Utilização de técnicas de *Machine Learning* e de *Deep Learning* para a predição de casos de internações causadas por dengue em municípios da Paraíba

Ewerthon Dyego de Araújo Batista [1,2], Wellington Candeia de Araújo [1], Romeryto Vieira Lira [2], Laryssa Izabel de Araújo Batista [3]

[1] Núcleo de Tecnologias Estratégicas em Saúde – Universidade Estadual da Paraíba
Caixa Postal 781/791–58429-500 – Campina Grande – PB – Brasil

[2] Instituto Federal de Educação, Ciência e Tecnologia da Paraíba – Itaporanga, Brasil.

[3] Universidade Federal da Paraíba, PB – Brasil

ewerthon.batista@aluno.uepb.edu.br, wcandeia@uepb.edu.br,
romeryto.lira@academico.ifpb.edu.br, laryssa.izabel@gmail.com

Abstract. *Dengue is a public health problem in Brazil, cases of the disease started to grow in Paraíba. The epidemiological bulletin of Paraíba, released in August 2021, reports an increase of 53% of cases compared to the previous year. Machine Learning (ML) and Deep Learning techniques are being used as tools to predict the disease and support its combat. Using Random Forest (RF), Support Vector Regression (SVR), Multilayer Perceptron (MLP), Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) techniques, this article presents a system capable of predicting hospital admissions caused by dengue in the cities Bayeux, Cabedelo, João Pessoa and Santa Rita. The system managed to perform forecasts for Bayeux with an error rate of 0.5290, while in Cabedelo the error was 0.92742, João Pessoa 9.55288 and Santa Rita 0.74551.*

Resumo. *Dengue é um problema de saúde pública no Brasil, os casos da doença voltaram a crescer na Paraíba. O boletim epidemiológico da Paraíba, divulgado em agosto de 2021, informa um aumento de 53% de casos em relação ao ano anterior. Técnicas de Machine Learning (ML) e de Deep Learning estão sendo utilizadas como ferramentas para a predição da doença e suporte ao seu combate. Por meio das técnicas Random Forest (RF), Support Vector Regression (SVR), Multilayer Perceptron (MLP), Long Short-Term Memory (LSTM) e Convolutional Neural Network (CNN), este artigo apresenta um sistema capaz de realizar previsões de internações causadas por dengue para as cidades Bayeux, Cabedelo, João Pessoa e Santa Rita. O sistema conseguiu realizar previsões para Bayeux com taxa de erro 0,5290, já em Cabedelo o erro foi 0,92742, João Pessoa 9,55288 e Santa Rita 0,74551.*

1. Introdução

Dengue é uma doença endêmica, causada pelo vírus DENV, e transmitida através do mosquito *Aedes aegypti*. Atualmente, existem quatro tipos sorológicos do vírus (1, 2, 3 e 4) em circulação no Brasil [Pham et al. 2018]. Embora não seja uma doença nova, ainda não existe vacina eficaz para a imunização da população contra todos os tipos do vírus. Uma vez infectado por uma sorologia, o paciente adquire imunidade a essa variação, porém, continua suscetível às demais [Swaminathan and Khanna 2019]. Adicionalmente, de acordo com o último boletim epidemiológico, a Paraíba passa por crescimento nos índices de doenças causadas por arboviroses [Paraíba 2021].

Para combater a doença, os sistemas governamentais investem em campanhas de conscientização da população, sendo uma delas o correto descarte de pneus e recipientes, que estão a céu aberto. Isso acontece, visto que, eles podem acumular água e se tornar, futuramente, berço para proliferação do mosquito.

As técnicas de ML e de DL vêm sendo utilizadas como ferramentas de apoio ao combate da dengue Doni and Sasipraba (2020), Xu et al. (2020) e Mussumeci and Codeço (2020). Por meio da utilização de dados epidemiológicos, climáticos e fatores sociais, pesquisadores estão desenvolvendo modelos de previsão de casos da doença. Com base nas previsões, os governantes podem direcionar melhor os esforços e os recursos contra a doença [Guo et al. 2017].

Diante desse cenário, o objeto deste trabalho é, por meio das técnicas RF, SVR, MLP, LSTM e CNN, criar e avaliar modelos para predição de casos internações causadas por dengue para as cidades: Bayeux, Cabedelo, João Pessoa e Santa Rita.

