UNIVERSIDADE NOVE DE JULHO - UNINOVE
PROGRAMA DE PÓS GRADUAÇÃO EM INFORMÁTICA E GESTÃO DO CONHECIMENTO - PPGI

MARCO ANTONIO RUSSI FONTOURA

MODELAGEM PREDITIVA DE SURTOS EPIDÊMICOS USANDO REDES NEURAIS LSTM: UMA APLICAÇÃO PARA A COVID-19

São Paulo
2024

MARCO ANTONIO RUSSI FONTOURA

MODELAGEM PREDITIVA DE SURTOS EPIDÊMICOS USANDO REDES NEURAIS LSTM: UMA APLICAÇÃO PARA A COVID-19

Exame de defesa apresentado ao Programa de Pós-Graduação em Informática e Gestão do Conhecimento (PPGI) da Universidade Nove de Julho - UNINOVE, como parte dos requisitos para a obtenção do título de Mestre em Informática e Gestão do Conhecimento.

Prof. Orientador: Dr(a). Pedro Henrique Triguis Schimit

Linha de pesquisa: LP1 - Modelagem e Otimização Computacional

São Paulo
2024

DEDICATÓRIA

Dedico este trabalho à minha família, sobretudo minha avó, mãe e tia, cujos amor e apoio incondicional foram fundamentais para minha jornada. Sempre me incentivaram a estudar e alcançar voos mais altos. A cada um de vocês, meu eterno agradecimento por estarem sempre ao meu lado, acreditando em mim nos momentos mais difíceis e celebrando comigo as conquistas.

À minha esposa Alinne, que sempre esteve ao meu lado e me apoiou em cada sonho e objetivo, independente do sacrifício que fosse necessário ser feito.

Aos meus filhos, Isabella e Chris, que me proporcionam momentos de alegria e inspiração. Por eles, luto qualquer batalha.

A todos que, de maneira direta ou indireta, contribuíram para que eu pudesse alcançar este objetivo, a minha mais sincera gratidão.

AGRADECIMENTOS

Agradeço, primeiramente, a Deus, por me dar força, coragem e sabedoria para enfrentar os desafios durante essa jornada.

À minha família, que sempre me apoiou incondicionalmente, com amor e compreensão, e foram fundamentais para minha formação.

Ao meu orientador, Pedro Henrique Triguis Schimit, por sua orientação, paciência e apoio constante durante toda a realização da minha pesquisa. Seu conhecimento e experiência foram essenciais para o sucesso deste trabalho.

A todos os professores do Programa de Pós-Graduação em Informática e Gestão do Conhecimento da Universidade Nove de Julho (UNINOVE), por compartilharem seu conhecimento e por contribuírem para minha formação acadêmica.

Aos colegas de pesquisa e amigos, que sempre estiveram presentes com palavras de incentivo, apoio e, principalmente, boas conversas que tornaram os momentos de trabalho mais leves. Uma citação também ao Wellington de Lima Barreto, que além de colega foi coorientado junto a mim pelo Professor Pedro Schmidt, cujo foi possível ter bastante parceria em nossas pesquisa.

À Universidade Nove de Julho (UNINOVE), pela oportunidade da concessão da bolsa de estudos, que permitiu o desenvolvimento deste trabalho e proporcionou um ambiente acadêmico enriquecedor.

Por fim, a todos que, de alguma forma, contribuíram para a realização deste trabalho e para o meu crescimento pessoal e profissional. A todos vocês, minha gratidão eterna.

EPÍGRAFE

"A mente que se abre a uma nova ideia jamais voltará ao seu tamanho original."

Albert Einstein (1915)
Físico.

RESUMO

A pandemia causada pela COVID-19, nos seus três anos iniciais, resultou em uma série temporal de difícil predição devido a ações de políticas públicas e ao surgimento de novas variantes. Restrições de locomoção, novas variantes, vacinas e diferenças culturais resultaram em estágios distintos de disseminação do vírus, dificultando a aplicação de um único modelo de predição epidemiológica para prever toda a série temporal. Com isso, diferentes modelos foram necessários para os estágios da pandemia. Com base na aplicação de redes neurais em estudos epidemiológicos para predição de séries temporais, este trabalho busca desenvolver uma metodologia baseada em uma rede neural *Long Short-Term Memory* (LSTM) multicamadas, capaz de ser aplicada a quase três anos de dados da pandemia para prever a quantidade de novos casos diários da doença. A metodologia aplica uma série de testes, combinando diferentes dados de entrada da rede, como casos e vacinações diárias, buscando prever a quantidade de casos para alguns dias no futuro. Os experimentos foram aplicados ao Brasil e a outros países, apresentando bons resultados para previsões com quinze dias de antecedência, com potencial para identificar mudanças de tendência na linha temporal de casos diários. Essa capacidade é útil para detectar o início de novas ondas de contaminação, contribuindo para sistemas de alerta de órgãos de saúde pública e tomada de decisão imediata.

Palavras-chave: análise de séries temporais; aprendizado de máquina; COVID-19; modelagem epidemiológica; redes LSTM.

ABSTRACT

The COVID-19 pandemic, in its initial three years, resulted in a time series that was difficult to predict due to public policy actions and the emergence of new variants. Mobility restrictions, new variants, vaccines, and cultural differences led to distinct stages of virus spread, making it challenging for a single epidemiological prediction model to forecast the entire time series, thus necessitating different models. Based on the application of neural networks in epidemiological studies for time series prediction, this work aims to develop a methodology based on a multi-layer *Long Short-Term Memory* (LSTM) neural network, capable of being applied to nearly three years of pandemic data to predict the number of new daily cases of the disease. The methodology applies a series of tests, combining different network input data, such as daily cases and vaccinations to predict the number of cases for several days into the future. The experiments were conducted for Brazil and other countries, showing good results for predictions up to fifteen days in advance, with the potential to identify trend changes in the daily case timeline. This capability is useful for detecting the onset of new waves of infection, contributing to public health alert systems and immediate decision-making.

Keywords: COVID-19 forecasting; epidemiological modelling; LSTM networks; machine learning; time series analysis.

