Dengue Recognition System for Srilanka

By using machine learning

Name - Savindu Mahasen Ruhunuhewa

Cardiff ID - st20212770

ICBT ID - KD/BSCSE-CMU/03/02

**Introduction**

In srilanka there are lot of viruses are spread. For an example dengue, malaria, HIV, etc. Among these viruses dengue is most threaten virus.

**Problem statement**

* Based on symptoms,machine learning existing applications or systems are fully created for health care professionals and doctors. At that moment there is no specific machine learning application or system is available for patient to detect the dengue by their own device.
* Lot of urban areas don’t have laboratory facilities for earlier detecting of dengue. Based on that urban area people need to go long distance for meet the doctor or health care professional to check the dengue or not. So at that time Dengue fever can be highly increase for them.

**Literature Review**

**Introduction**

In now a days dengue is a common health issue in the world. Because dengue is spread though the word faster. Especially in srilanka. For an example in srilanka in may 2023 61,361 total dengue patients are reported(Sri Lanka: Dengue Outbreak - May 2023). Earlier days lot of laboratory testing methods were doing for detecting the dengue. For example Reverse transcriptase-polymerase chain reaction (RT-PCR), viral isolation and identification, nuclear acid sequence-based amplification (NASBA), IgM and IgG seroconversion, PCR testing are used. But these tests have several disadvantages. for example cost, complexity, and takes lot of time to get the result etc. Based on these disadvantages different systems, and machine learning techniques already created for dengue detection. But now a days most of these created machine learning systems or applications created for assisting to doctors and healthcare professional to identify the patient dengue. Which means there is no specifically created the any particular machine learning system or application for patient to dengue identify. In this literature review mainly focused on what are the already existing application, systems by using machine learning techniques and their problems. And also consider the other normal application, or systems.And their problems.

**Problems**

**Dengue prediction and diagnosis using machine learning techniques**

According to this “A dengue disease prediction and diagnosis model using sentiment analysis and machine learning algorithms” research paper, for the predicting the dengue KNN classifier, decision tree, random forest, Gaussian naive Bayes, and support vector classifier (SVC)(Gupta et al., DDPM: A dengue disease prediction and diagnosis model using sentiment analysis and machine learning algorithms 2023).

**The main limitations of this application**

Unbalance data set - when it comes to the dengue dataset. Where that contains positive cases of dengue may be rare compare to the negative cases.

Hyper parameter tuning problem - when using this models for increasing the model accuracy need to take lot of time to find out the best parameters.

Scalability - when using this KNN classifier, random, SVC are high expensive when we are dealing with larger amount of dataset (Jamwal & Bhatia, Prevalence of vector borne diseases in Jammu division, Jammu and Kashmir, India 2021).

**Classification of dengue application using machine learning techniques**

According to the “Classification of dengue using machine learning techniques” research paper, Simple Cart, C-4.5, Multi-layer perception algorithms used for classifying the dengue is infected or not (Sajana et al., Classification of dengue using Machine Learning Techniques 2018).

The main limitations of this models

* Overfitting - These simple cart, C-4.5, willing to overfitting, when trees becomes to the more deep, and complex. But this leads to generate the inaccurate predictions.
* Sensitive to small changes - these simple cart, C-4.5 algorithms highly respond to the training dataset. That leading to the different tree structure and potentially different predictive performance.
* Imbalance dataset - when it comes to the dengue dataset. Where that contains positive cases of dengue may be rare compare to the negative cases.
* Privacy and security problems - extracting the human sensitive information raise some privacy problems.
* Model accuracy - According to the “Dengue outbreaks prediction in bangaladesh perspective using distinct multi-layer perceptron NN and decision tree” research paper. Predicting accuracy of Multi-layer perceptron 68.5%. which means it not sufficient for using the predicting purpose (Khan et al., Dengue outbreaks prediction in Bangladesh perspective using distinct multilayer Perceptron NN and decision tree 2022).

**Presumptive diagnosis System for dengue fever by using machine learning**

Presumptive diagnosis system used the Decision Tree, Random Forest, Naive Bayes algorithms used for predicting the dengue fever. What ever this Random Forest , Naive Bayes, Decision Tree algorithms use for early detection of dengue in this system (Khan & Raza, Development and evaluation of a predictive diagnostic system for dengue fever using Machine Learning Techniques 2023).

The main Limitation of this system

* Less number of data points are available
* Accuracy of this models are very low
* Sensitivity - presumptive diagnosis system is lack with sensitivity for early stage of detecting (Khan & Raza, Development and evaluation of a predictive diagnostic system for dengue fever using Machine Learning Techniques 2023).

