Application of Machine Learning to Inform Clinical Management of Infectious Diseases in Vietnam

Dynamic Severe Dengue Predictors

Imperial College London

Thi Anh Tho Nguyen

1. Introduction

- Dengue is an infectious disease originally endemic in tropical countries such as South East Asia. Recently, Dengue was found in over 188 countries across different continents, putting 3 billion people at risk annually.
- If Dengue progress to Severe Dengue, the mortality rate can be up to 50%, and reduce to 1% if patients get proper treatment.
- Dengue places a heavy burden on the healthcare systems and economy of endemic countries. In Vietnam, every year 2 million people are infected and cost the economy $95 million annually.

2. Objective

Using clinical data collected over 10 years by OUCRU, a machine learning model is hoped to be developed, deployed and aid clinicians in diagnosing Severe Dengue and relieve the strain of overcrowding in hospital

3. Machine Learning Model

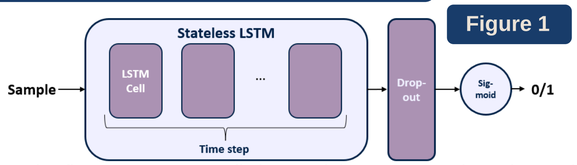
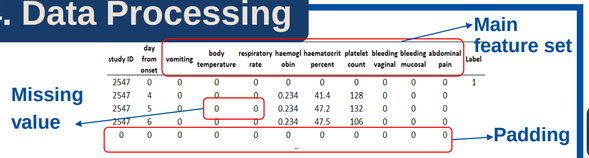


Figure 1

- The first Degue ML predictive model to use RNN based neural network architecture.
- LSTM is proven to work well with clinical data (*) through multiple research. It is capable of processing time-series data. And unlike traditional RNN, it is able to decide which information to withhold for long-term and which for short-term.
- Model is optimised with RMSProp algorithm which is good in preventing vanishing and exploding gradient in complex network

4. Data Processing



study ID	day from onset	vomiting	body temperature	respiratory rate	haemoglobin	haematocrit percent	platelet count	bleeding vaginal	bleeding mucosal	abdominal pain	Label
2547	0	0	0	0	0	0	0	0	0	0	1
2547	4	0	0	0	0.234	41.4	128	0	0	0	
2547	5	0	0	0	0.234	47.2	132	0	0	0	
2547	6	0	0	0	0.234	47.5	106	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	

1. Data Imputation: to simulate the reality that not all medical tests are done on a daily basis, missing values from data is not interpolated but only filled with 0 values. Since 0 does not associated to any meaningful value in any numerical variable, it should not hinder the learning ability of LSTM model (**)

2. Data formatting: Data is formatted to different type known as step-ahead format. 1-step-ahead format means the data is to be used to predict if Severe Dengue will happen in roughly 1 day. Each sequence is then padded to the decided common data-length so the training process would be easier.

3. Data Standardisation

5. Preliminary Results

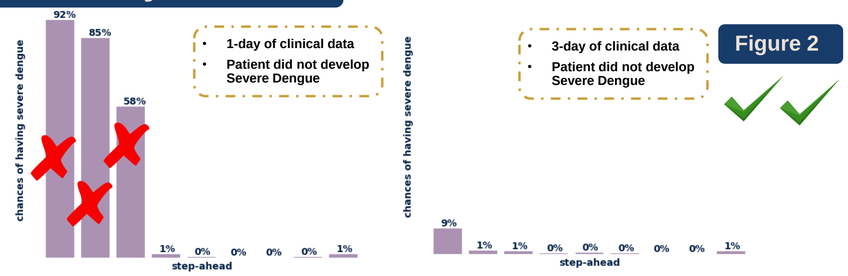


Figure 2

Simulation Result on Negative Sample from Patient 641 using Maximum Data Length

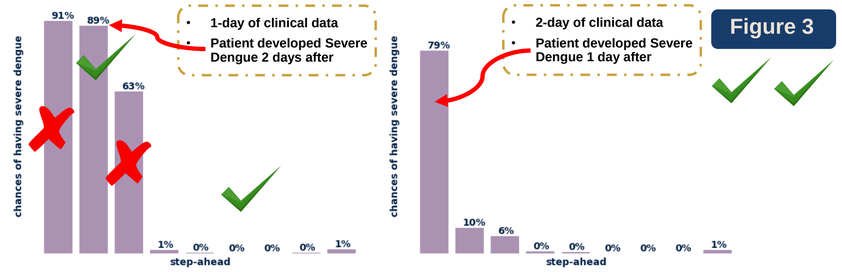


Figure 3

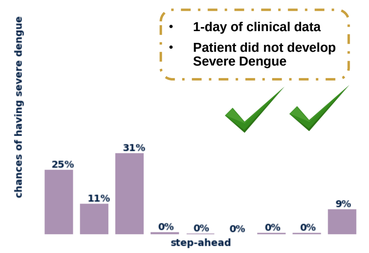
Simulation Result on Positive Sample from Patient 602 using Maximum Data Length

- The predictors started out with high performance in most of the metrics. However, since the amount of data decreases as the step-ahead increases, the model started to struggle significantly at 7-step-ahead format.
- Figure 2 & 3 show the simulation of all 9 predictors on ML web-based application when inputting 1-3 days worth of data into the simulation. Although the prediction results were not too far off, there is still some delay. The delay was hypothesized to be a result of data being formatted with maximum data length, which results in a typical/sufficient input sequence going into the model to be longer than what the clinicians or patients expect.
- Thus, the models were retrained with median data length, with results shows in Table 1.2