SUMÁRIO

LISTA DE ABREVIATURAS

AdaGrad	Adaptive Gradient Algorithm
AI	Artificial Intelligence
API	Application Programming Interface
ARIMA	AutoRegressive Integrated Moving Average
AWS	Amazon Web Services
BTT	Backpropagation Through Time
CNN	Convolutional Neural Network
COVID-19	Corona Virus Disease 2019
GRU	Gated Recurrent Unit
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAPE	Absolute Percentage Error
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
MLP	Multilayer Perceptron
RMSprop	Root Mean Square Propagation
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
SGD	Stochastic Gradient Descent

Variáveis e símbolos matemáticos

θ	Janela de observação no modelo de previsão.
τ	Horizonte temporal de entrada de dados.
η	Horizonte temporal de saída de dados.
Φ	Número de características de entrada por etapa de tempo.
$x_c(t)$	Casos diários de COVID-19 (nova variável em série temporal).
$x_v(t)$	Número de novas vacinações diárias.
Y	Saída prevista pelo modelo em formato univariado para cada janela.
X	Tensor tridimensional de entrada representando a série temporal multivariada.
$tanh(x)$	Função tangente hiperbólica usada nas células LSTM.
$\sigma(x)$	Função sigmoide usada nas redes neurais.
$g(x)$	Função de ativação para entrada (geralmente $tanh$).
$h(x)$	Função de ativação para saída (geralmente $tanh$).
$\odot$	Produto escalar.
R^2	Coeficiente de determinação.
$p.p.$	Pontos percentuais, diferença entre dois valores percentuais.

1 INTRODUÇÃO

1.1 Contextualização

A pandemia de COVID-19 (***Co****rona* ***Vi****rus* ***Di****sease* - ***19*** - ano em que a doença foi identificada pela primeira vez), iniciada em 2020, destacou a importância da epidemiologia matemática como uma ferramenta crucial para prever a dinâmica de disseminação de doenças (CAMPILLO-FUNOLLET et al., 2021; KWUIMY et al., 2020; PEIRLINCK et al., 2020a; PEIRLINCK et al., 2020b; PUGA; MONTEIRO, 2021; RĂDULESCU; WILLIAMS; CAVANAGH, 2020; RAMBAUT et al., 2020). Contudo, a rápida disseminação da SARS-CoV-2 (***S****evere* ***A****cute* ***R****espiratory* ***S****yndrome* - ***Co****rona****v****irus* - ***2*** - indicando que este é o segundo coronavírus identificado que causa SARS), combinada com a emergência de inúmeras variantes ao longo do tempo, apresentou um grande desafio para os modelos epidemiológicos tradicionais. Esses modelos, embora eficazes em certos contextos, enfrentaram dificuldades em prever os diferentes picos de infecção causados pelas novas variantes, resultando em respostas inadequadas em momentos críticos da pandemia (GUAN et al., 2020; HAAS et al., 2021; MONTEIRO; FANTI; TESSARO, 2020; NYBERG et al., 2022; ABU-RADDAD et al., 2022; SCHIMIT, 2021; SINGANAYAGAM et al., 2022; VOLZ et al., 2021).

Os picos de infecção tiveram origens e características distintas. Quanto à origem, os picos frequentemente surgiram com o aparecimento de novas variantes, que possuíam graus de infecciosidade e letalidade variados (MOURA et al., 2022; SLAVOV et al., 2022). Além disso, as mudanças em políticas de isolamento social também contribuízram para o aumento de casos em alguns instantes da pandemia (ALBRECHT, 2022). Em relação às características da COVID-19, muitos casos apresentaram sintomas como dor de cabeça, coriza, febre e perda de paladar, que se assemelham aos sintomas de outras doenças respiratórias. Essas manifestações clínicas mudavam de acordo com a variante do vírus, sendo que nem todos os indivíduos infectados manifestavam todos os sintomas. Além disso, houve também indivíduos que contraíam e propagavam a doença, mas eram assintomáticos (HUANG et al., 2020).

Considere a evolução temporal dos casos de COVID-19 no Brasil, conforme ilustrado na Figura 1.1. Nesta figura, as linhas tracejadas vermelhas indicam as datas aproximadas em que diversas variantes começaram a se espalhar no Brasil, iniciando com as variantes B.1.1.28 (Beta) e B.1.1.33 no início da pandemia e seguindo até as ondas de variantes Ômicron a partir de janeiro de 2022, até o quase fim da pandemia, que foi oficialmente declarado pela Organização Mundial da Saúde (OMS) em maio de 2023 (MOURA et al., 2022; SLAVOV et al., 2022). Esta figura mostra os diversos estágios da pandemia, durante os quais foram necessários múltiplos modelos matemáticos para representar com precisão todo o período. Alguns modelos abordaram os estágios iniciais da pandemia

em vários países (ACEMOGLU et al., 2021; DIAGNE et al., 2021; DIN et al., 2020; FUDOLIG; HOWARD, 2020; KHYAR; ALLALI, 2020; MONTEIRO; FANTI; TESSARO, 2020; PANWAR; UDUMAN; GóMEZ-AGUILAR, 2021; SCHIMIT, 2021; XU et al., 2023), outros focaram no impacto da vacinação, e outros ainda examinaram os picos periódicos (KUNIYA, 2022; KUNIYA, 2023; MOURA et al., 2022), semelhantes aos últimos três picos mostrados na figura.

