Research, Society and Development, v. 10, n. 10, e33101119347, 2021
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v10i11.19347

Previsão de casos de dengue através de *Machine Learning* e *Deep Learning*: uma revisão sistemática

Predicting dengue cases through Machine Learning and Deep Learning: a systematic review

Predicción de casos de dengue a través del aprendizaje automático y el aprendizaje profundo: una revisión sistemática

Recebido: 12/08/2021 | Revisado: 16/08/2021 | Aceito: 18/08/2021 | Publicado: 22/08/2021

Ewerthon Dyego de Araújo Batista
ORCID: https://orcid.org/0000-0003-4993-9900
Universidade Estadual da Paraíba, Brasil
E-mail: edabew@gmail.com
Wellington Candeia de Araújo
ORCID: https://orcid.org/0000-0003-2102-7993
Universidade Estadual da Paraíba, Brasil
E-mail: wcandeia@uepb.edu.br
Romeryto Vieira Lira
ORCID: https://orcid.org/0000-0003-2567-0839
Instituto Federal de Educação, Ciência e Tecnologia da Paraíba, Brasil
E-mail: romeryto.lira@academico.ifpb.edu.br
Laryssa Izabel de Araújo Batista
ORCID: https://orcid.org/0000-0002-0188-9425
Universidade Federal da Paraíba, Brasil
E-mail: laryssa.izabel@gmail.com

Resumo

Introdução: a dengue é uma arbovirose causada pelo vírus DENV e transmitida para o homem através do mosquito *Aedes aegypti*. Atualmente, não existe uma vacina eficaz para combater todas as sorologias do vírus. Diante disso, o combate à doença se volta para medidas preventivas contra a proliferação do mosquito. Os pesquisadores estão utilizando *Machine Learning (ML)* e *Deep Learning (DL)* como ferramentas para prever casos de dengue e ajudar os governantes nesse combate. Objetivo: identificar quais técnicas e abordagens de ML e de DL estão sendo utilizadas na previsão de dengue. Métodos: revisão sistemática realizada nas bases das áreas de Medicina e de Computação com intuito de responder as perguntas de pesquisa: é possível realizar previsões de casos de dengue através de técnicas de ML e de DL, quais técnicas são utilizadas, onde os estudos estão sendo realizados, como e quais dados estão sendo utilizados? Resultados: após realizar as buscas, aplicar os critérios de inclusão, exclusão e leitura aprofundada, 14 artigos foram aprovados. As técnicas *Random Forest* (RF), *Support Vector Regression* (SVR), e *Long Short-Term Memory* (LSTM) estão presentes em 85% dos trabalhos. Em relação aos dados, na maioria, foram utilizados 10 anos de dados históricos da doença e informações climáticas. Por fim, a técnica *Root Mean Absolute Error* (RMSE) foi a preferida para mensurar o erro. Conclusão: a revisão evidenciou a viabilidade da utilização de técnicas de ML e de DL para a previsão de casos de dengue, com baixa taxa de erro e validada através de técnicas estatísticas.
Palavras-chave: Dengue; Previsão; Machine learning; Deep learning.

Abstract

Introduction: dengue is an arbovirus caused by the DENV virus and transmitted to humans through the Aedes aegypti mosquito. Currently, there is no effective vaccine to combat all serology of the virus. Therefore, the fight against the disease turns to preventive measures against the proliferation of the mosquito. Researchers are using Machine Learning (ML) and Deep Learning (DL) as tools to predict cases of dengue and help governments in this fight. Objective: to identify which ML and DL techniques and approaches are being used to predict dengue. Methods: systematic review carried out on the bases of the areas of Medicine and Computing in order to answer the research questions: it is possible to make predictions of dengue cases using ML and DL techniques, which techniques are used, where the studies are being performed, how and what data is being used? Results: after performing the searches, applying the inclusion, exclusion and in-depth reading criteria, 14 articles were approved. The Random Forest (RF), Support Vector Regression (SVR), and Long Short-Term Memory (LSTM) techniques are present in 85% of the works. Regarding the data, most were used 10 years of historical data on the disease and climate information. Finally, the Root Mean Absolute Error (RMSE) technique was preferred to measure the error. Conclusion: the review showed the feasibility of using ML and DL techniques to predict dengue cases, with a low error rate and validated through statistical techniques.
Keywords: Forecast; Machine learning; Deep learning.

Resumen

Introducción: el dengue es un arbovirus causado por el virus DENV y transmitido al ser humano a través del mosquito Aedes aegypti. Actualmente, no existe una vacuna eficaz para combatir todas las serologías del virus. Por tanto, la lucha contra la enfermedad se convierte en medidas preventivas contra la proliferación del mosquito. Los investigadores están utilizando Machine Learning (ML) y Deep Learning (DL) como herramientas para predecir casos de dengue y ayudar a los gobiernos en esta lucha. Objetivo: identificar qué técnicas y enfoques de LD y LD se están utilizando para predecir el dengue. Métodos: revisión sistemática realizada sobre las bases de las áreas de Medicina y Computación para dar respuesta a las preguntas de investigación: es posible realizar predicciones de casos de dengue utilizando técnicas de ML y DL, qué técnicas se utilizan, dónde se están realizando los estudios, ¿cómo y qué datos se utilizan? Resultados: luego de realizar las búsquedas, aplicando los criterios de inclusión, exclusión y lectura en profundidad, se aprobaron 14 artículos. Las técnicas Random Forest (RF), Support Vector Regression (SVR) y Long Short-Term Memory (LSTM) están presentes en el 85% de los trabajos. En cuanto a los datos, la mayoría se utilizaron 10 años de datos históricos sobre la enfermedad y la información climática. Finalmente, se prefirió la técnica de Root Mean Absolute Error (RMSE) para medir el error. Conclusión: la revisión mostró la viabilidad de utilizar técnicas de LD y LD para predecir casos de dengue, con una baja tasa de error y validadas mediante técnicas estadísticas.

