Um Experimento de Engenharia de Features para Geração de Modelos Preditivos para Casos de Dengue

Ramon Garcia[1], Eduardo Ogasawara[1], Jorge Soares[1], Amaury de Souza[2], Rejane Sobrino[3], Eduardo Bezerra[1]

[1]Programa de Pós-graduação em Ciência da Computação
Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ)
Rio de Janeiro, RJ – Brasil

[2]Instituto de Física
Universidade Federal do Mato Grosso do Sul (UFMS)
Campo Grande, MS – Brasil

[3]Instituto de Estudos em Saúde Coletiva
Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro, RJ – Brasil

ramon.garcia@aluno.cefet-rj.br,rejane@iesc.ufrj.br

{eogasawara,jorge.soares,ebezerra}@cefet-rj.br

amaury.souza@ufms.br

Abstract. *This study aims to develop machine learning models to predict the number of dengue cases in a given health facility. Our approach involves feature engineering by integrating data from various domains. Specifically, we combine data from Brazil's Unified Health System with meteorological data from the National Institute of Meteorology and the GOES-16 weather satellite. We train Long Short-Term Memory neural networks to generate predictive models that capture climatic patterns and their influences on dengue incidence, considering both spatial and temporal data.*

Resumo. *Este estudo tem como objetivo desenvolver modelos de aprendizado de máquina para prever o número de casos de dengue em uma determinada unidade de saúde. Nossa abordagem envolve a engenharia de features por meio da integração de dados de vários domínios. Especificamente, combinamos dados do Sistema Único de Saúde do Brasil com dados meteorológicos do Instituto Nacional de Meteorologia e do satélite meteorológico GOES-16. Treinamos redes neurais do tipo Long Short-Term Memory (LSTM) para gerar modelos preditivos que capturam padrões climáticos e suas influências na incidência de dengue, considerando dados espaciais e temporais.*

1. Introdução

A dengue é uma doença viral transmitida principalmente pelo mosquito *Aedes aegypti*, que se tornou uma grande preocupação de saúde pública global. Em 2024, o Brasil registrou mais de seis milhões de casos suspeitos de dengue, refletindo a seriedade do problema e a necessidade de entender melhor os fatores que influenciam a proliferação dos mosquitos e a transmissão do vírus [Ministério da Saúde, 2024a].

Estudos indicam que temperaturas altas aceleram o desenvolvimento do mosquito e aumentam a taxa de reprodução, enquanto a precipitação fornece locais de criadouro, aumentando a densidade populacional do vetor [Reinhold et al., 2018]. Condições climáticas favoráveis, como temperaturas entre 25-30°C e chuvas periódicas, são cruciais para a proliferação do *Aedes aegypti*.

Além dos fatores climáticos, aspectos socioeconômicos e ambientais, como o saneamento básico inadequado e a urbanização desordenada, também contribuem para a proliferação do mosquito. A combinação desses fatores cria um ambiente propício para o aumento dos casos de dengue, destacando a necessidade de intervenções integradas [Machado et al., 2013].

O Sistema Único de Saúde (SUS) é a rede de atendimento de saúde pública no Brasil, oferecendo acesso universal e gratuito a serviços de saúde para toda a população. Dentro do SUS, o Sistema de Informação de Agravos de Notificação (SINAN) é uma ferramenta que coleta, transmite e dissemina dados sobre doenças de notificação compulsória, incluindo a dengue. Outro componente do SUS é o Cadastro Nacional de Estabelecimentos de Saúde (CNES), que mantém um registro atualizado de todas as unidades de saúde no Brasil [Ministério da Saúde, 2024b].

O objetivo deste estudo é investigar a criação de modelos de Aprendizado de Máquina para prever a quantidade de casos de dengue. Nossa abordagem emprega engenharia de *features* integrando dados provenientes de diferentes domínios. Em particular, realizamos a fusão de dados do Sistema Único de Saúde (SUS), incluindo o Sistema de Informação de Agravos de Notificação (SINAN), com dados meteorológicos do Instituto Nacional de Meteorologia (INMET) e do satélite meteorológico GOES-16. Nossa hipótese é que essa combinação de dados permite uma análise mais abrangente dos fatores que influenciam a incidência de dengue. Utilizamos *Long Short-Term Memory* (LSTM) para gerar modelos preditivos que capturem padrões climáticos e suas influências na incidência de dengue, considerando tanto dados espaciais quanto temporais [Schmidhuber, 2015; Hochreiter and Schmidhuber, 1997].