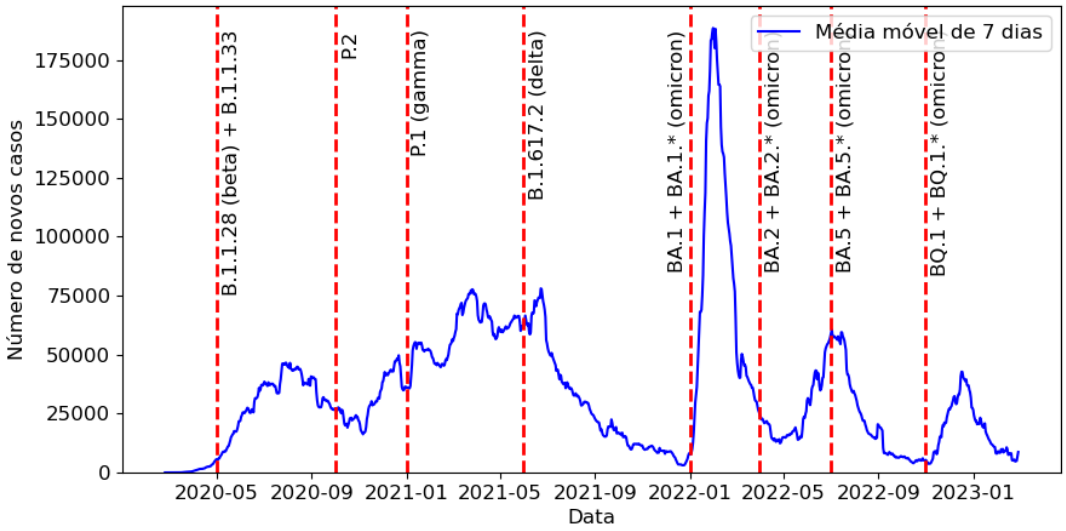


Figura 1.1 – *Evolução dos novos casos de COVID-19 no Brasil, mostrando uma média móvel de 7 dias desde o início da pandemia até fevereiro de 2023. As linhas tracejadas vermelhas representam a introdução de diversas variantes do SARS-CoV-2 na população brasileira, começando pelas variantes B.1.1.28 (Beta) e B.1.1.33 e culminando em múltiplas ondas de variantes Ômicron. A linha do tempo ilustra os desafios dinâmicos enfrentados na modelagem epidemiológica em diferentes estágios da pandemia, destacando a necessidade de abordagens matemáticas diversas para capturar as complexidades da disseminação do vírus e das medidas de controle.*

Uma aplicação do aprendizado de máquina (ML) e da inteligência artificial (IA) em epidemiologia é a estimativa de parâmetros em modelos. Essas tecnologias abordam o desafio não linear da estimativa de parâmetros ao explorar eficientemente o espaço de soluções e evitar mínimos locais, proporcionando uma vantagem computacional em relação aos métodos tradicionais (VRIES et al., 2006; SIMSKE, 2013; GLOVER; KOCHENBERGER, 2003). Técnicas como algoritmos genéticos, busca harmônica, redes neurais e aprendizado profundo são empregadas para aproximar soluções quase ótimas de forma robusta, tornando-as especialmente úteis no tratamento de problemas inversos e no mapeamento da progressão de doenças (MARINOV; MARINOVA, 2020; XIANG; LIU, 2015; MONTEIRO; GANDINI; SCHIMIT, 2020; GOPAL; LEE; SEOW, 2021; FOKAS; DIKAIOS; KASTIS, 2020; PEREIRA; SCHIMIT; BEZERRA, 2021).

Outra aplicação desses métodos em epidemiologia é a previsão e projeção da evolução temporal de casos e mortes, amplamente utilizada em modelos de COVID-19 durante a pandemia. Redes neurais de memória de curto e longo prazo (LSTM, do inglês *Long*

Short-Term Memory) estiveram entre as técnicas mais empregadas devido à sua capacidade de classificar, processar e prever dados temporais. Como um tipo de arquitetura de rede neural recorrente (RNN), o modelo LSTM tem sido utilizado em diversas áreas, incluindo economia financeira (BAO; YUE; RAO, 2017), processamento de dados biomédicos (GRAVES; SCHMIDHUBER, 2005; MA et al., 2015; OH et al., 2018; ORDóñEZ; ROGGEN, 2016; YILDIRIM, 2018; ZHAO; MAO; CHEN, 2019), agricultura (ZHANG et al., 2018a), energia (BOUKTIF et al., 2018; CHEMALI et al., 2018; ZHANG et al., 2018b; ABDEL-NASSER; MAHMOUD, 2019; SAGHEER; KOTB, 2019; ZHA et al., 2022) e processamento de imagens (ULLAH et al., 2017; WANG et al., 2019). Durante a pandemia de COVID-19, muitos pesquisadores utilizaram redes LSTM para prever as tendências da COVID-19 em diversos países.

Dessa forma, o presente trabalho desenvolve um modelo baseado em redes neurais LSTM multi-camadas com o objetivo de superar as limitações dos modelos epidemiológicos tradicionais. O modelo foi projetado para prever com maior precisão as ondas de contágio (novos casos), levando em conta as variações abruptas observadas nas séries temporais de disseminação de doenças, como ocorreu durante a pandemia de COVID-19. Sua aplicação ao longo de todo o período da pandemia no Brasil, aliada a análises detalhadas de diferentes configurações de previsão, evidencia seu potencial para aprimorar as respostas a futuras crises de saúde pública.

1.2 Problema de Pesquisa

Modelos epidemiológicos tradicionais se mostraram limitados na previsão dos picos de infecção ao longo de toda a série temporal, bem como não conseguiram se adaptar adequadamente às rápidas mudanças nas dinâmicas de contágio causadas pela emergência de novas variantes do vírus e por fatores externos, como mudanças nas políticas de saúde pública e nas condições socioeconômicas. Assim, o problema de pesquisa deste trabalho é: como desenvolver um modelo preditivo que seja capaz de prever a série temporal de casos da COVID-19, considerando as rápidas mudanças na dinâmica de propagação da doença e suas dependências a longo e curto prazo?

1.3 Hipótese

Propõe-se que a aplicação de modelos baseados em redes neurais profundas, especificamente a arquitetura LSTM multi-camadas, pode prever com boa precisão a série temporal de casos da COVID-19. A arquitetura LSTM, com sua capacidade de capturar dependências de longo e curto prazo, permite a integração eficiente de dados históricos e recentes, equilibrando padrões sazonais e variações abruptas. Além disso, supõe-se que a identificação de um conjunto reduzido de variáveis-chave possibilitará maior precisão e

flexibilidade nas previsões. Essa abordagem, ao ser avaliada em diferentes momentos da pandemia de COVID-19 e aplicada a diferentes países, deverá demonstrar que é possível alcançar uma boa acurácia preditiva de forma consistente, mesmo em contextos epidemiológicos distintos.

1.4 Justificativa

A pandemia de COVID-19 mostrou a necessidade de múltiplos modelos matemáticos para trabalhar com os diferentes estágios da pandemia. Como destacado na introdução, foi necessária uma série de modelos matemáticos para representar as diferentes fases da pandemia (ACEMOGLU et al., 2021; DIAGNE et al., 2021; DIN et al., 2020; FUDOLIG; HOWARD, 2020; KHYAR; ALLALI, 2020; MONTEIRO; FANTI; TESSARO, 2020; PANWAR; UDUMAN; GóMEZ-AGUILAR, 2021; SCHIMIT, 2021; XU et al., 2023). Por exemplo, alguns modelos foram eficazes em capturar os estágios iniciais da pandemia, enquanto outros abordaram os impactos da vacinação e os picos periódicos (KUNIYA, 2022; KUNIYA, 2023; MOURA et al., 2022). Essa dependência de múltiplas abordagens ressaltou a ausência de um modelo único capaz de lidar de forma robusta com a complexidade de uma pandemia como a de COVID-19.