Palabras clave: Dengue; Pronóstico; Aprendizaje automático; Aprendizaje profundo.

1. Introdução

Em circulação no Brasil desde 1981, a dengue foi considerada uma doença extinta, porém reapareceu e passou a ser classificada como endêmica (Câmara et al., 2007). Os principais sintomas da dengue são: febre alta, dores musculares, mal-estar, falta de apetite e dores de cabeça. Os casos mais graves da dengue podem causar hemorragias e levar o paciente a óbito (Graciano et al., 2017).

A arbovirose dengue pode ser originada por cinco variantes do vírus DENV (Mustafa et al., 2015) e, no Brasil, estão em circulação os sorotipos DENV-1, DENV-2, DENV-3 e DENV-4. Uma vez infectado por algum dos quatro tipos do vírus, o paciente continua vulnerável aos demais (Neto, Nascimento, Sousa, 2016). O mosquito *Aedes aegypti* encontra em países de clima tropical, como o Brasil, ambientes propícios para sua reprodução. Adicionalmente, problemas sociais e sanitários como, por exemplo, o descarte incorreto de lixo e o indevido lançamento de esgotos, aumenta o número de locais para as fêmeas do mosquito depositarem seus ovos (de Souza & Albuquerque, 2018).

No Brasil, existe a regulamentação da vacina Dengvaxia para combater a dengue. Contudo, a vacina está disponível apenas na rede particular e o seu uso é indicado, exclusivamente, para pessoas que já tiveram a doença (da Silveira, 2019). Logo, para combater a doença, os sistemas governamentais utilizam campanhas de conscientização e ações contra a proliferação do vetor (Ferreira et al., 2019). No ano de 2019, a *World Health Organization* (WHO) contabilizou cerca de 4.2 milhões de manifestações de dengue em todo o planeta. Anteriormente, esse mesmo órgão, emitiu um alerta classificando a dengue como uma das principais doenças para o ano de 2019. No Brasil, em 2019, devido ao aumento da circulação de um nova variante do vírus DENV-2, houve um novo surto de dengue com crescimento de 149% dos casos em alguns estados (Brasil, 2019; De Jesus et al., 2020).

A criação de ferramentas para prever dengue é uma tarefa complexa, pois vários fatores contribuem para o aparecimento e proliferação da doença. Entretanto, técnicas de *Machine Learning* e de *Deep Learning* vêm ajudando as pesquisas nessa área. Doni e Sasipraba (2020), conduziram um estudo na Índia, utilizando técnicas de *Deep Learning*, para analisar dados climáticos como temperatura, dados de precipitação e umidade. O trabalho conseguiu prever casos de dengue com 89% de acurácia. Outro exemplo da utilização de predição através de ML e de DL ocorreu na Tailândia. Em seu trabalho, os pesquisadores (Puengpreeda, Yhusumrarn, Sirikulvadhana, 2020) utilizaram dados climáticos e informações de pesquisas realizadas no Google a respeito de dengue.

Ao realizar pesquisa sobre *Machine Learning* e *Deep Learning* para a doença em estudo, na base *Scopus*, mais de 250 artigos foram retornados. Esse fato demonstra a atenção dada pela ciência ao tema. Portanto, combinando a importância da ciência e os problemas de saúde aqui listados, é justificável realizar uma revisão sistemática e verificar quais temas,

abordagens e técnicas estão sendo utilizadas nessa área.

O objetivo da pesquisa é, através de uma revisão sistemática, verificar a viabilidade de utilizar técnicas de *Machine Learning* e de *Deep Learning* durante a previsão de dengue, quais técnicas são utilizadas, onde os estudos estão sendo realizados, como e quais dados estão sendo utilizados na previsão e, por fim, quais técnicas estão demonstrando melhores resultados.

2. Metodologia

Esta revisão foi estruturada com base no modelo definido por Levac (Levac, Colquhoun, O'brien, 2010). Em sua metodologia, a autora elencou cinco fases obrigatórias e uma fase opcional para a confecção de uma revisão sistemática. Para este trabalho, adotamos apenas as fases obrigatórias, são elas: 1 – Identificar a questão da pesquisa, 2 – Identificar os estudos relevantes, 3 – Selecionar os estudos, 4 – Mapear os dados e 5 – Coleta, sumarização e relato dos dados. Além da metodologia citada anteriormente, os pesquisadores deste trabalho utilizaram o *software* StArt (Fabbri et al., 2016) como forma de auxílio na definição e execução do protocolo de pesquisa.