O artigo está organizado conforme a seguir. Na Seção 2, exploramos trabalhos anteriores que formam a base teórica e metodológica do nosso estudo. Na Seção 3, descrevemos o processo de coleta de dados e ajuste dos modelos. A Seção 4 apresenta os resultados obtidos, assim como fornece uma análise das limitações da nossa abordagem atual. A Seção 5 apresenta considerações finais.

2. Trabalhos Relacionados

Em nossa busca por trabalhos relacionados, encontramos três estudos relevantes que abordam aspectos similares em termos de domínio do problema e abordagem.

Zhao et al. [2020] compararam o uso de *Random Forests* (RF) e *Artificial Neural Networks* (ANN) para prever a carga de dengue em níveis nacional e subnacional na Colômbia. O estudo destaca que os modelos RF superaram os modelos ARIMA tradicionais em termos de precisão de previsão, especialmente quando variáveis sociodemográficas foram incluídas.

Salim et al. [2021] utilizaram técnicas de aprendizado de máquina para prever surtos de dengue em Selangor, Malásia. Os autores empregaram múltiplos modelos de

aprendizado de máquina, incluindo *Support Vector Machines* (SVM) e *Gradient Boosting*. Nesse estudo, os autores concluíram que a inclusão de dados climáticos e ambientais melhora significativamente a precisão das previsões.

Alves et al. [2021] desenvolveram um modelo de transmissão da dengue impulsionado por dados climáticos para analisar a dinâmica da dengue em diferentes regiões do Brasil. Foi utilizada uma abordagem estatística baseada no modelo SEI-SIR, incorporando transmissão vertical do vetor e o compartimento de ovos do vetor, permitindo que a precipitação modulasse a eclosão dos ovos. Dados de temperatura e precipitação de satélites foram utilizados como entradas climáticas do modelo.

Buscamos nos destacar dos trabalhos relacionados pela integração de dados provenientes de múltiplas fontes e pela abordagem orientada a dados para a previsão de casos de dengue. Nossa metodologia, como detalhado na Seção3, utiliza informações sobre o ciclo de vida do mosquito e seu comportamento para a engenharia de *features* que serão usadas no treinamento dos modelos.

3. Metodologia

Em nossa metodologia para construção de modelos preditivo para casos de Dengue, seguimos uma abordagem estruturada que incluiu a coleta de dados, pré-processamento, derivação de variáveis climáticas, treinamento de modelos e validação. Este estudo focou-se especificamente em dados coletados no estado do Rio de Janeiro, no período de 2020 a 2023. A área geográfica considerada abrangeu as coordenadas delimitadas entre 23.3702°S e 20.7634°S de latitude, e entre 44.7930°W e 40.7635°W de longitude.

Os dados utilizados neste estudo foram obtidos a partir de três fontes: (1) o banco de dados SINAN, que fornece dados epidemiológicos de notificações de casos de dengue, incluindo informações sobre local (i.e., estabelecimento de saúde), data e número de casos diários; (2) as estações meteorológicas do INMET (Instituto Nacional de Meteorologia), que fornecem dados horários de temperatura e precipitação; e (3) o satélite meteorológico GOES-16, especificamente os produtos LST (*Land Surface Temperature*) e RRQPE (*Rainfall Rate and Quantitative Precipitation Estimation*).

Os dados do SINAN representam notificações de dengue em diversos estabelecimentos de saúde. As notificações, incluindo todas as suspeitas independentemente de confirmação, foram contabilizadas por estabelecimento de saúde e dia. Em seguida, separamos as unidades que tiveram casos em todos os anos do período considerado neste estudo. Finalmente, selecionamos as dez unidades que mais tiveram casos e as dez que menos tiveram casos. As unidades de saúde selecionadas, bem como suas respectivas quantidades de casos, se encontram na Tabela 1.