Além disso, a necessidade de modelos que mantenham consistência preditiva em diferentes contextos e sob condições de variabilidade elevada é particularmente relevante em pandemias futuras. Desenvolver um modelo que consiga prever surtos de contágio com base em um conjunto reduzido de variáveis-chave, aplicável a diferentes regiões e estágios de disseminação facilita sua escalabilidade para novos cenários, e pode servir como uma ferramenta decisiva para estratégias de controle e mitigação de crises sanitárias. Isso oferece subsídios mais precisos para a tomada de decisão por parte de autoridades públicas, profissionais de saúde e pesquisadores.

1.5 Objetivos

1.5.1 Objetivos gerais

Desenvolver um modelo preditivo baseado em redes neurais LSTM multi-camadas para ser usado nos diferentes estágios de propagação de uma doença, incluindo períodos de estabilidade, picos e redução de casos. O modelo deve ser flexível e robusto para aplicação em diferentes contextos temporais e geográficos. Por fim, será testado com dados epidemiológicos da pandemia de COVID-19 do Brasil e outros países.

1.5.2 Objetivos Específicos

- Levantar e organizar dados epidemiológicos detalhados da COVID-19, com foco em variáveis relevantes para a modelagem preditiva;
- Desenvolver um modelo de rede neural recorrente com arquitetura multi-camadas LSTM, capaz de considerar dependências de curto e longo prazo nas séries temporais;
- Identificar e validar um conjunto reduzido de variáveis de entrada que seja eficaz para prever as variáveis de saída, garantindo simplicidade e eficiência computacional;
- Avaliar o desempenho do modelo em diferentes estágios da pandemia de COVID-19 e em diferentes países, comparando os resultados para verificar sua generalização e consistência preditiva.

1.6 Delimitação da Pesquisa

Esta investigação se dedica ao estudo e avaliação de um modelo de previsão construído com redes neurais LSTM multi-camadas que é aplicado aos dados epidemiológicos da pandemia de COVID-19. O estudo se concentra especialmente em antecipar os diversos estágios da propagação da doença em séries temporais com alta dinâmica de mudança, que inclui momentos de estabilidade, picos e diminuição dos casos.

Inicialmente será analisado o conjunto de dados do Brasil proveniente do portal *Our World in Data* (MATHIEU et al., 2020), que fornece uma compilação completa e uniformizada sobre casos confirmados da doença COVID 19, mortes pela doença e informações relacionadas com a vacinação entre outras variáveis associadas à pandemia do coronavírus. Essas informações são essenciais para compreender em profundidade os dados disponíveis e avaliar os desafios específicos decorrentes dos surtos causados pelas diferentes variantes do vírus SARS-CoV - 02 e influências externas como alterações nas políticas públicas ou nas condições econômico-financeiras das populações afetadas. Além disso, o estudo identifica um conjunto reduzido de variáveis-chave para uma boa acurácia das previsões. Esse estudo visa não apenas desenvolver um modelo robusto para a COVID-19, mas também criar uma abordagem escalável e aplicável a futuras epidemias ou pandemias.

1.7 Organização do trabalho

Este trabalho está estruturado da seguinte forma: o próximo capítulo apresenta a revisão da literatura, abordando os principais conceitos e estudos relacionados ao tema. Em seguida, a metodologia detalha os procedimentos adotados para o desenvolvimento e avaliação do modelo proposto, seguida pela apresentação e conclusão dos resultados obtidos.

2 REVISÃO DA LITERATURA

Neste capítulo, é realizada uma revisão da literatura abordando os temas de modelos epidemiológicos tradicionais e redes LSTM em epidemiologia matemática.

2.1 Modelos epidemiológicos tradicionais

Os primeiros modelos epidemiológicos tinham como objetivo principal descrever a evolução temporal de casos e indivíduos infectados por uma doença em uma população. Em 1760, Daniel Bernoulli desenvolveu um modelo pioneiro para avaliar o impacto da técnica de inoculação contra a varíola na expectativa de vida de uma população Bernoulli (1760). Em 1916, Ronald Ross introduziu um modelo probabilístico que marcou o início das abordagens compartimentais, dividindo a população em diferentes estados e analisando como parâmetros, como a taxa de transmissão e recuperação, influenciam a dinâmica de uma epidemia Ross (1916). Posteriormente, em 1927, Kermack e McKendrick apresentaram um modelo matemático baseado em equações diferenciais ordinárias (EDOs), aplicado a surtos de peste bubônica em Londres e Mumbai. Esse modelo, conhecido como SIR (Suscetível-Infectado-Recuperado), trouxe análises quantitativas, permitindo avaliar as condições para o surgimento de uma epidemia com base nos parâmetros do modelo Kermack e McKendrick (1927).

O modelo SIR é amplamente usado como referência para o desenvolvimento de novos modelos. Durante a pandemia de COVID-19, esses modelos foram aprimorados para incluir mais compartimentos, como indivíduos expostos ou infectados assintomáticos, refletindo a complexidade da dinâmica da transmissão viral. Por exemplo, para considerar diferentes variantes, um modelo SEIR (Suscetível-Exposto-Infectado-Recuperado) foi usado para avaliar combinações de estratégias de controle da pandemia, como quarentena e vacinação. O estudo ressaltou a importância de se considerar diferentes compartimentos para diferentes variantes, e o correto uso de dados reais da doença (KHYAR; ALLALI, 2020). Em outro estudo que considerou diversas variantes do vírus da COVID-19, foram avaliadas condições para coexistência e cenários endêmicos com diferentes variantes em uma população (FUDOLIG; HOWARD, 2020).