	Sensitivity	Specificity	PPV	NPV
1-step-ahead	96 ± 4	94 ± 3	95 ± 3	95 ± 6
2-step-ahead	98 ± 2	92 ± 3	91 ± 2	98 ± 2
3-step-ahead	92 ± 4	90 ± 4	85 ± 6	92 ± 3
4-step-ahead	84 ± 5	87 ± 3	81 ± 4	86 ± 4
5-step-ahead	85 ± 5	87 ± 5	80 ± 5	88 ± 3
6-step-ahead	82 ± 4	86 ± 3	78 ± 6	86 ± 3
7-step-ahead	73 ± 12	82 ± 4	70 ± 5	82 ± 6
8-step-ahead	75 ± 9	82 ± 5	65 ± 8	84 ± 9
9-step-ahead	53 ± 12	89 ± 1	77 ± 9	86 ± 3

Table 1.1: Performance of 9 Dyanmic Severe Dengue Predictor using Maximum-Data-Length

6. Updated Results



Simulation Result on Negative Sample from Patient 641 using Median DT

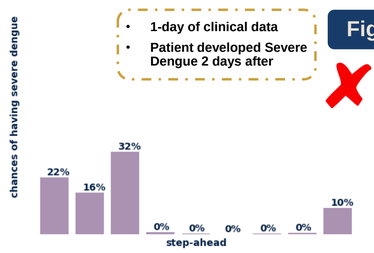


Figure 4

Simulation Result on Positive Sample from Patient 602 using Median DT

- Table 1.2 shows the improved performance of predictors when trained with median-data-length format. The PPV and NPV metrics overall appear to be more stable throughout 1-7 step-ahead prediction. However, at 1-step-ahead prediction, the NPV decreased to 80%, which can be explained by the highly skewed label distribution towards the positive labels. 8 and 9-step-ahead predictors suffer worst from imbalanced and lack of data since median-data-length format deletes some profiles according to their sequence length while maximum-data-length format does not
- Figure 4 shows the updated simulation of the deployment. There was significant increase in performance in the negative examples, where the patients were correctly diagnose to not develop Severe Dengue with just 1 day of clinical data. Unfortunately for the positive example, it performs worse on this updated model because of the decrease in NPV of 1-step-ahead model.

	Sensitivity	Specificity	PPV	NPV
1-step-ahead	97 ± 1	79 ± 10	97 ± 2	80 ± 5
2-step-ahead	97 ± 2	85 ± 4	95 ± 1	94 ± 4
3-step-ahead	97 ± 2	84 ± 8	92 ± 4	93 ± 4
4-step-ahead	97 ± 2	85 ± 7	90 ± 2	93 ± 5
5-step-ahead	99 ± 1	83 ± 6	86 ± 3	98 ± 2
6-step-ahead	97 ± 2	82 ± 4	85 ± 3	96 ± 3
7-step-ahead	88 ± 6	78 ± 4	81 ± 4	88 ± 5
8-step-ahead	67 ± 18	87 ± 1	66 ± 16	88 ± 4
9-step-ahead	15 ± 15	88 ± 3	26 ± 22	83 ± 3

Table 1.2: Performance of 9 Dyanmic Severe Dengue Predictor using Median-Data-Length

- Throughout many random simulations, it is observed that the NPV rate is higher than that of PPV. And although Figure 3 suggests that the model can narrow down roughly the exact date Severe Dengue would develop, other simulations suggest that it is only a rare occurrence. Generally, for positive samples, most of the predictors will show high chances of developing Severe Dengue.

7. Discussion

- From the performance results shown in Table 1.1 and 1.2, the predictors shows promising results in classifying Severe Dengue cases. With adequately high PPV and NPV, the deployed model is hoped to help clinicians in clinical management, relieve burden on hospital systems during an outbreak, and giving severely at-risk patients with proper treatment. This result demonstrates the ability of LSTM in learning the dependencies between variables, through times, even when the training data is missing up to 50% data points.
- Figure 2 shows a more desirable performance of the model where for a positive sample, the model was able to narrow down the date that Severe Dengue would develop. Since the current sequences of data are often filled with 0, it is suggested that the lack of data points (missing values) hinders the LSTM ability to expand on its hypotheses. However, during simulation, the models still exhibit performance consistent with test result shown in Table 1.1 and 1.2
- For future work, a model trained on imputed data set would be a good starting point. Other data formats which promote the reality of short input data being provided into the models should also be explored. A smaller set of features will also increase efficiency of the since they would require less amount of clinical test which may be costly and time consuming. Instead of binary classification, a regression model that predicts the day Severe Dengue is worth looking into. In this project, the regression model was attempted. However, date features were either missing a lot of data points or incomprehensibly recorded, which hinder the feasibility of developing and training the regression model.

(*) J. Xia, S. Pan, M. Zhu, G. Cai, M. Yan, Q. Su, J. Yan, and G. Ning, “A long shortterm memory ensemble approach for improving the outcome prediction in intensive care unit,” Computational and Mathematical Methods in Medicine, vol. 2019, pp. 1–10, 2019.

(*) T. Pham, T. Tran, D. Phung, and S. Venkatesh, “Predicting healthcare trajectories from medical records: A deep learning approach,” Journal of Biomedical Informatics, vol. 69, pp. 218–229, 2017.

(**) F. Chollet, Deep learning with Python. Manning, 2018

NPV: Negative Predictive Value

PPV: Positive Predictive Value

oucru VITAL wellcome