As medições de satélite foram coletadas a partir do satélite meteorológico GOES-16, na resolução *full disk*. Esses dados são disponibilizados em tempo real, e estão disponível em um formato de grade. Dentre os vários produtos disponibilizados por este satélite, neste estudo utilizamos o LST e o RRPQE. O produto LST mede a temperatura da superfície terrestre. Ele é útil para o monitoramento climático e ambiental, ajudando a compreender e prever fenômenos como ondas de calor, secas e variações sazonais. O produto RRQPE fornece estimativas da quantidade de precipitação. Os produtos LST e RRQPE são disponibilizados por hora e a cada dez minutos, respectivamente. Para este

CNES	2020	2021	2022	2023	Total	Estação	Distância (Km)
7427549	51	85	390	733	1259	A621	14.68
2268922	287	45	147	678	1157	A667	28.69
7149328	15	2	109	1021	1147	A609	2.14
2299216	50	273	117	695	1135	A609	1.71
0106453	39	27	167	639	872	A608	16.29
6870066	118	129	5	581	833	A609	6.29
6042619	15	744	4	15	778	A626	19.06
2288893	223	74	14	466	777	A609	1.94
5106702	39	72	479	85	675	A608	3.16
6635148	264	47	2	334	647	A626	13.04
2269481	2	1	1	6	10	A627	22.11
2708353	3	2	2	3	10	A652	5.44
7591136	5	3	1	1	10	A621	14.34
2283395	4	1	1	2	8	A606	9.24
2287579	2	2	1	3	8	A607	3.67
2291533	4	2	1	1	8	A627	6.66
2292386	1	1	1	5	8	A618	2.68
0012505	4	1	1	2	8	A627	3.26
2292084	1	2	1	2	6	A627	6.47
6518893	1	1	1	1	4	A621	9.46
Total	**1128**	**1560**	**1651**	**6307**	**10646**		

Tabela 1. Estabelecimentos de saúde selecionados para este estudo. A parte superior lista os estabelecimentos com mais casos reportados, enquanto a parte inferior lista aqueles com menos casos.

estudo, esses dados foram agregados temporalmente para produzir uma série com resolução temporal diária, gerando as temperaturas mínima, média e máxima de cada dia a partir do LST e a precipitação acumulada a partir do RRQPE. As resoluções espaciais dos produtos LST e RRQPE são de 10 Km e 2 Km, respectivamente.

Os dados das estações meteorológicas foram obtidos através da API do INMET. Primeiramente, foram selecionadas estações meteorológicas localizadas dentro das coordenadas da área geográfica considerada para estudo. As estações escolhidas foram as automáticas, capazes de medir temperatura e precipitação. Esta seleção resultou em 10 estações. As estações do sistema INMET produzem dados com resolução temporal horária. Por outro lado, as observações de casos de Dengue provenientes do SINAN são diárias. Dessa forma, para fins de uniformização de resoluções temporais, as medições consideradas foram agregadas para gerar as temperaturas mínima, média e máxima diárias, bem como a precipitação acumulada diária. Lacunas nesses dados foram preenchidas usando interpolação linear de dados. Foram interpolados 0.8% dos dados de temperatura. Nenhum dado de chuva precisou ser interpolado.

No caso dos dados de satélite, determinar a temperatura e precipitação próxima à unidade de saúde foi relativamente simples: utilizamos a latitude e longitude do endereço

da unidade de saúde para selecionar a medição mais próxima realizada pelo satélite. Entretanto, devido à natureza dos dados coletados por esse sensor, pode haver momentos em que não há medições disponíveis para determinadas localidades. Nesses casos, as lacunas nos dados foram compensadas usando interpolação linear de dados. Para os dados das estações meteorológicas, utilizamos as medições da estação mais próxima da unidade de saúde, embora algumas unidades estejam situadas a uma distância considerável das estações meteorológicas. Uma porcentagem de 41.75% dos dados de temperatura e 18.4% dos dados de precipitação passaram por este procedimento de interpolação.