Indivíduos assintomáticos, que podiam transmitir o vírus da COVID-19 sem perceberem que estão infectados, tiveram um impacto significativo durante a pandemia. Por isso, foram considerados como um compartimento específico na modelagem proposta por Monteiro, Fanti e Tessaro (2020), denominada SAIR (Suscetível-Assintomático-Infectado-Recuperado). Nesse estudo, os autores fizeram um paralelo do modelo em autômatos celulares probabilistícos, com o mesmo modelo descrito em equações diferenciais ordinárias. A avaliação de diferentes estratégias de controle baseadas em regimes de quarentena também foi realizada no estudo (MONTEIRO; FANTI; TESSARO, 2020).

Outras abordagens foram utilizadas para analisar estratégias de controle da pandemia de COVID-19. Din et al. (2020) investigaram a dinâmica global da COVID-19 por meio de um modelo SEIR que incorporou múltiplas exposições de indivíduos suscetíveis ao vírus. Nesse tipo de análise, são avaliados os pontos de equilíbrio globais do sistema, ou seja, estados que, independentemente das condições iniciais da simulação, conduzem o sistema a um regime permanente correspondente. Os resultados indicaram que medidas como quarentena e vacinação podem reduzir a incidência da doença na população, conduzindo o sistema ao equilíbrio livre da doença (DIN et al., 2020).

Intervenções contínuas não farmacêuticas e vacinação também foram consideradas em um estudo para o Senegal (DIAGNE et al., 2021); além da avaliação de políticas de isolamento social para um modelo SIR estruturado, ou seja, que divide os indivíduos de uma população por faixa etária (ACEMOGLU et al., 2021). Nesse último modelo, confirmou-se a importância de restrições mais severas para indivíduos acima de 65 anos. Um modelo compartimental que considerou diferentes indivíduos hospitalizados foi proposto por Schimit (2021) para avaliar o impacto da disponibilidade de equipamento adequado em unidades de terapia intensiva nos casos da doença para o Brasil (SCHIMIT, 2021).

Diferente dos modelos apresentados anteriormente, os trabalhos de Kuniya (2022), Kuniya (2023), Moura et al. (2022) propuseram modelos para momentos de oscilação da COVID-19 em uma população. A partir do modelo SIR, Kuniya (2022) investigaram os mecanismos essenciais por trás das ondas epidêmicas recorrentes observadas em muitos países, introduzindo um modelo com efeito psicológico e atraso de tempo distribuído (KUNIYA, 2022). Em seguida, Kuniya (2023) exploraram um modelo epidêmico com quarentena e atraso distribuído no tempo, demonstrando que o isolamento e o atraso desempenham papéis podem influenciar a ocorrência de surtos epidêmicos recorrentes (KUNIYA, 2023).

Com outra perspectiva, Moura et al. (2022) analisaram a dinâmica temporal da COVID-19 no Brasil, identificando ondas epidemiológicas associadas a características intrínsecas da doença e à resposta imunológica da população. O trabalho sugere que a periodicidade das ondas pode ser explicada por fatores internos, como o esgotamento temporário de suscetíveis e a recuperação gradual de suscetibilidade em grupos previamente expostos (MOURA et al., 2022).

Esses estudos ressaltaram os fatores que podem levar a uma periodicidade de populações em relação aos surtos epidêmicos. Além disso, mostra uma necessidade de modelos específicos para tais situações. Dessa forma, percebe-se uma necessidade de modelos que consigam ser ajustados para diversos estágio de uma epidemia em uma população.

2.2 Redes LSTM em epidemiologia matemática

A rede neural recorrente do tipo *Long-Short Term Memory* superou algumas limitações de redes neurais recorrentes no tratamento de séries temporais. Usando uma notação semelhante à usada em (HOUDT; MOSQUERA; NáPOLES, 2020), um bloco LSTM possui a arquitetura apresentada na Figura 2.1.

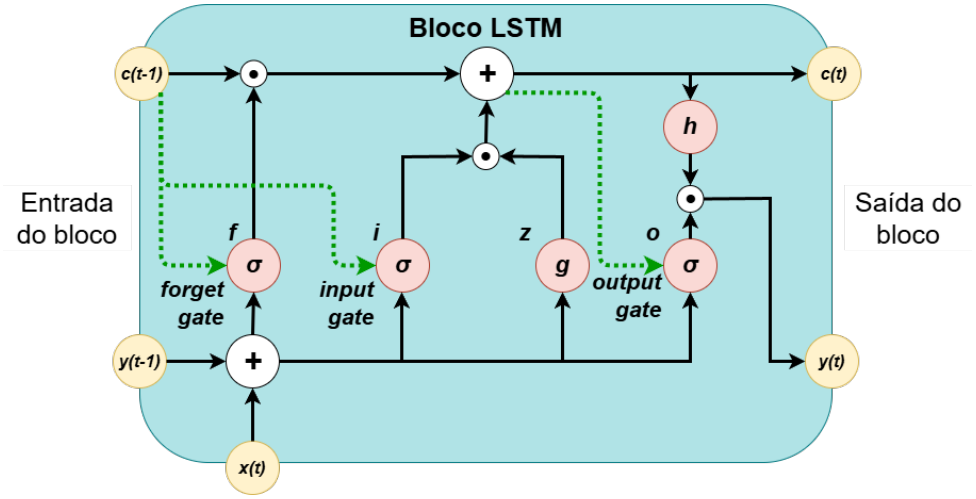


Figura 2.1 – *Representação de uma célula LSTM*

O bloco consiste em uma entrada de bloco $z(t)$, input gate $i(t)$, forget gate $f(t)$, cell $c(t)$, output gate $o(t)$ e saída de bloco $y(t)$. As expressões da entrada do bloco para a saída são:

$$z(t) = g(W_z x(t) + R_z y(t-1) + b_z) \tag{2.1}$$
$$i(t) = \sigma(W_i x(t) + R_i y(t-1) + p_i \odot c(t-1) + b_i) \tag{2.2}$$
$$f(t) = \sigma(W_f x(t) + R_f y(t-1) + p_f \odot c(t-1) + b_f) \tag{2.3}$$
$$c(t) = z(t) \odot i(t) + c(t-1) \odot f(t) \tag{2.4}$$
$$o(t) = \sigma(W_o x(t) + R_o y(t-1) + p_o \odot c(t) + b_o) \tag{2.5}$$
$$y(t) = g(c(t)) \odot o(t), \tag{2.6}$$

