dengue hemorrhagic fever in thailand. *Nature*, 427:344347Cushing, 2004.

[29] K Vimala. Stress causing arrhythmia detection from ecg signal using hmm. *Stress. IJARCCE*, 2(10):6079–6085, 2014.

[30] S Wongkoon, M Pollar, M Jaroensutasinee, and K Jaroensutasinee. Predicting dhf incidence in northern thailand using time series analysis technique. *International Journal of Biological and Medical Sciences*, 4(3):117–121, 2009.

[31] Hang Wu, Sahong Kim, and Keunsung Bae. Hidden markov model with heart sound signals for identification of heart diseases. In *Proceedings of 20th International Congress on Acoustics (ICA), Sydney, Australia*, pages 23–27, 2010.

[32] Yuhanis Yusof and Zuriani Mustaffa. Dengue outbreak prediction: A least squares support vector machines approach. *International Journal of Computer Theory and Engineering*, 3(4):489, 2011.