2.1 Identificação da questão de pesquisa

O objetivo principal desta revisão é verificar a viabilidade da utilização de técnicas de *Machine Learning* e de *Deep Learning* na previsão de casos da doença dengue. Objetivando um melhor direcionamento do assunto, a equipe subdividiu o problema de pesquisa em quatro frentes: análise de técnicas, quais dados foram utilizados na previsão, quais abordagens foram utilizadas, como foram feitas as validações dos resultados e, por fim, quais foram os melhores resultados. Feito isso, foram criadas as seguintes perguntas derivadas:

1. Quais técnicas de *Machine Learning* e de *Deep Learning* são utilizadas nas previsões?
2. Em qual país foi realizado o estudo? Como foram coletados os dados?
3. Quantos anos de dados foram utilizados nos modelos e quais itens foram considerados na criação dos modelos? Exemplo: Fatores climáticos, econômicos, dados de redes sociais, entre outros.
4. Como foi feita a validação dos modelos? Quais técnicas estatísticas o estudo utilizou na avaliação?
5. Qual técnica ou combinação de técnicas obtiveram os melhores resultados?

2.2 Identificação de estudos relevantes

Após a definição da questão primária da pesquisa e suas derivações, o próximo passo foi definir quais bases de dados seriam relevantes para a revisão. Durante o levantamento dos estudos, foram utilizadas as principais bases eletrônicas nas áreas da Saúde e da Ciência da Computação: *Scopus*, *IEEE Xplore*, *PubMed*, *ACM Digital Library* e *Web of Science*. Outras bases, como *Cochrane*, foram testadas, porém não tiveram adequada indexação de artigos para o tema desta revisão ou o acesso aos artigos estava limitado.

Definidas as bases, a equipe iniciou o estudo sobre os termos para a formação das *strings* de busca. Os termos utilizados nas *strings* de busca fazem referência à previsão, à *Machine Learning* (aqui entende-se *Deep Learning* como sendo um tipo de ML) e à dengue. Segue a listagem das palavras-chave, o conjunto e o escopo abordado por cada uma delas: *predict** (referenciam termos como *predict*, *prediction*, *predicted*), *forecast** (contemplando as palavras *forecast*, *forecasting*, *forecasted*), *Machine Learning* (não há variação para esse termo), *Deep Learning* (não há variação para esse termo) e dengue (referência à dengue *fever*, *fever hemorrhagic* dengue e, em alguns países, apenas dengue).

Utilizando a estratégia de *rounds* e a variação nos operadores lógicos, as combinações das *strings* foram testadas e, a cada *round*, pequenos ajustes foram feitos. Para a avaliar a qualidade da *string*, a equipe elencou dezessete artigos como indispensáveis no retorno das bases. Caso algum desses artigos não fosse retornado, a alteração na *string* era descartada. Por fim, após quatro *rounds*, a equipe chegou no consenso e definiu as seguintes *strings*:

- Base *Scopus*: *TITLE-ABS-KEY ((predict* or forecast*) AND ("machine learning" or "deep learning") AND (dengue))*;
- Base *IEEE: ((("Full Text & Metadata":predict* or forecast*) AND "Full Text & Metadata":machine learning or deep learning) AND "Full Text & Metadata":dengue);*
- Base *PubMed: All Fields ((predict* or forecast*) AND ("machine learning" or "deep learning")) AND (dengue)*;
- Base *ACM: [All: predict or forecast] AND [All: machine learning or deep learning] AND [All: dengue] AND [All: predict* or forecast*] AND [All: machine learning or deep learning] AND [All: dengue]*;
- Base *Web of Science: TOPIC: (predict* or forecast*) AND TOPIC: (machine learning or deep learning) AND TOPIC: (dengue).*

2.3 Seleção dos estudos

Machine Learning e de *Deep Learning* são temas atuais, estão em constante evolução, e são cada vez mais empregados na resolução de problemas complexos. Adicionalmente, são amplamente pesquisados nas áreas de Medicina e de Ciência da Computação. Primordialmente, foram selecionados artigos contendo previsões de casos de dengue realizados através de técnicas de ML e de DL. A equipe definiu os critérios de inclusão (INC) e exclusão (EXC) para ter um melhor direcionamento junto à pergunta desta pesquisa. São eles:

- INC01 – O estudo utiliza técnicas de *Machine Learning*;
- INC02 – O estudo utiliza técnicas de *Deep Learning*;
- INC03 – O estudo foi validado estatisticamente;
- INC04 – Existe no estudo a comparação e a utilização de mais de uma técnica ou modelo de ML ou DL;
- INC05 – O estudo contém previsões de casos de dengue;
- INC06 – O artigo deve ser escrito no idioma inglês ou português;
- EXC01 – Publicações além de 5 anos;
- EXC02 – Publicações que utilizam apenas um método de *Machine Learning* ou *Deep Learning*;
- EXC03 – Publicações que utilizam apenas técnicas clássicas da estatística;
- EXC04 – Estudos com baixo grau de validação estatística;
- EXC05 – Trabalhos sem resultados de previsão ou internação de casos de dengue;
- EXC06 – Estudos que não utilizaram bases oficiais ou formais;
- EXC07 – Estudos utilizados para classificar a doença de acordo com os sintomas;
- EXC08 – Estudos não primários;
- EXC09 – Problema de acesso ou não acesso total aos dados;
- EXC10 – Linguagem diferente de inglês ou português.

Após a definição dos critérios, dois pesquisadores iniciaram a triagem dos artigos utilizando os conceitos e as técnicas descritos pelo PRISMA (Moher et al., 2009). Com intuito de evitar influência nos resultados, cada um deles fez a classificação

individualmente e sem saber o resultado do outro pesquisador. Durante a seleção dos estudos, as seguintes atividades foram realizadas: 1 – eliminar os artigos duplicados, 2 – leitura rápida do título, *abstract* e resultados dos artigos e 3 – aplicar os critérios de inclusão e exclusão definidos. As exclusões por não obediência aos critérios tiveram o seu registro realizado. Finalizado esse processo, houve comparação dos resultados obtidos. Os artigos com opiniões conflitantes foram submetidos ao crivo de um terceiro pesquisador. O terceiro pesquisador discutiu os pontos levantados preliminarmente, e, por fim, deferiu o parecer final.