Com os dados coletados e organizados, prosseguimos para realizar engenharia de *features* baseada nos dados meteorológicos. Esse passo foi realizado de acordo com as informações sobre o impacto da temperatura no mosquito obtidas em Reinhold et al. [2018] e Carrington et al. [2013] e sobre o impacto da precipitação no mosquito obtido em Edillo et al. [2024].

A temperatura tem um impacto significativo no desenvolvimento do mosquito, na velocidade e distância de voo, e na quantidade de vezes que o mosquito se alimenta. Inicialmente, criamos uma *feature* para acompanhar a média móvel da temperatura nos últimos 7, 14 e 21 dias. Os mosquitos têm uma taxa de reprodução e desenvolvimento mais estável entre as temperaturas de 20°C a 35°C. Nossa hipótese é que, dentro dessa faixa de temperatura, a população de mosquitos aumente, levando a um crescimento no número de casos de dengue. Denominamos essas temperaturas como "temperatura ideal" e utilizamos os dias com temperatura média nessa faixa para criar *lag features* de 7, 14 e 21 dias, acompanhando o ciclo de vida do mosquito. Além disso, uma alta amplitude térmica tem impacto negativo no desenvolvimento do mosquito. A média móvel da amplitude térmica também foi monitorada em intervalos de 7, 14 e 21 dias. Por outro lado, temperaturas abaixo de 14°C e acima de 38°C são desfavoráveis para os mosquitos. Dias com essas temperaturas foram usados para criar a *feature* "temperatura adversa", que é monitorada a cada 7 dias.

A precipitação influencia diretamente a disponibilidade de locais de criadouro para o mosquito. A média móvel da precipitação é monitorada em intervalos de 7, 14 e 21 dias. Uma precipitação superior a 10mm contribui para a formação de novos locais onde o mosquito pode depositar seus ovos. Com base nessa informação, durante a engenharia de *features*, criamos a *feature* binária denominada "precipitação significativa". Nossa hipótese é que períodos constantes de precipitação significativa aumentem os criadouros dos mosquitos e, consequentemente, os casos de dengue. Portanto, acompanhamos os dias com precipitação superior a 10mm e a precipitação acumulada a cada 7, 14 e 21 dias. Uma precipitação superior a 150mm em um dia pode destruir os ovos existentes. Com base nisso, nas nossas *features*, identificamos esses dias como "precipitação extrema", e esses eventos são monitorados a cada 7 dias.

Finalmente, a quantidade de casos de Dengue influencia no surgimento de novos casos, devido à infecção horizontal do mosquito. Nessa modalidade de infecção, o mosquito se infecta com o vírus da dengue ao se alimentar de um ser humano infectado. Com o ciclo típico da doença, a quantidade de casos e a média móvel de casos são monitoradas a cada 14 e 21 dias, criando *lag features*. Os casos são agrupados e contabilizados por cada unidade de saúde.

Com base nos critérios descritos acima, geramos um total de 45 *features*, sendo 40 derivadas de fontes meteorológicas, geradas tanto para dados de satélite (20) quanto para dados de estações meteorológicas (20), e 5 relacionadas ao número de casos de dengue. Concretamente, as *features* usadas para treinamento dos modelos preditivos são: (i) Média móvel da temperatura (7, 14 e 21 dias); (ii) Média móvel da precipitação (7, 14 e 21 dias); (iii) Precipitação acumulada (7, 14 e 21 dias); (iv) Amplitude térmica (7, 14 e 21 dias); (v) Temperatura ideal (7, 14 e 21 dias); (vi) Temperatura adversa (7 dias); (vii) Precipitação significativa (7, 14 e 21 dias); (viii) Precipitação extrema (últimos 7 dias); (ix) Quantidade de casos no dia; (x) Quantidade de casos acumulados (14 e 21 dias); (xi) Média móvel de casos de dengue (14 e 21 dias).