Inicialmente, a entrada do bloco é atualizada combinando a entrada atual $x(t)$ com a saída $y(t-1)$ da iteração anterior, resultando no valor $z(t)$, no qual W_z e R_z são os pesos, e b_z é o vetor de bias, sendo $\odot$ o produto escalar. Em seguida, o valor do input gate $i(t)$ é calculado com base em $x(t)$, $y(t-1)$ e o valor da célula $c(t-1)$, com pesos W_i, R_i, p_i e o vetor de bias b_i. O forget gate $f(t)$ é determinado usando $x(t)$, $y(t-1)$, $c(t-1)$, com pesos W_f, R_f, p_f e o vetor de bias b_f. O valor da célula $c(t)$ é então calculado combinando $z(t)$, $i(t)$, $f(t)$ e $c(t-1)$. O output gate $o(t)$ combina $x(t)$, $y(t-1)$, $c(t)$, com pesos W_o, R_o, p_o e o vetor de bias b_o. Finalmente, a saída do bloco $y(t)$ é determinada pelo produto

da ativação da célula $g(c(t))$ e o valor de $o(t)$. As funções de ativação utilizadas são σ para a função de ativação do gate (*sigmoid*) , g é a função de ativação de input (*tanh*), e h é a função de ativação de output (*tanh*).

As redes neurais recorrentes do tipo LSTM vêm sendo usadas em diversas áreas, como o estudo de séries temporais financeiras (SEZER; GUDELEK; OZBAYOGLU, 2020; BAO; YUE; RAO, 2017), processamento de dados biomédicos (GRAVES; SCHMIDHUBER, 2005; MA et al., 2015; OH et al., 2018; ORDóñEZ; ROGGEN, 2016; YILDIRIM, 2018; ZHAO; MAO; CHEN, 2019), agricultura (ZHANG et al., 2018a), energia (BOUKTIF et al., 2018; CHEMALI et al., 2018; ZHANG et al., 2018b; ABDEL-NASSER; MAHMOUD, 2019; SAGHEER; KOTB, 2019; ZHA et al., 2022) e processamento de imagens (ULLAH et al., 2017; WANG et al., 2019). Pesquisadores da área de matemática epidemiológica começaram a obter bons resultados na avaliação de séries temporais epidêmicas usando diversas arquiteturas de redes e modelos estatísticos, como Perceptron Multicamadas Profundo (*Deep Multilayer Perceptron* - DMLP), Rede Neural Convolucional (*Convolutional Neural Network* - CNN), Rede Neural de Regressão Generalizada (*Generalized Regression Neural Network* - GRNN), Rede Neural Não Linear Autorregressiva (*Nonlinear Autoregressive Neural Network* - NARANN), Unidade Recorrente com Portas (*Gated Recurrent Unit* - GRU), Regressão Ponderada Geograficamente (*Geographically Weighted Regression*) - GWR), Modelo Auto-Regressivo Integrado de Médias Móveis (*AutoRegressive Integrated Moving Average* - ARIMA), e Suavização Exponencial (*Exponential Smoothing* - ES), mas a rede LSTM é tem sido escolhida por sua eficácia no tratamento de séries temporais.

Em comparação com outras abordagens, as redes recorrentes LSTM mostraram melhores resultados em alguns estudos epidemiológicos. Em um estudo sobre a incidência do vírus HIV (Vírus da Imunodeficiência Humana, do inglês *Human Immunodeficiency Virus*) em Guangxi, China, uma rede LSTM apresentou menores valores de erro na previsão de dados temporais de casos, em comparação a modelos como ARIMA, GRNN e ES (WANG et al., 2019). O modelo ARIMA é um método estatístico para análise de séries temporais, enquanto o ES (*Exponential Smoothing*) é uma técnica de suavização exponencial, ambos voltados para a previsão de padrões em dados históricos. A comparação de uso de redes LSTM com ARIMA também foi realizada por Elsheikh et al. (2021). Nesse estudo, que também usou NARANN para a previsão de casos confirmados da COVID-19 e óbitos em seis países (incluindo o Brasil), confirmou as redes LSTM como a arquitetura que retornou os menores valores de erros nas previsões (ELSHEIKH et al., 2021).

Com dados da China, Wang et al. (2020) compararam o uso de redes LSTM com um modelo tradicional SEIRD (Suscetível-Exposto-Infectado-Recuperado-Falecido (*Dead*)) e com o GWR. De novo, as redes LSTM obtiveram os menores erros ao prever dados temporais da COVID-19 (WANG et al., 2020). As redes LSTM foram comparadas com redes neurais recorrentes para predição de casos da COVID-19 em três países: Malásia,

Marrocos e Arábia Saudita. O modelo baseado em redes LSTM teve uma acurácia de cerca de 5% maior que as redes neurais recorrentes. Por fim, comparações entre as redes LSTM e modelos clássicos de aprendizagem de máquina, como a floresta aleatória (*random forest*) e a máquina de vetores de suporte (*support vector machine*) também deram vantagem às redes LSTM na previsão de casos da COVID-19, ainda que o custo computacional tenha sido maior (ZHOU et al., 2023).

O modelo Bi-LSTM também foi aplicado a dados da COVID-19. Said et al. (2021) e Shahid, Zameer e Muneeb (2020a) exploraram o uso de LSTM, GRU e Bi-LSTM para prever casos confirmados, mortes e recuperações em 10 países, e o Bi-LSTM apresentou o melhor desempenho. O modelo Bi-LSTM (*Bidirectional Long Short-Term Memory*) é uma extensão da arquitetura LSTM, capaz de processar informações em ambas as direções da sequência temporal, permitindo capturar relações tanto passadas quanto futuras nos dados, o que pode melhorar a precisão das previsões em séries temporais. Esse modelo também retornou bons resultados com uso de dados de mobilidade e medidas de isolamento (ALASSAFI; JARRAH; ALOTAIBI, 2022).

Outros tipos de dados de entrada foram testados em alguns estudos. Dados epidemiológicos cumulativos da COVID-19, bem como informações sobre isolamento social e quarentena, foram utilizados como entrada em uma rede LSTM com dados da Itália, apresentando bons resultados (YAN et al., 2020; JIN et al., 2022). Além disso, dados postados por usuários na rede social Twitter foram empregados como entrada da rede e contribuíram para melhorar a previsão de surtos de dengue e gripe, utilizando informações das Filipinas e de países vizinhos (SHAHID; ZAMEER; MUNEEB, 2020b).