2.4 Mapeamento dos dados

Na etapa de mapear os dados, a equipe definiu as informações a serem extraídas dos artigos. Inicialmente, ocorreu uma reunião para a equalização das informações a serem extraídas. Como resultado da reunião, os pesquisadores entraram em consenso para extrair: dados bibliográficos, quais técnicas de ML ou de DL foram utilizadas, qual país o estudo foi feito, como os dados foram coletados e se são oriundos de bases oficiais, quantos anos de dados foram utilizados nas amostras, quais técnicas estatísticas foram utilizadas na validação das predições e, enfim, qual técnica obteve melhor performance. Ademais, os revisores acresceram as suas considerações em relação aos artigos.

À medida que os artigos foram lidos, os revisores, individualmente, incluíram as informações na ferramenta StART. A etapa final do mapeamento foi sumarizar em um arquivo de planilha eletrônica as extrações. Aqui, mais uma vez, surgiram conflitos. Novamente, as divergências foram resolvidas por um terceiro avaliador.

2.5 Coleta, sumarização e relato dos dados

Durante a coleta, sumarização e relato dos resultados, (Levac, Colquhoun, O'brien, 2010) sugere a quebra nos tópicos *analysis*, *reporting* e *implications*. No primeiro, a equipe utilizou a análise quantitativa. Durante a análise quantitativa, os artigos foram classificados e verificada a forma de resposta às subquestões da pesquisa. O relato da revisão foi feito através de quadros contendo as respostas e os dados coletados. Como último passo desta revisão, a equipe emitiu recomendações em relação ao tema estudado e aos achados da revisão.

2.6 Registro de protocolo

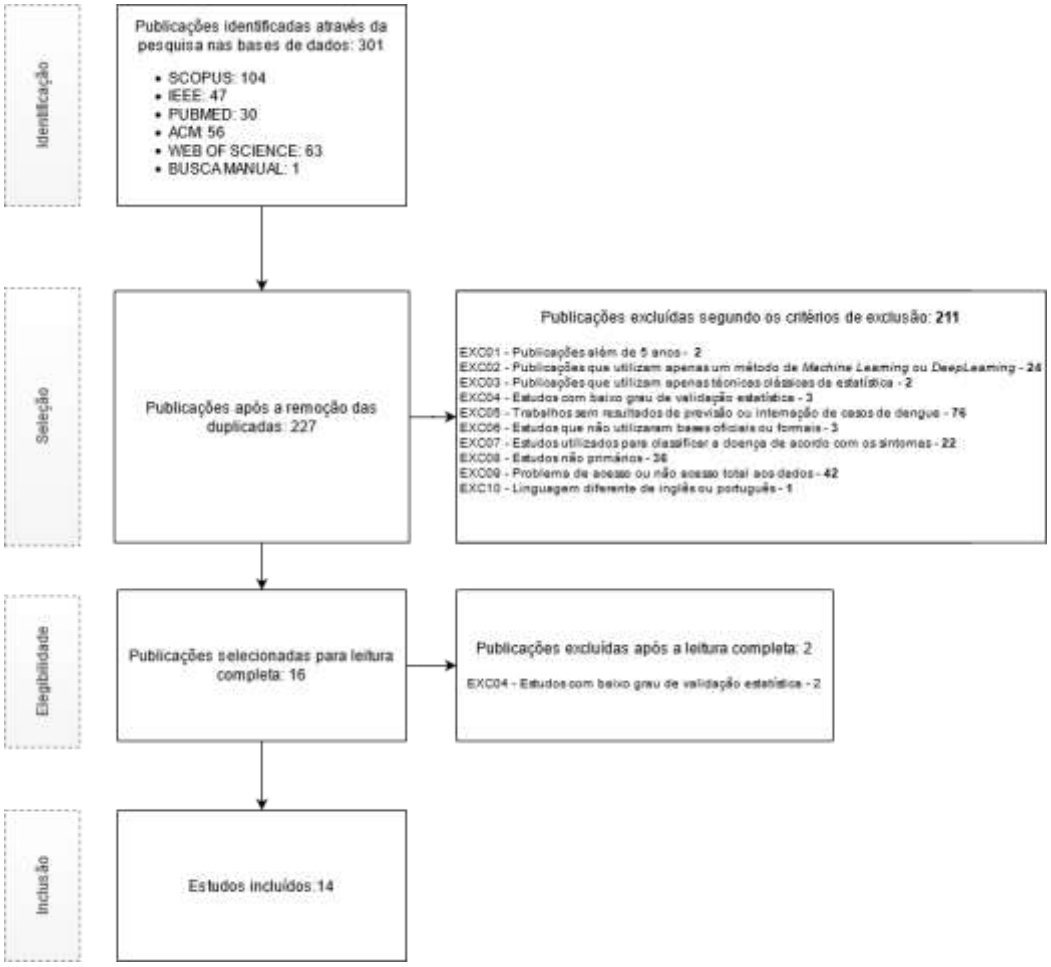
Esta revisão sistemática foi guiada através da recomendação *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) (Moher et al., 2009). O protocolo de estudo foi registrado no *Open Science Framework* (OSF) e está disponível através do *link*: https://osf.io/fqa57.