2. Fundamentação Teórica

2.1. Dengue

A arbovirose dengue é causada pelo vírus DENV e transmitida através do mosquito *Aedes aegypti*. Os principais sintomas da dengue são: febre alta, dores musculares, mal-estar, falta de apetite e dores de cabeça. Em alguns casos mais graves, a dengue pode causar hemorragias e levar o paciente a óbito [Carvalho et al 2019].

O ciclo da doença dengue inicia com o mosquito *Aedes aegypti* picando um humano infectado. Uma vez infectado, o mosquito é capaz de transmitir dengue até o fim da sua vida. O mosquito se reproduz através do depósito de seus ovos em águas paradas e encontra, em países de clima tropical, ambiente ideal para sua reprodução. Estudos apontam que o vírus vem sofrendo mutações e, além de reproduzir em águas limpas, vem obtendo sucesso em ambientes com águas sujas, como, por exemplo, em esgotos [Beserra et al. 2019].

As principais ações de combate à dengue se voltam contra a não proliferação do seu vetor. Para isso, há campanhas de conscientização da população solicitando o correto descarte de objetos, orientações sobre como armazenar água e a utilização de pesticidas [Norrby 2014].

2.2. Técnicas de *Machine Learning* e de *Deep Learning*

Random Forest (RF) é uma técnica de ML que, por meio da criação e de treinamento de diversas árvores de decisão, consolida os resultados das predições das árvores. *Support*

Vector Regression (SVR) é um método de regressão que utiliza vetores de suporte para encontrar uma função capaz de traçar um hiperplano contendo a maior parte dos dados de treinamento [Appice et al. 2020].

A *Multilayer Perceptron* (MLP) é uma rede neural, do tipo *feedforward*, formada por neurônios. A MLP é estruturada em uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. Com exceção da camada de saída, todos os nós estão totalmente conectados com os nós da camada seguinte [Awad and Khanna 2015]

Long Short-Term Memory (LSTM) faz parte das redes neurais recorrentes. A ideia por trás das LSTMs é a utilização de células capazes de decidir a curto e longo prazo quais dados devem ser incorporados ou esquecidos. O papel de decisão sobre a incorporação dos dados ou o seu descarte são, respectivamente, papéis do *input gate* e *forget gate* [Doni and Sasipraba 2020]. *Convolutional Neural Network* é uma técnica de DL muito empregada em processamento de imagens. Tradicionalmente, as redes neurais convolucionais são utilizadas para a extração de características importantes em imagens. Contudo, a convolução linear também vem sendo utilizada em previsões em razão da sua capacidade de capturar padrões nas séries temporais [Zhao, K. *and* Wang 2017].

3. Trabalhos Relacionados

A previsão de doenças não é uma tarefa fácil. Existem vários fatores influenciadores e impactantes durante a predição, como, por exemplo, fatores climáticos, fatores econômicos, fatores sociais, mobilidade urbana, entre outros [Mussumeci and Codeço 2020]. Devido à complexidade, inúmeros trabalhos estão utilizando ML e DL durante a predição de doenças.

[Doni and Sasipraba 2020], na Índia, conduziram um estudo utilizando as técnicas LSTM, SVR, *Extreme Gradient Boosting*(XGboost), RF e *Generalized additive model* (GAM). Na criação dos modelos foram utilizados dados epidemiológicos e climáticos, fornecidos pelo governo, entre os anos de 2015 e 2018. A técnica LSTM obteve o melhor resultado com um *Root mean square error* (RMSE) de 42,00.

[Xu et al. 2020] realizaram predições de casos de dengue na China. Utilizando dados meteorológicos e as técnicas LSTM, *Back propagation neural network* (BPNN), GAM e SVR. A LSTM obteve a menor taxa de erro, RMSE 36,50.

[Mussumeci and Codeço 2020] propuseram a previsão de casos de dengue nos municípios do estado do Rio de Janeiro, utilizando as técnicas de LSTM, RF *e Least absolute shrinkage and selection operator* (LASSO), dados climáticos e históricos da doença. Como resultado, a técnica LSTM obteve uma taxa de erro de 0,45.

[Carvajal et al. 2018], nas Filipinas, utilizaram dados climáticos para a previsão de dengue. As técnicas utilizadas foram RF e GAM e a verificação dos valores ficou por meio do RMSE. RF obteve a menor taxa de erro: 0,29.

[Guo et al. 2017], utilizando dados históricos de dengue, entre 2011 e 2014, na China, conseguiram prever casos da doença com um RMSE 0,2861 através da SVR.