**Detection of dengue disease by using fused machine learning**

In this case use the PFDM model. In this PFDM model use two main ML-based procedures. For an example SVM and ANN. And this PFDM has two basic components. the training layer and the testing layer. In the training layer,

It has five steps (collecting, preprocessing, classification, efficiency, and machine-level combination). If the model fails to meet learning criteria, it is reassigned until satisfactory results are achieved. The outputs of ANN and SVM are then fused, and the trained model is stored in a cloud system. In the testing layer, data is obtained from a database, and preprocessed trained models are loaded from the cloud. The fused model is used to predict whether a Dengue diagnosis is progressive or destructive, with the predicted outcome compared to the actual result to measure accuracy (Al Nasar et al., Detection of dengue disease empowered with fused machine learning 2022).

The main limitation

* One important factor in classification is class imbalance, and classification efficiency cannot be entirely determined by accuracy rates alone (Al Nasar et al., Detection of dengue disease empowered with fused machine learning 2022).
* Feature engineering - SVM(Support vector machine) is heavily depend on feature engineering to identify the relevant patterns (Hoyos et al., Dengue models based on machine learning techniques: A systematic literature review 2021).
* Complexity and overfitting - when dealing with the noisy and small datasets. Overfitting can generate unreliable predictions (Hoyos et al., Dengue models based on machine learning techniques: A systematic literature review 2021).
* Computational resources - when using ANN it required the computational resources including high performance GPU, and memory (Chakraborty et al., Forecasting dengue epidemics using a hybrid methodology 2019).

**Decision support system for dengue detection**

The use of decision support systems (DSS) is essential for the early diagnosis of dengue. A viable method for creating DSS for dengue care is case-based reasoning (CBR), which uses clinical guidelines and historical cases to forecast a patient's present state based on their vital signs and symptoms (binti Mohd Zainee & Chellappan, A preliminary dengue fever prediction model based on vital signs and blood profile 2020).

Limitation of decision support system

* Imbalance dataset - which means dengue infected positive cases significantly are lower than negative cases.
* Data is limited - dengue dataset size is small .
* Complexity of system - Dengue has several symptoms. That symptoms can be overlapped with similar similar illness (Lopez et al., An intelligent decision support system to prevent and control of Dengue 2018).

**Solution**

According to the above mentioned machine learning systems, applications and normal systems like decision support system there are significant limitations are available. Base on that machine learning system, applications limitations are complexity of model and system, low accuracy of models, overfitting, unbalance dataset etc. And also above mention systems are mainly created for doctor and healthcare professional. Then some urban area people need to go long distance for meet the doctor or laboratories for detecting the dengue is infected or not. Based on that urban area people cannot be identify the dengue in earlier stage. These are the major problems existing in machine learning dengue systems and application. When it comes to the normal systems like decisions support system there is some issue like complexity of the system, data limitation, in balance dataset available.

Based on above mentioned machine learning systems, applications and normal systems like decision support systems problems or limitations, dengue detection system for patient is going to be implemented for patient by using decision tree, random forest, logistic regression algorithm as solution including the “grid\_search cv” , proper dataset and models parameter like random state.

The main reasons of this support vector machine model is used in dengue detection system are easily work with high dimensionality features, and memory efficient which means support vector machines don’t need all the data. It wants only near the decision boundary. And also support vector machines can be easily work with different data types by using the mathematical tricks called kernals(Nordin et al., The classification performance using support vector machine for endemic dengue cases 2020).

The reasons of decision tree algorithm used for dengue detection system are efficiency is very high. Which means by using this decision tree algorithm can be making the understandable decision. And also easy to apply this decision tree algorithm, and work with larger dataset (Jayasundara et al., Developing a decision support testing algorithm to detect severity level of Dengue 2017).

The reasons of random forest algorithm used for dengue detection system are efficiency is very high. Which means by using this random forest algorithm can be making the correct prediction. And also easy to apply this decision tree algorithm, and work with larger dataset (Katta et al., An efficient learning model selection for dengue detection 2022).

**Methodology**

**Data Collection**

For the dengue detection system based on symptoms by using machine learning technique, data sets are received by websites and hospitals, and other medical institutions. Combine these datasets together and created the full dataset.

**Data Cleaning**

Checking the null values and duplicate values in created full dataset. Null values are going to be filled by using mean value of that particular column. And duplicated values are removed from that full dataset.

**Splitting the dataset**

This dengue dataset is going to be divided into the two part by using “random splitting function” Training set and testing set. Training set size is 80%. Test set size is 20%. This training set is used for model training and testing is used for evaluate model performance.

**Models**

For classifying this dengue, support vector machine, decision tree algorithm, random forest algorithms are the models going to be used.

Support vector machine model - The main reasons of this support vector machine model is going to be used in dengue detection system are easily work with high dimensionality features, and memory efficient which means support vector machines don’t need all the data. It wants only near the decision boundary. And also support vector machines can be easily work with different data types by using the mathematical tricks called kernals (Nordin et al., The classification performance using support vector machine for endemic dengue cases 2020).

Decision tree model - The main reasons of this decision tree model is going to be used in this dengue detection system are this model can be easily used, higher accuracy models can be created, and higher predictions can be produced (Huynh-Cam et al., Using decision trees and random forest algorithms to predict and determine factors contributing to first-year university students’ learning performance 2021).