3. Resultados e Discussão

Ao executar o protocolo de revisão, os pesquisadores identificaram 300 artigos. Adicionalmente, 1 artigo foi adicionado manualmente. Após a exclusão dos duplicados, esse número reduziu a 227. Durante as leituras de título, resumo, resultados e aplicando os critérios de inclusão e exclusão, foram selecionados 16 artigos para a leitura aprofundada. Nessa etapa, dois artigos foram excluídos e, enfim, o número de trabalhos aprovados ficou em 14 artigos.

A Figura 1 detalha o processo de seleção, listando o quantitativo de artigos por base e a minúcia das exclusões.

Figura 1 - Resultado da seleção dos estudos após a execução do protocolo da revisão sistemática**.**



Fonte: Autores (2021)

As repostas para o questionamento “Quais técnicas de *Machine Learning* e *Deep Learning* são utilizadas nas previsões?” estão representadas no Quadro 1. Como se pode observar, há uma grande variação no emprego das técnicas de *Machine Learning* e de *Deep Learning*. Contudo, técnicas como *Support Vector Machine for regression* (SVR), *Random forest* (RF), *Long short-term memory* (LSTM) estão presentes em 12 dos 14 trabalhos.

Quadro 1 - Listagem das técnicas de *Machine Learning* e *Deep Learning* utilizada para os trabalhos aprovados.

Estudo	**Técnicas ML e DL**
(Manogaran & Lopez, 2018)	*Gaussian process regression* (GPR), RF, SVR, *Multiple regression* (MR).
(Appice et al., 2020)	*AUTOencoding based TIme series Clustering with Nearest. Neighbour (*AutoTiC-NN), *K-Nearest Neighbourhood* (KNN), SVR, *Autoregressive integrated moving average* (ARIMA), *M5 regression* (M5).
(Dhaka & Singh, 2020)	RF, *Decision Tree* (DT), *Multiple linear regression* (MLR) e SVR.
(Mishra, Tiwari, Ajaymon, 2019)	*Neural Network* (NN), RF, *Boosted Trees, Least absolute shrinkage and selection operator* (LASSO), *Ridge*, *Extreme Gradient Boosting* (XGBoost) e SVR.
(Guo et al., 2017)	SVR, *Gradient boosted* (GBM), LASSO e *Generalized additive model* (GAM).
(Raju et al., 2019)	SVR, *Ridge* e *Linear Regression* (LR).
(Xu et al., 2020)	LSTM, *Back propagation neural network* (BPNN), *Generalized additive model* (GAM), SVR e *Gradient boosted* (GBM).
(Kerdprasop, Kerdprasop, Chuaybamroong, 2019)	*Chi-squared automatic interaction detection* (CHAID), LR, GLM, *Artificial Neural Network* (ANN) e SVR.
(Pham et al., 2018)	*Genetic Algorithm Enhanced Recurrent Neural Network* (GA_RNN), LR e DT.
(Mussumeci & Codeço Coelho, 2020)	LSTM, RF e LASSO.
(Doni & Sasipraba, 2020)	LSTM, SVR, XGboost, RF, GAM, BPNN
(Carvajal et al., 2018)	RF, GAM e GB.
(Dharmawardana et al., 2018)	ANN E XGBoost.
(Puengpreeda, Yhusumrarn, Sirikulvadhana, 2020)	RF e *Ridge*.

Fonte: Autores (2021).

O Quadro 2 contém os esclarecimentos para os questionamentos 2 e 3, respectivamente, “Em qual país foi realizado o estudo? Como foram coletados os dados?” e “Quantos anos de dados foram utilizados nos modelos e quais itens foram considerados na criação dos modelos”. Para cada pesquisa, foram elencados o país alvo, quantos anos de dados foram utilizados, quais informações foram utilizadas e, por fim, as suas origens.

Quadro 2 - Levantamento dos países, período, dados utilizados e suas origens para os trabalhos selecionados.

Estudo	País	Período	Dados utilizados	Origem dos dados
(Manogaran & Lopez, 2018)	Índia	1998 a 2006	Meteorológicos e epidemiológicos	Ministério da saúde
(Appice et al., 2020)	México	1985 a 2010	Epidemiológicos e temperatura	Governo do México
(Dhaka & Singh, 2020)	Índia	2013 a 2017	Epidemiológicos e climáticos	Ministério da saúde da Índia
(Mishra, Tiwari Ajaymon, 2019)	Peru	Não informado	Epidemiológicos, sociais e climáticos	*Dataset* dos USA
(Guo et al., 2017)	China	2011 a 2014	Epidemiológicos	Sistema nacional de vigilância da China
(Raju et al., 2019)	Índia	2001 a 2018	Epidemiológicos e climáticos	Dados do governo de Kerala
(Xu et al., 2020)	China	2005 a 2018	Meteorológicos	Dados do Centro nacional da China
(Kerdprasop, Kerdprasop, Chuaybamroong, 2019)	Tailândia	2003 a 2017	Epidemiológicos e climáticos	Ministério da saúde
(Pham et al., 2018)	Malásia	2002 a 2012	Epidemiológicos e climáticos	Câmara municipal de Kuala Lumpur
(Mussumeci & Codeço Coelho, 2020)	Brasil	2010 a 2018	Epidemiológicos e climáticos	Base InfoDengue
(Doni & Sasipraba, 2020)	Índia	2015 a 2018	Epidemiológicos e climáticos	Governo da Índia
(Carvajal et al., 2018)	Filipinas	2009 a 2013	Climáticos	Ministério da saúde
(Dharmawardana et al., 2018)	Sri Lanka	2012 a 2017	Epidemiológicos e de telefonia móvel	Centro nacional de controle a dengue
(Puengpreeda, Yhusumrarn, Sirikulvadhana, 2020)	Tailândia	2014 a 2018	Epidemiológicos e Google *Trend topics*	Google e Departamento de meteorologia

Fonte: Autores (2021).