[Appice et al. 2020], no México, propuseram a criação de modelos de previsão de casos de dengue utilizando as técnicas *AUTOencoding based TIme series Clustering with Nearest Neighbour* (AutoTic-NN), *K-Nearest Neighbourhood* (KNN), SVR e *Autoregressive integrated moving average* (ARIMA). Para a produção dos seus

modelos, foram utilizados dados históricos da doença e climáticos entre 1985 e 2010.O AutoTic-NN obteve a menor taxa de erro, RMSE 5,18.

4. Metodologia

4.1. Base de dados

Para a criação dos modelos de predição deste estudo, foram utilizados os números mensais das internações causadas por dengue nos municípios. As internações tiveram como fonte o SIH/SUS entre os anos de 2010 e 2019. Além dos dados históricos da doença, para cada município, foi utilizado o valor mensal de precipitação. A pluviometria foi fornecida pela AESA (Agência Executiva de Gestão das Águas do Estado da Paraíba)

Do SNIS (Sistema Nacional de Informações sobre Saneamento), foram coletadas informações sobre água e esgoto entre os anos de 2010 e 2019 e incorporadas ao trabalho. Após a análise, optou-se por utilizar os dados de índice de coleta de esgoto e índice de tratamento de esgoto. Por fim, foi criado um banco de dados relacional, chamado DashDengue.

4.2. Sistema para predição de casos de internação

O desenvolvimento do sistema de predição foi realizado utilizando a linguagem Python. Ademais, foram empregadas as bibliotecas *Scikit-learn* e *TensorFlow*. As fases de desenvolvimento foram divididas em: criação, treinamento e escolha dos modelos, ajustes nos hiperparâmetros, geração das previsões e, por fim, validação estatística dos resultados.

4.2.1 Criação, treinamento e escolha dos modelos

A primeira etapa do processo foi recuperar os dados da base DashDengue e normalizá-los. De posse das informações, foram criados quatros cenários, variando a quantidade de anos dos dados e se haveria ou não tratamento de observações fora do comum (*outliers*). Na sequência, foram propostos 8 modelos candidatos contendo as combinações dos atributos: número de internação, pluviometria mensal, índice de coleta de esgoto e índice de tratamento de esgoto. Finalmente, para cada atributo, foram adicionadas de 0 a 4 *lags* (informações do passado para os atributos) com os dados dos últimos meses.

Definidos os modelos, os mesmos foram submetidos ao treinamento através das técnicas RF, SVR, MLP, LSTM e CNN. O RF foi executado com a configuração *n_estimators*=50. Para o SVR, foram utilizados os parâmetros padrões do *Scikit-learn*. Em relação à MLP, aqui chamada de RN, foram adicionadas duas camadas ocultas com 250 neurônios. Parâmetros não citados seguiram os valores padrão do *Scikit-Learn*.

A técnica LSTM foi implementada com base na biblioteca *TensorFlow* e possui a seguinte característica: camada de entrada, camada LSTM com 50 neurônios e uma camada densa. A quantidade de épocas foi definida em 50 e o *batch_size* em 12.

Finalmente, a técnica CNN é a combinação de camadas convolucional e LSTM. A primeira camada é a de entrada seguida de uma camada Conv1D, com 16 filtros. Na sequência foi adicionada uma camada LSTM com 50 neurônios e, por fim, uma camada densa. A quantidade de épocas foi de 50 e o *batch_size*=12. Os Parâmetros não listados assumiram os valores padrão do *TensorFlow*.

Os dados foram separados entre treino e teste, seguindo a proporção de 80% para treino e 20% para teste. Finalizado o treinamento, foram geradas previsões e, utilizando a técnica RMSE [Carvajal et al 2018], a taxa de erro foi mensurada. Sobre o RMSE, quanto mais baixo for o seu valor melhor é a previsão. O modelo e cenário com melhores resultados foram os escolhidos para seguir para a próxima etapa.

4.2.2 Ajustes nos hiperparâmetros

Na fase de ajuste de parâmetros, os modelos vencedores foram submetidos a variação de parâmetros. A saber: RF (n_*estimators*, *min_samples_split* e *min_samples_leaf*), SVR (*kernel*, *gamma*, C e *epsilon*), RN (*activation*, *solver*, *batch_size*, *learning*, neurônios), LSTM (*neurons*, *activation*, *recurrent_activation*, *dropout* e *batch_size*) e CNN (*filters*, *neurons*, *activation*, *recurrent_activation*, *dropout* e *batch_size*). A configuração com melhor taxa de erro foi utilizada para a previsão dos casos de internação

4.2.3 Geração de previsões, avaliação dos resultados e comprovações estatísticas

De posse da melhor configuração de modelo, cenário e parâmetros, os dados foram submetidos ao treinamento, à geração das previsões numéricas de internações e ao cálculo da taxa de erro. Por questões estocásticas de algumas técnicas, separados em quatros *rounds*, foram realizadas 100 execuções para cada uma das configurações. Com intuito de evitar o *overfitting*, a técnica de parada antecipada foi implementada.