Após a geração das *features*, ajustamos modelos preditivos por meio de redes neurais LSTM. Para isso, dividimos os dados da seguinte forma: as observações feitas no período de 01/01/2020 até 31/12/2022 foram usadas para treinamento, enquanto que as realizadas no período de 01/01/2023 até 31/12/2023 foram usadas para teste. Todos os modelos foram treinados para prever a quantidade de casos de Dengue no dia seguinte em cada unidade de saúde. Visto que a variável alvo é um valor numérico, os modelos foram avaliados utilizando a métrica *Mean Squared Error* (MSE). Além disso, como um passo de pós-processamento, arredondamos a saída dos modelos para o valor inteiro mais próximo.

Para fins de comparação, para cada unidade de saúde, treinamos modelos considerando quatro conjuntos de *features* distintas: (F1) todas as *features*; (F2) combinação de *features* de estações meteorológicas com *features* de casos; (F3) combinação de *features* de satélites com *features* de casos; (F4) combinação sem *features* provenientes de dados meteorológicos.

4. Resultados e Discussão

Os experimentos para geração dos diversos modelos foram implementados utilizando a biblioteca Keras[1]. A rede neural utilizada para treinar todos os modelos foi definida com uma única camada LSTM de 50 unidades. A camada de saída foi definida com uma única unidade. O otimizador utilizado foi o Adam com taxa de aprendizado igual a 0.001, e a função de perda usada foi a MSE, que é apropriada para problemas de regressão.

A Tabela 2 resume os valores de MSE obtidos sobres os conjuntos de teste. Para cada unidade de saúde (CNES), são apresentadas os valores MSE para os diferentes conjuntos de *features*. O valor MSE do modelo vencedor está destacado em negrito. Os resultados indicam que o modelo treinado sem as *features* climáticas apresentou, em geral, um desempenho superior, conforme evidenciado pelos menores valores de MSE na maioria das unidades de saúde.

A análise detalhada dos resultados revela que modelos que não utilizam *features* climáticas apresentam desempenho preditivo superior na maioria das unidades de saúde estudadas. Este resultado é surpreendente, pois esperávamos que a inclusão de variáveis meteorológicas melhorasse a precisão dos modelos em todos os casos, dada a influência conhecida das condições climáticas sobre a proliferação do *Aedes aegypti* e a transmissão do vírus da dengue.

[1]https://keras.io

CNES	F1	F2	F3	F4
7427549	**0.0192**	0.0575	6.1945	0.0329
2268922	**0.7863**	0.8356	6.5699	1.3178
7149328	8.2247	7.9890	27.6137	**4.9397**
2299216	4.7562	5.1589	13.8055	**4.6110**
0106453	**0.9041**	0.6384	9.0904	0.9808
6870066	9.4795	8.8219	22.5123	**7.0438**
6042619	0.1589	0.1014	0.5041	**0.0630**
2288893	2.8356	2.2493	15.6	**1.8137**
5106702	1.1808	0.8082	2.8521	**1.0137**
6635148	**0.2384**	0.3342	3.7014	0.5288
Média	2.85837	2.69944	10.84439	2.23452
2269481	**0.0**	**0.0**	0.0164	**0.0**
2708353	**0.0**	**0.0**	0.0082	**0.0**
7591136	**0.0**	**0.0**	0.0027	**0.0**
2283395	0.0027	0.0027	**0.0110**	0.0027
2287579	**0.0**	**0.0**	0.0082	**0.0**
2291533	**0.0**	**0.0**	0.0027	**0.0**
2292386	**0.0**	**0.0**	0.0192	**0.0**
0012505	**0.0**	**0.0**	0.0055	**0.0**
2292084	**0.0**	**0.0**	0.0055	**0.0**
6518893	**0.0**	**0.0**	0.0027	**0.0**
Média	0.00027	0.00027	0.00821	0.00027
Média Geral	1.42932	1.349855	5.4263	1.117395

Tabela 2. Métricas de desempenho dos modelos preditivos para diferentes conjuntos de *features* e para diferentes unidades de saúde.

Ao examinarmos mais de perto, notamos que as *features* derivadas dos dados de satélite apresentaram um desempenho inferior em comparação com as *features* provenientes das estações meteorológicas do INMET. Isso nos leva a crer que essas *features* estejam impactando negativamente o modelo com todas as *features*.