Na sua grande maioria (79%), as pesquisas foram realizadas nos países da Ásia. O restante (11%) correspondeu a estudos realizados nas Américas. Com exceção do trabalho conduzido por (Mishra, Tiwari, Ajaymon, 2019), os artigos informaram o período de anos utilizado em seus modelos de predição. Ainda sobre os dados, a grande parte deles utilizou dados epidemiológicos ou climáticos fornecidos por órgãos oficiais do país onde o estudo foi conduzido. Os trabalhos de (Dharmawardana et al., 2018) e de (Puengpreeda, Yhusumrarn, Sirikulvadhana, 2020) chamam atenção por inovar e utilizar, respectivamente, dados de telefonia móvel e de pesquisas realizadas no Google.

Os resultados obtidos pelos estudos, qual abordagem foi utilizada para mensurar a taxa de erro, os resultados para cada técnica e, enfim, a técnica vencedora, estão listados no Quadro 3. Este quadro apresenta as repostas aos questionamentos 4 (“Como foi feita a validação dos modelos? Quais técnicas estatísticas o estudo utilizou na avaliação?”) e 5 (“Qual técnica ou combinação de técnicas obtiveram os melhores resultados?”).

Quadro 3 - Quadro contendo a forma de validação, resultados e técnica vencedora de cada estudo.

Estudo	Forma de Validação	Resultados	Vencedor
(Manogaran & Lopez, 2018)	RMSE	MR – RMSE 0,525 GPR – RMSE 0,281 SVR – RMSE 0,352 RF – RMSE 0,323	GPR
(Appice et al., 2020)	RMSE	AutoTiC-NN – RMSE 5,18 KNN – RMSE 8,20 SVR – RMSE 19,62 ARIMA – RMSE 12,23 M5- RMSE 115,34	AutoTiC-NN
(Dhaka & Singh, 2020)	*Sum of Absolute Difference* (SAD)	RF – SAD 91560,56 DT – SAD 101911,00 MLR – SAD 80901,42 SVR – SAD 11376,1	MLR
(Mishra, Tiwari, Ajaymon, 2019)	*Mean absolute error* (MAE)	NN – MAE 25,621 RF – MAE 25,012 *Boosted Trees* – MAE 24,985 LASSO – MAE 27,045 Ridge – MAE 28,052 XGBoost – MAE 24,802 SVR – MAE 25,011	XGBoost
(Guo et al., 2017)	RMSE	SVR – RMSE 0,2681 LASSO – RMSE 2,0621 GAM – RMSE 4,4973 GBM – RMSE 3,4527	SVR
(Raju et al., 2019)	MAE	SVR – MAE 180,61 Ridge – MAE 366,570 LR – MAE 190,04	SVR
(Xu et al., 2020)	RMSE	LSTM – RMSE 36,50 BPNN – RMSE 48,61 GAM – RMSE 41,95 SVR – RMSE 44,37 GBM – RMSE 42,33	LSTM
(Kerdprasop, Kerdprasop, Chuaybamroong, 2019)	Erro preditivo	CHAID – Erro preditivo 0,275 LR – Erro preditivo 0,598 GLM – Erro preditivo 0,598 ANN – Erro preditivo 0,901 SVR – Erro preditivo 1,034	CHAID
(Pham et al., 2018)	RMSE	GA-RNN – RMSE 13,06 LR – RMSE 22,99 DT – RMSE 34,89	GA-RNN
(Mussumeci & Codeço Coelho, 2020)	RMSE	LSTM – RMSE 0,45 RF – RMSE 0,47 LASSO – RMSE 0,50	LSTM
(Doni & Sasipraba, 2020)	RMSE	LSTM – RMSE 42,00 SVR – RMSE 49,00 XGboost – RMSE 48,00 RF – RMSE 51,00 GAM – RMSE 53,00 BPNN – RMSE 48,00	LSTM
(Carvajal et al., 2018)	RMSE	RF– RMSE 0,29 GAM – RMSE 0,33 GB – RMSE 0,30	RF
(Dharmawardana et al., 2018)	RMSE	ANN – RMSE 0,67 XGBoost – RMSE 0,54	XGBoost
(Puengpreeda, Yhusumrarn, Sirikulvadhana, 2020)	MAE	RF – MAE 10,98 Ridge – MAE 16,44	RF

Fonte: Autores (2021).

Com exceção de (Raju et al., 2019), as melhores técnicas produzem resultados de previsões bem similares aos valores reais. As técnicas SVR, RF, LSTM obtiveram os melhores resultados e venceram em 50% dos artigos estudados. Vale ressaltar que, para todas as técnicas de medição de erro aqui abordadas, quanto menor for o seu valor melhor é o resultado da previsão.

Não foi encontrado na literatura um consenso de quais técnicas ou algoritmos devem ser utilizados durante as atividades de previsão. Inúmeros fatores influenciam positivamente ou negativamente nos resultados de um modelo de previsão. Entre eles, pode-se citar a quantidade de dados, como eles são tratados e quais foram utilizados para a previsão. Contudo, os resultados desta revisão apontam, que, ao menos, uma das técnicas RF, SVR e LSTM são utilizadas em 85% das pesquisas.