Não foram encontrados na literatura estudos realizando previsões de internação para os municípios aqui listados. Com isso, adotou-se o procedimento de comparar os resultados com a abordagem *Naive Forecast* [Kirby et al 2015] e ARIMA [Appice et al 2020]. Em relação à validação estatística, foi verificada a normalidade dos resultados usando o teste de Shapiro-Wilk. Se os resultados obedecessem à curva normal, as diferenças estatísticas seriam computadas através dos testes Anova e Tukey. Caso contrário, foram usados Kruskall e Dunn.

5. Resultados e Discussão

O primeiro objetivo deste projeto foi verificar qual modelo e cenário produziu melhores resultados para as cidades alvo. Para as cidades em estudo, a Tabela 1 lista a melhor taxa de erro, parâmetros previsores, período e *lags*.

Tabela 1 - Resultados contendo os modelos vencedores e suas configurações.

Município	Melhor RMSE	Parâmetros	Período	Outlier	Qtd *lags*
Bayeux	0,629282246	Internações, precipitação e índice de coleta de esgoto	2010 - 2019	Não	4
Cabedelo	0,955240078	Internações, precipitação e índice de coleta de esgoto	2015 - 2019	Sim	4
João Pessoa	10,10272226	Internações	2015 – 2019	Sim	3
Santa Rita	1,081508751	Internações, precipitação e índice de coleta de esgoto	2015 - 2019	Sim	3

Como notado, a combinação de parâmetros de internações, precipitação e índice de coleta de esgoto obteve melhores resultados em três das quatros cidades. A não utilização de informações de esgoto, em João Pessoa, pode ser explicada devido a maior média de coleta de esgoto (69,37%) e de tratamento de esgoto (99,95%) dessa cidade comparada as outras três. O tratamento de *outliers* e quantidade de anos seguiram a mesma tendência. Enfim, o número de *lags* variou entre 4 e 3. A Tabela 2 ilustra os resultados obtidos após os ajustes de hiperparâmetros e qual técnica foi a vencedora.

Tabela 2 - Resultados após ajustes de parâmetros e técnica vencedora.

Município	Melhor RMSE	Técnica vencedora
Bayeux	0,529017139	LSTM
Cabedelo	0,927428107	LSTM
João Pessoa	9,552880644	CNN
Santa Rita	0,745519952	RF

A estratégia de ajuste de parâmetros melhorou a taxa de erro para todas as cidades. A técnica LSTM venceu em 50% das cidades, com CNN e RF aparecendo em uma cidade cada. Durante a validação estatística, para todas as cidades, a curva normal foi seguida. Com isso, o teste de Tukey foi realizado e os resultados estão na Figura 1. Na sequência, as figuras 2, 3 e 4 demonstram as previsões obtidas e o comparativo entre o resultado das previsões e o número real de internações causadas por dengue.

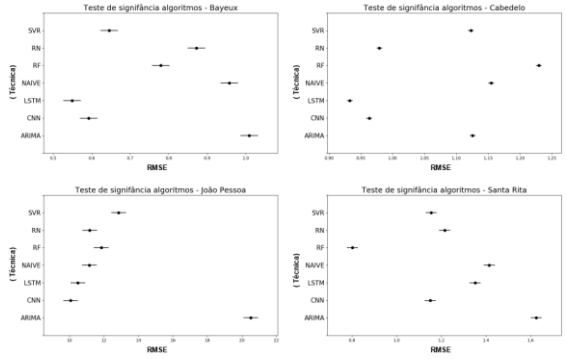

Figura 1 – Diferença estatística entre as técnicas de ML e DL.