.

Random forest model -

The main reasons of this random forest is used in this dengue detection system are increase the model accuracy, reduce the overfitting level, and handling the larger dataset (George & Ganesan, Advanced cuttlefish optimizer-random decision forest (ACORDF) based design of fractional order PID controller for higher-order time-delay system 2021).

**Statistical Testing**

**Normality testing**

Normality testing means check whether data is normally distributed or not.

**Conclusion**

**References**

*Sri Lanka: Dengue Outbreak - May 2023* (no date) *ReliefWeb*. Available at: https://reliefweb.int/disaster/ep-2023-000084-lka (Accessed: 28 February 2024).

Pruszczyk, P. (2018) ‘Diagnosis: Clinical prediction rules and laboratory tests’, *ESC CardioMed*, pp. 2758–2761. doi:10.1093/med/9780198784906.003.0658.

Gupta, G. *et al.* (2023) ‘DDPM: A dengue disease prediction and diagnosis model using sentiment analysis and machine learning algorithms’, *Diagnostics*, 13(6), p. 1093. doi:10.3390/diagnostics13061093.

Jamwal, N. and Bhatia, S. (2021) ‘Prevalence of vector borne diseases in Jammu division, Jammu and Kashmir, India’, *Journal of Vector Borne Diseases*, 58(1), p. 28. doi:10.4103/0972-9062.321746.

Khan, Md.A. *et al.* (2022) ‘Dengue outbreaks prediction in Bangladesh perspective using distinct multilayer Perceptron NN and decision tree’, *Health Information Science and Systems*, 10(1). doi:10.1007/s13755-022-00202-x.

Sajana, T. *et al.* (2018) ‘Classification of dengue using Machine Learning Techniques’, *International Journal of Engineering &amp; Technology*, 7(2.32), p. 212. doi:10.14419/ijet.v7i2.32.15570.

Khan, J.R. and Raza, S.K. (2023) *Development and evaluation of a predictive diagnostic system for dengue fever using Machine Learning Techniques* [Preprint]. doi:10.21203/rs.3.rs-2473833/v2.

Al Nasar, M.R. *et al.* (2022) ‘Detection of dengue disease empowered with fused machine learning’, *2022 International Conference on Cyber Resilience (ICCR)* [Preprint]. doi:10.1109/iccr56254.2022.9996009.

Hoyos, W., Aguilar, J. and Toro, M. (2021) ‘Dengue models based on machine learning techniques: A systematic literature review’, *Artificial Intelligence in Medicine*, 119, p. 102157. doi:10.1016/j.artmed.2021.102157.

Chakraborty, T., Chattopadhyay, S. and Ghosh, I. (2019) ‘Forecasting dengue epidemics using a hybrid methodology’, *Physica A: Statistical Mechanics and its Applications*, 527, p. 121266. doi:10.1016/j.physa.2019.121266.

binti Mohd Zainee, N. and Chellappan, K. (2020) ‘A preliminary dengue fever prediction model based on vital signs and blood profile’, *2016 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES)* [Preprint]. doi:10.1109/iecbes.2016.7843530.

Lopez, D. *et al.* (2018) ‘An intelligent decision support system to prevent and control of Dengue’, *Journal of Ambient Intelligence and Humanized Computing* [Preprint]. doi:10.1007/s12652-018-0829-9.

Jayasundara, S.D.P., Perera, S.S.N. and Rathnayaka, N.S. (2017) ‘Developing a decision support testing algorithm to detect severity level of Dengue’, *European Scientific Journal*, 13(9). doi:10.19044/esj.2017.v13n9p137.

Rocha, F.P. and Giesbrecht, M. (2022) ‘Machine learning algorithms for dengue risk assessment: A case study for São Luís do Maranhão’, *Computational and Applied Mathematics*, 41(8). doi:10.1007/s40314-022-02101-z.

Katta, M. *et al.* (2022) ‘An efficient learning model selection for dengue detection’, *Smart Innovation, Systems and Technologies*, pp. 439–453. doi:10.1007/978-981-16-9873-6\_40.

Haque, M.A., Shivaprasad, G. and Guruprasad, G. (2021a) ‘Passenger Data Analysis of Titanic using machine learning approach in the context of chances of surviving the disaster’, *IOP Conference Series: Materials Science and Engineering*, 1065(1), p. 012042. doi:10.1088/1757-899x/1065/1/012042.

George, T. and Ganesan, V. (2021) ‘Advanced cuttlefish optimizer-random decision forest (ACORDF) based design of fractional order PID controller for higher-order time-delay system’, *Journal of Engineering, Design and Technology* [Preprint]. doi:10.1108/jedt-11-2020-0451.

Nordin, N.I. *et al.* (2020) ‘The classification performance using support vector machine for endemic dengue cases’, *Journal of Physics: Conference Series*, 1496(1), p. 012006. doi:10.1088/1742-6596/1496/1/012006.