As técnicas de *Deep Learning*, como é o caso da LSTM, vêm ganhando espaço nos problemas de previsão frente às demais técnicas de *Machine Learning*. Cientistas atribuem o sucesso da LSTM devido à capacidade de possuir uma memória sobre o que já foi calculado, bem como capacidade de decidir quais dados precisam ser utilizados no futuro ou esquecidos (Doni & Sasipraba, 2020). Essa capacidade credencia as LSTMs na utilização de modelagem de problemas complexos como a previsão de dengue, visto que, vários fatores contribuem para a proliferação da doença: climáticos, sanitários, econômicos e sociais.

Em relação aos locais dos estudos desta revisão, eles foram realizados em países da Ásia e Américas. Embora exista uma disparidade espacial entre os continentes, as características do clima e chuva se assemelham: altos níveis de precipitações, temperaturas elevadas e elevados níveis de umidade. Combinando com registros epidemiológicos, os dados climáticos mostraram a sua capacidade em participar da criação de modelos de previsão de dengue.

Ademais, o trabalho conduzido por (Dharmawardana et al., 2018) mostra uma abordagem interessante sobre a migração de pessoas entre os países e a sua capacidade de proliferação de doenças. O estudo foi conduzido em 2018 e, a partir de 2020, o fenômeno foi bastante observado durante a epidemia do coronavírus. Sendo assim, fica a pergunta: quantas vidas poderiam ter sido salvas na epidemia do COVID19 caso medidas sanitárias e efetivos rastreios da doença fossem feitos antes da sua infestação pandêmica?

Em média, as previsões foram feitas com 10 anos de informação base e a coleta dos dados foi realizada em fontes oficiais dos governos daqueles países. Logo, problemas relacionados à qualidade ou viés dos dados, podem ser minimizados, pois os dados foram oriundos de fontes oficiais e em uma quantidade significativa. Sobre a significância estatísticas dos resultados obtidos, todos os trabalhos aprovados possuem técnicas de validação estatística atreladas à análise dos seus resultados. Artigos sem validação estatística ou com baixa qualidade no reporte foram excluídos desta revisão.

Outro achado da revisão é a grande aceitação da técnica RMSE para a validação dos resultados previstos pelos modelos, sendo utilizado em 64% dos artigos. Adicionalmente, a utilização do RMSE demonstra com mais fidelidade as discrepâncias entre o resultado previsto versus o resultado esperado (Carvajal et al., 2018).

4. Considerações Finais

A partir desta revisão, pode-se inferir que é possível prever, com baixa taxa de erro, casos de dengue através de técnicas de *Machine Learning* e de *Deep Learning*. A grande maioria dos estudos envolvendo ML e DL na previsão de dengue ocorreu em países asiáticos, embora também tenhamos trabalhos nas américas.

Apesar da existência de uma gama de técnicas de ML e de DL, podemos destacar as técnicas RF, SVR e LSTM como recorrentes nos estudos. Embora cada estudo tenha a sua particularidade, vale destacar os ótimos resultados da LSTM. Sempre que usada, essa técnica saiu vitoriosa.

É perceptível também o padrão de utilização de dados com período de, em média, 10 anos. Outro destaque desta revisão é a padronização da utilização dos dados para a confecção dos modelos. Na maioria deles, foi usado dados históricos e

climáticos. Finalmente, em relação à validação estatística, os pesquisadores têm preferência por medir através do RMSE.

Ainda que esta pesquisa tenha sido ampla, realizada em bases de referências e conduzida por pesquisadores experientes, vale ressaltar que, ao realizar revisão com base em artigos publicados, os resultados produzidos pela revisão são direcionados por eles. Por fim, devido às similaridades entre as variantes dos vírus em circulação e semelhança entre os climas, boa parte dos trabalhos aqui listados podem ser reproduzidos no Brasil.

Como sugestões para trabalhos futuros, indicamos a realização da pesquisa em outras bases, como, por exemplo, a *Springer* e realizar variações na *string* de busca. Por fim, sugerimos uma nova execução do protocolo definido por este artigo com intuito de verificar as novas soluções utilizadas para a predição de casos de dengue.

Referências

Appice, A., Gel, Y. R., Iliev, I., Lyubchich, V. & Malerba, D. (2020). A multi-stage machine learning approach to predict dengue incidence: a case study in mexico. *Ieee access*, 8, 52713–52725.

Brasil. (2019). Ministério da saúde alerta para aumento de 149% dos casos de dengue no país. Ministério da saúde, brasil, p. 2020.

Câmara, F. P., Theophilo, R. L. G., Santos, G. T. D., Pereira, S. R. F. G., Câmara, D. C. P., & Matos, R. R. C. D. (2007). Estudo retrospectivo (histórico) da dengue no Brasil: características regionais e dinâmicas. *Revista da Sociedade Brasileira de Medicina Tropical*, *40*, 192-196.