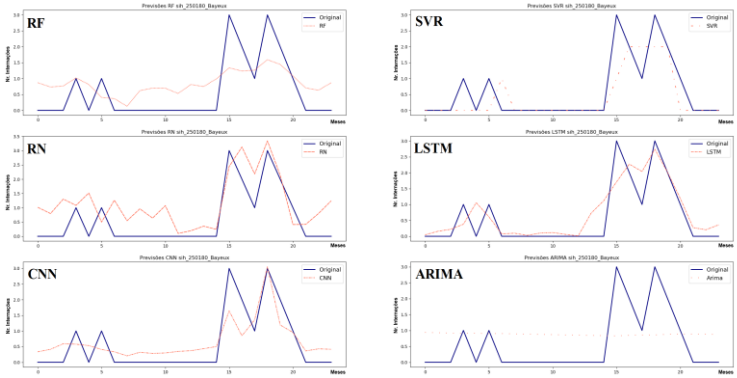

Figura 2 – Previsões feitas pelas técnicas para a cidade de Bayeux.

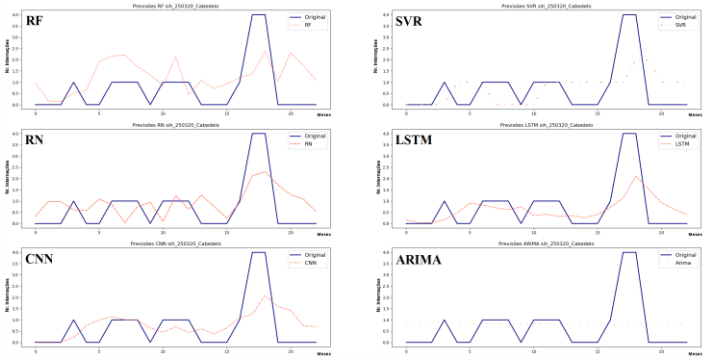

Figura 3 - Previsões feitas pelas técnicas para a cidade de Cabedelo.

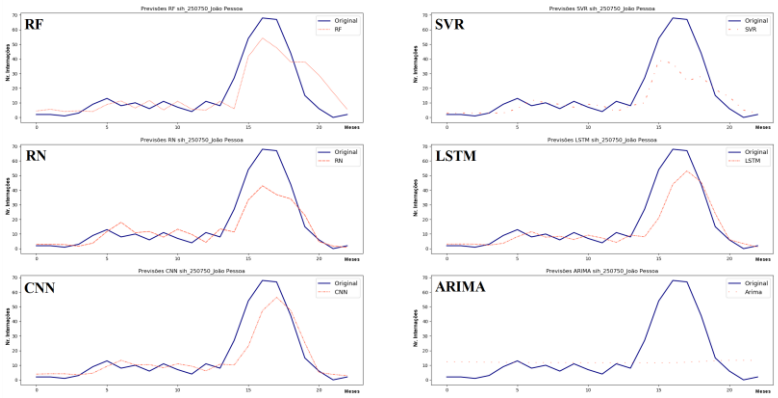

Figura 4 - Resultados das previsões feitas para a cidade de João Pessoa.

Como pode ser observado, LSTM obteve os melhores resultados para a cidade de Bayeux. Contudo, não há, estatisticamente, diferença entre ela e a CNN. Sobre Cabedelo, LSTM venceu e os testes estatísticos demonstraram a sua superioridade. Para João Pessoa, não ficou comprovada a diferença estatística entre as primeiras colocadas, respectivamente, CNN e LSTM. Por fim, a técnica RF obteve menor taxa de erro em Santa Rita e a sua superioridade ficou comprovada.

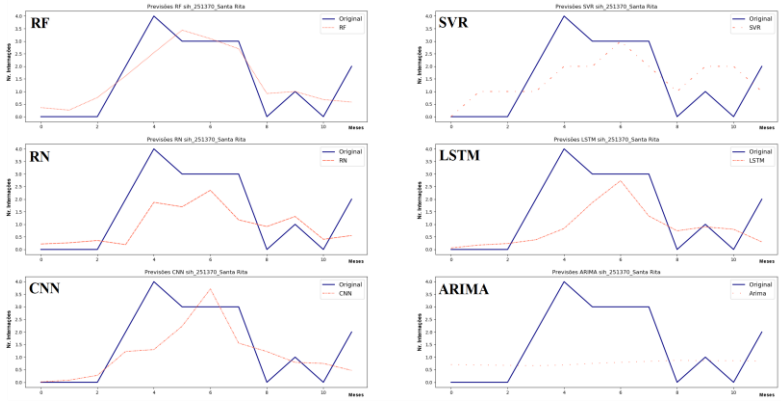

Figura 5 - Resultados das previsões feitas para a cidade de Santa Rita.