Os modelos que utilizaram dados do INMET apresentaram uma melhoria moderada em comparação com os dados de satélite, mas ainda assim foram superados pelos modelos sem *features* climáticas em 6 das 10 unidades no grupo de unidades com muitos casos. Isso sugere que, embora os dados meteorológicos sejam relevantes, a precisão desses dados tem impacto significativo nos modelos.

As unidades de saúde com menos casos de dengue apresentaram resultados com menor MSE, indicando que a variabilidade nos dados de casos pode influenciar a performance dos modelos preditivos. A baixa variabilidade nos dados pode ter facilitado a modelagem e previsão dos casos de dengue, enquanto unidades com maior número de casos e variabilidade mais alta apresentaram maiores desafios para os modelos.

5. Conclusão

Este estudo propôs a utilização de modelos de Aprendizado de Máquina para prever a incidência de casos de dengue, integrando dados de múltiplas fontes, SINAN, CNES, e dados meteorológicos. Os resultados preliminares aqui reportados indicam que a fusão de dados provenientes de diferentes domínios, como saúde pública e meteorologia, pode melhorar a precisão dos modelos preditivos em alguns casos. Embora os modelos sem *features* climáticas tenham apresentado melhor desempenho na maioria dos casos testados, há potencial para melhorias significativas por meio do refinamento das *features* climáticas e da inclusão de outras variáveis relevantes. A integração de dados de diferentes fontes e a modelagem adequada das condições locais podem aprimorar a previsão dos casos de dengue.

Como trabalhos futuros, pretendemos realizar experimentos com dados meteorológicos mais precisos espacialmente, como o CHIRPS (*Climate Hazards Group InfraRed Precipitation with Station*). Outra continuação planejada é criar modelos para outras arboviroses (Zika e Chikungunya). Pretendemos também investigar modelos baseados na técnica *Physhics Informed Machine Learning* para produzir modelos que incorporem de maneira mais acurada a dinâmica de evolução dos casos de Dengue.

Referências

Alves, L., Lana, R., and Coelho, F. (2021). A framework for weather-driven dengue virus transmission dynamics in different brazilian regions. *Int. J. Environ. Res. Public Health*, 18:9493.

Carrington, L. B., Armijos, M. V., Lambrechts, L., Barker, C. M., and Scott, T. W. (2013). Effects of fluctuating daily temperatures at critical thermal extremes on aedes aegypti life-history traits. *PLoS ONE*, 8(3):e58824.

Edillo, F., Ymbong, R. R., Navarro, A. O., Cabahug, M. M., and Saavedra, K. (2024). Detecting the impacts of humidity, rainfall, temperature, and season on chikungunya, dengue and zika viruses in aedes albopictus mosquitoes from selected sites in cebu city, philippines. *Virology Journal*, 21:42.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Machado, C. J. S., Miagostovich, M. P., Leite, J. P. G., and Vilani, R. M. (2013). Promoção da relação saúde-saneamento-cidade por meio da virologia ambiental. *Revista de Informação Legislativa*, 50(199):321–345.

Ministério da Saúde (2024a). Indicadores de dengue.

Ministério da Saúde (2024b). Sistema Único de saúde (sus).

Reinhold, J. M., Lazzari, C. R., and Lahondère, C. (2018). Effects of the environmental temperature on aedes aegypti and aedes albopictus mosquitoes: A review. *Insects*, 9(4):158.

Salim, N. A. M., Samsudin, N. A., Ismail, R., et al. (2021). Prediction of dengue outbreak in selangor malaysia using machine learning techniques. *Sci. Rep.*, 11:79193.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117.

Zhao, N., Charland, K., Carabali, M., Nsoesie, E. O., Maheu-Giroux, M., Rees, E., et al. (2020). Machine learning and dengue forecasting: Comparing random forests and artificial neural networks for predicting dengue burden at national and sub-national scales in colombia. *PLOS Neglected Tropical Diseases*, 14(9).