Carvajal, T. M., Viacrusis, K. M., Hernandez, L. F. T., Ho, H. T., Amalin, D. M., & Watanabe, K. (2018). Machine learning methods reveal the temporal pattern of dengue incidence using meteorological factors in metropolitan Manila, Philippines. *BMC infectious diseases*, *18*(1), 1-15.

da Silveira, L. T. C., Tura, B., & Santos, M. (2019). Systematic review of dengue vaccine efficacy. *BMC infectious diseases*, *19*(1), 1-8.

de Jesus, J. G., Dutra, K. R., Sales, F. C. D. S., Claro, I. M., Terzian, A. C., Candido, D. D. S., & Faria, N. R. (2020). Genomic detection of a virus lineage replacement event of dengue virus serotype 2 in Brazil, 2019. *Memórias do Instituto Oswaldo Cruz*, 115.

de Souza, R. F., & da Cunha Albuquerque, A. R. (2018). Geografia Da Dengue: Uma Análise Das Políticas De Controle E Monitoramento Do Aedes Aegypti Em Manaus/Geography of Dengue: an analysis of the control and monitoring policies of Aedes aegypti in Manaus. *Revista Geonorte*, *9*(31), 68-76.

Dhaka, A., & Singh, P. (2020, January). Comparative Analysis of Epidemic Alert System using Machine Learning for dengue and Chikungunya. In *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 798-804). IEEE.

Dharmawardana, K. G. S., Lokuge, J. N., Dassanayake, P. S. B., Sirisena, M. L., Fernando, M. L., Perera, A. S., & Lokanathan, S. (2017, December). Predictive model for the dengue incidences in Sri Lanka using mobile network big data. In *2017 IEEE International Conference on Industrial and Information Systems (ICIIS)* (pp. 1-6). IEEE.

Doni, A. R., & Sasipraba, T. (2020). LSTM-RNN Based Approach for Prediction of Dengue Cases in India. *Ingénierie des Systèmes d'Information*, *25*(3).

Fabbri, S., Silva, C., Hernandes, E., Octaviano, F., Di Thommazo, A., & Belgamo, A. (2016, June). Improvements in the StArt tool to better support the systematic review process. In *Proceedings of the 20th international conference on evaluation and assessment in software engineering* (pp. 1-5).

Ferreira, V. M., Nunes, R. C., Ferreira, J. M. S., & Herrera, K. M. S. (2019). Um mosquito e três doenças: ação de combate ao Aedes aegypti e conscientização sobre Dengue, Chikungunya e Zika em Divinópolis/MG, BRASIL. *Revista Brasileira de Extensão Universitária*, *10*(2), 49-54.

Graciano, A. R., de Assis, L. P. F., Cozer, A. M., Amâncio, V. C., & de Oliveira, J. M. R. (2017). Morbimortalidade da dengue em idosos no Brasil-Dengue morbidity and mortality in elderly in Brazil. *Revista Educação em Saúde*, *5*(1), 56-65.

Guo, P., Liu, T., Zhang, Q., Wang, L., Xiao, J., Zhang, Q., & Ma, W. (2017). Developing a dengue forecast model using machine learning: A case study in China. *PLoS neglected tropical diseases*, *11*(10), e0005973.

Kerdprasop, K., Kerdprasop, N., & Chuaybamroong, P. (2019, December). Forecasting Dengue Incidence with the Chi-squared Automatic Interaction Detection Technique. In *Proceedings of the 2019 2nd Artificial Intelligence and Cloud Computing Conference* (pp. 37-42).

Levac, D., Colquhoun, H., & O'Brien, K. K. (2010). Scoping studies: advancing the methodology. *Implementation science*, *5*(1), 1-9.

Manogaran, G., & Lopez, D. (2018). A Gaussian process based big data processing framework in cluster computing environment. *Cluster Computing*, *21*(1), 189-204.

Mishra, V. K., Tiwari, N., & Ajaymon, S. L. (2019, December). Dengue disease spread prediction using twofold linear regression. In *2019 IEEE 9th International Conference on Advanced Computing (IACC)* (pp. 182-187). IEEE.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Prisma Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS medicine*, *6*(7), e1000097.

Mussumeci, E., & Coelho, F. C. (2020). Large-scale multivariate forecasting models for Dengue-LSTM versus random forest regression. *Spatial and Spatio-temporal Epidemiology*, *35*, 100372.

Mustafa, M. S., Rasotgi, V., Jain, S., & Gupta, V. J. M. J. A. F. I. (2015). Discovery of fifth serotype of dengue virus (DENV-5): A new public health dilemma in dengue control. *Medical journal armed forces India*, *71*(1), 67-70.

Neto, A. S. L., do Nascimento, O. J., & de Sousa, G. D. S. (2016). Dengue, zika e chikungunya-desafios do controle vetorial frente à ocorrência das três arboviroses-parte I. *Revista Brasileira em Promoção da Saúde*, *29*(3), 305-312.

Pham, D. N., Aziz, T., Kohan, A., Nellis, S., Khoo, J. J., Lukose, D., & Ong, H. H. (2018, October). How to efficiently predict dengue incidence in kuala lumpur. In *2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)* (pp. 1-6). IEEE.

Puengpreeda, A., Yhusumrarn, S., & Sirikulvadhana, S. (2020). Weekly Forecasting Model for Dengue Hemorrhagic Fever Outbreak in Thailand. *Engineering Journal*, *24*(3), 71-87.

Raju, N. G., Krishna, P. G., Manognya, K., Kiran, G. R., Rohit, P., & Likhith, K. (2019, July). Evolution of predictive model for Dengue incidence by using machine learning algorithms. In *2019 International Conference on Communication and Electronics Systems (ICCES)* (pp. 51-59). IEEE.

Xu, J., Xu, K., Li, Z., Meng, F., Tu, T., Xu, L., & Liu, Q. (2020). Forecast of dengue cases in 20 Chinese cities based on the deep learning method. *International journal of environmental research and public health*, *17*(2), 453.