6. Considerações e Trabalhos Futuros

Empregando dados epidemiológicos, climáticos e sanitários foi possível criar, avaliar e realizar previsões de casos internações causadas por dengue por meio de ML e DL.

Foi constatado que as técnicas LSTM e CNN saíram vencedoras em 75 % das cidades. Com exceção da cidade de Bayeux, modelos com tratamento de *outlier* e a utilização de dados entre 2015 e 2019 obtiveram os melhores resultados. Por fim, ficou demonstrada, estatisticamente, a diferença entre as técnicas e a superioridade das abordagens de ML e DL.

Para trabalhos futuros, é sugerido a verificação de utilização de outros parâmetros para a criação dos modelos, tais como: índice de coleta de resíduos, indicadores de umidade e temperatura média. Além disso, é válido a utilização de outras técnicas como redes neurais recorrentes e novas variações das técnicas aqui estudadas. Como limitação deste trabalho, destaca-se a ausência de dados do ano 2020.

Referências

Appice, A., Gel, Y.R., Iliev, I., Lyubchich and V.,Malerba, D. (2020). A Multi-Stage Machine Learning Approach to Predict Dengue Incidence: A Case Study in Mexico. *In IEEE Access*, v. 8, p. 52713–52725.

Awad, M. and Khanna, R. (2015) Efficient learning machines: Theories, concepts, and applications for engineers and system designers. In *Apress Media LLC*.

Beserra, Eduardo B., Freitas, Eraldo M. de, Souza, José T. de, Fernandes, Carlos R. M. and Santos, Keliana D. (2009). Ciclo de vida de Aedes (Stegomyia) aegypti (Diptera, Culicidae) em águas com diferentes características. In *Iheringia. Série Zoologia*.

Carvajal, T.M., Viacrusis, K.M., Hernandez, L.F.T., et al. (2018). Machine learning methods reveal the temporal pattern of dengue incidence using meteorological factors in metropolitan Manila, Philippines. In *BMC Infectious Diseases*, v. 18, n. 1.

Carvalho, T. M., Tenório, G. L., Figueiredo, K., Vellasco and M., Caarls, W. (2019) Comparison of Machine Learning Models for Total Dengue Cases Prediction.

Doni, A. R. and Sasipraba, T. (2020). Lstm-Rnn Based Approach for Prediction of Dengue Cases in India. In *Ingenierie des Systemes d'Information.*

Guo, Pi, Liu, Tao, Zhang, Qin, Wang, Li, Xiao, Jianpeng, Zhang, Qingying, Luo, Ganfeng, Li, Zhihao, He, Jianfeng, Zhang, Yonghui and Ma, Wenjun. (2017) Developing a dengue forecast model using machine learning: A case study in China. In *PLoS Neglected Tropical Diseases*, v. 11, n. 10.

Kirby, Simon, Paramaguru, Kanya and Warren, James. (2015) The Accuracy of NIESR's GDP Growth Forecasts. In *National Institute Economic Review.*

Mussumeci, E. and Codeço Coelho, F. (2020). Large-scale multivariate forecasting models for Dengue - LSTM versus random forest regression. Spatial and Spatio-temporal Epidemiology, v. 35.

Norrby, R. (2014) Outlook for a dengue vaccineClinical Microbiology and Infection. Disponível: <https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue>. Acesso em: 21 set. 2020

PARAÍBA.(2021) Boletins Epidemiológicos. Disponível em: <https://paraiba.pb.gov.br/diretas/saude/consultas/vigilancia-em-saude-1/boletins-epidemiologicos>. Acesso em: 21 jul. 2021.

Pham, Duc Nghia, Aziz, Tarique, Kohan, Ali, Nellis, Syahrul , Jamil, Juraina Binti Abd ,Khoo, Jing Choco ,Lukose, et al. (2018) How to Efficiently Predict Dengue Incidence in Kuala Lumpur. Proceedings - 2018 In *4th International Conference on Advances in Computing, Communication and Automation, ICACCA*

Swaminathan, S. and Khanna, N. (2019) Dengue vaccine development: Global and Indian scenarios. In *International Journal of Infectious* Diseases, v. 84, p. S80–S86.

Xu, Keqiang, Li, Zhichao, Meng, Fengxia, Tu, Taotian, Xu, Lei. et al. (2020). Forecast of dengue cases in 20 chinese cities based on the deep learning method. In *International Journal of Environmental Research and Public Health*, v. 17, n. 2.

Zhao, K. and Wang, C. (2017). Sales Forecast in E-commerce using Convolutional Neural Network.