As redes LSTM também foram usadas em combinações com outros modelos na busca da melhora das previsões envolvidas. Em comparação com uma rede LSTM simples, a combinação do modelo probabilístico de Markov com a rede LSTM reduziu os erros das previsões em cerca de 70% (MA et al., 2021). O uso de redes LSTM com K-means teve bons resultados na previsão de surto de casos no estado de Louisiana, EUA, com erros 80% menores que um modelo tradicional SEIR, e uso de dados climáticos e demográficos da região (VADYALA et al., 2021).

Yudistira et al. (2021) propuseram uma análise multivariada a partir de uma rede híbrida LSTM e convolucionária. A primeira camada dessa rede é uma camada típica de redes neurais convolucionais com uma dimensão, seguida por uma camada LSTM. Ao todo, mais de cinquenta variáveis foram usadas, de áreas como ambiente, educação, governo, saúde e economia. Os resultados mostraram um modelo capaz de prever a evolução da COVID-19, mas que também mostrou a relação entre as séries temporais de diversas variáveis (YUDISTIRA et al., 2021).

Vale ressaltar que esses estudos não costumam usar um período de tempo de análise dos dados muito maior que algumas semanas. Usando redes LSTM, Kumar et al. (2021) consideraram pouco mais de três semanas de dados para suas análise sobre a disseminação

da COVID-19 na Nova Zelândia (KUMAR et al., 2021); cerca de cinco meses foram considerados para avaliação de longo prazo da COVID-19 para Rússia, Peru e Irã (AUNG et al., 2023); e nove meses foram considerados por Alassafi, Jarrah e Alotaibi (2022) no estudo descrito anteriormente neste capítulo.

Com base na revisão dos trabalhos apresentados nesse capítulo, vale ressaltar o uso de redes LSTM para diferentes tipos de modelos híbridos, com diferentes tipos de dados de entrada e séries temporais de dados com grande variabilidade. Isso mostra a flexibilidade do modelo para trabalhar com séries temporais. A partir dessa base teórica, o próximo capítulo contém detalhes do desenvolvimento do modelo proposto nessa dissertação.

3 METODOLOGIA

Neste capítulo, a metodologia aplicada nesse projeto é proposta. A pesquisa possui caráter aplicado, que propõe o desenvolvimento e análise de modelos preditivos baseados em redes neurais LSTM multi-camadas, para a previsão de novos casos de COVID-19. Usando dados da pandemia disponibilizados pelo projeto Our World in Data (MATHIEU et al., 2020), investiga-se combinações de variáveis, ajustes de hiperparâmetros e diferentes configurações de tuplas temporais, com o objetivo de construir um modelo capaz de identificar tendências e picos epidêmicos. A acurácia do modelo é avaliada por meio de métricas estatísticas, e o desempenho é testado em cenários distintos. Com essa abordagem, pretende-se não apenas prever a evolução da pandemia no Brasil, mas também estabelecer um método escalável que seja possível aplicar a outros contextos epidemiológicos no futuro.

3.1 Fonte de Dados

Os dados relacionados à COVID-19 foram obtidos por meio do projeto Our World in Data (MATHIEU et al., 2020), uma iniciativa do Global Change Data Lab, uma organização sem fins lucrativos dedicada a estudar as transformações nas condições de vida global e no ambiente terrestre. O projeto Our World in Data compilou dados diários sobre a pandemia de COVID-19, disponibilizando perfis epidemiológicos detalhados para 207 países. Esses perfis incluem informações variadas, como casos totais e diários, mortes, dados de testagem, vacinação e respostas governamentais à crise. As informações podem ser acessadas diretamente no site do Our World in Data ou por meio de seu repositório no GitHub.

As variáveis disponíveis no conjunto de dados incluem o número acumulado e os novos casos diários de COVID-19, o total de mortes e novas mortes diárias, dados de testagem (como o número total de testes e a taxa de positividade) e detalhes sobre a vacinação, como o número de doses administradas e a porcentagem da população vacinada. Além disso, a base de dados abrange informações sobre medidas governamentais, como políticas de lockdown e restrições de viagem, embora possam haver lacunas em algumas dessas informações para determinados países.

Para esta pesquisa, após a aquisição dos dados, foi realizada uma análise específica para o Brasil, com foco em variáveis selecionadas que estão descritas na Tabela 4.1. Após testes iniciais do modelo para alguns conjuntos de dados, que buscavam validar a estrutura do *script*, duas variáveis se mostraram mais promissoras para serem usadas como entrada do modelo: novos casos diários ($x_c(t)$) e novas vacinações diárias ($x_v(t)$).

3.2 A rede LSTM multi-camadas

O bloco LSTM único mostrado no capítulo 2 pode então ser integrado em uma rede para corresponder ao formato dos dados e aos requisitos da aplicação. O número de passos de tempo usados como entrada e o número de passos de tempo que a rede irá prever são dois hiperparâmetros importantes da rede. Aqui, foram combinadas multipla camadas LSTM, configuradas para reter as características de uma série temporal. A figura 3.1 contém a estrutura de rede utilizada neste estudo. A combinação de blocos LSTM são implementados com uma única camada LSTM Keras. Keras (CHOLLET et al., 2015) é uma API Python que trabalha com redes LSTM usando TensorFlow (ABADI et al., 2015) em seu backend.

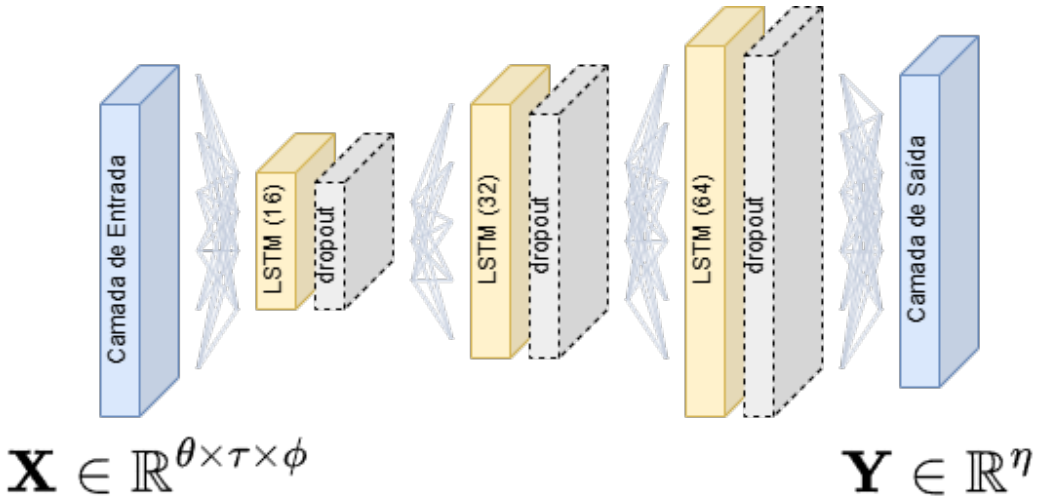


Figura 3.1 – *Representação da rede LSTM Multi-Camadas*

O total de inputs para a rede é uma multivariada série temporal, representada por um tensor tridimensional:

$$\mathbf{X} \in \mathbb{R}^{\theta \times \tau \times \phi},$$

em que:

- θ: o total de número de exemplos de treinamento utilizados no dataset;
- τ: o número de passos anteriores considerados;
- ϕ: o número de características input.

A rede gerada é uma predição univariada para cada tupla temporal. O output para cada batch é representado por:

$$\mathbf{Y} \in \mathbb{R}^{\eta}$$

O treinamento da rede para ajustar os pesos nas equações 2.1-2.6 usa o algoritimo Adam (*Adaptive Moment Estimation*). Adam é uma variação do *Stochastic Gradient Descent*, combinando características do AdaGrad (*Adaptive Gradient Algorithm*) e RMSprop (*Root Mean Square Propagation*). AdaGrad adapta a taxa de aprendizado baseado na taxa de atualização, enquanto RMSprop normaliza o gradiante mantendo uma média móvel do gradiente quadrado. Adam alavanca tanto a média móvel dos gradientes (primeiro momento) e gradiente quadrado (segundo momento) para ajustar a taxa de aprendizagem para cada parâmetro.

Considerando que está sendo aplicado uma avaliação por janela móvel para os casos de COVID-19 na janela temporal, uma importante variável para ser avaliada é o tamanho do conjunto de dados treinados, denotado por θ. Seu tamanho precisa ser selecionado cuidadosamente para garantir que a rede possa trabalhar efetivamente com uma serie temporal que exiba características diferentes para a janela temporal. Assim, a tupla (θ, τ, η) representa as variáveis para a quantidade de dias utilizados para o treinamento da rede, o número de dias de entrada da rede, e o número de dias a serem previstos.

Devido aos diferentes cenários da COVID-19, é necessário retreinar o conjunto de dados frequentemente, mesmo quando utilizando a arquitetura LSTM. Nesse contexto, aplica-se uma janela móvel de θ dias, configurada individualmente em cada teste. Para cada instância de teste i, a janela é definida a partir do dia $i - \theta + 1$ até o dia i, em que θ é o tamanho da janela de treinamento. Além disso, são considerados os intervalos para a realização de previsões, como a janela de entrada de dados (τ) — que determina quantos dias de dados serão fornecidos ao modelo para predição — e a janela de saída (η), que define o horizonte de previsão, ou seja, quantos dias à frente o modelo tentará prever. Tal representação pode ser observada na figura 3.2.

Para avaliar a precisão das previsões geradas pelo modelo, usa-se o *MAPE* (*Mean Absolute Percentage Error*). O *MAPE* é uma medida de precisão comum em previsões, que compara os valores previstos pelo modelo com os dados reais, oferecendo uma perspectiva sobre a acurácia das previsões em termos percentuais. Essa métrica é particularmente útil para entender a eficácia do modelo em capturar as tendências dos dados de COVID-19. O *MAPE* é calculado pela fórmula:

$$\text{MAPE} = \left(\frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \right) \times 100\%$$

em que n representa o número total de itens, y_i são os valores reais e $\hat{y}_i$ são os valores previstos pelo modelo.

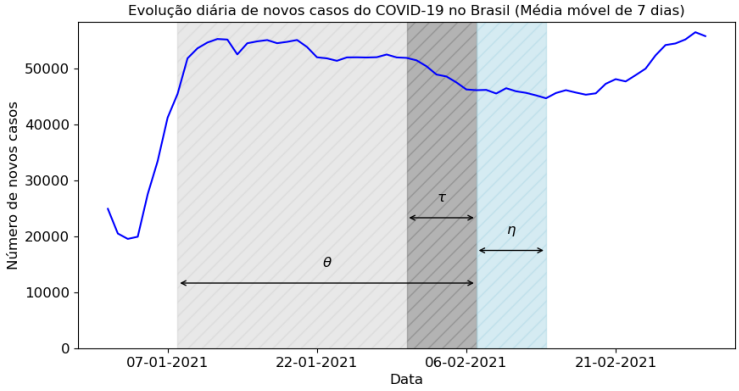


Figura 3.2 – *Exemplo da tupla de variáveis* (θ, τ, η)*. Neste caso, o número de dias usados para treinar a rede é* $\theta = 30$*, o número de dias de entrada para a rede é* $\tau = 7$*, e o número de dias que a rede prevê a evolução dos casos de COVID-19 é* $\eta = 7$*.*

Além do *MAPE*, também foi utilizado o R^2 (coeficiente de determinação). O R^2 é uma medida estatística que indica a proporção da variância dos dados reais que é explicada pelo modelo de predição. Essa métrica oferece uma perspectiva sobre a qualidade do ajuste do modelo aos dados observados, sendo particularmente útil para entender a eficácia do modelo em capturar as tendências dos dados de COVID-19. O R^2 é calculado pela fórmula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

em que n representa o número total de itens, y_i são os valores reais, $\hat{y}_i$ são os valores previstos pelo modelo, e $\bar{y}$ é a média dos valores reais. Um valor de R^2 próximo de 1 indica que o modelo explica bem a variância dos dados, enquanto um valor próximo de 0 indica que o modelo não explica bem a variância.

3.3 Seleção de características e Hiperparametros

Foram realizados experimentos com as características selecionadas, utilizando a arquitetura LSTM. A Tabela 3.1, demonstra as tuplas usadas nessa primeira análise. Para suavizar os picos e vales nos dados, foi aplicada uma média móvel de 14 dias às características.

Além disso, seleciona-se um período de pouco mais de três anos para a realização dos testes. A importância dessa decisão para a análise está diretamente relacionada à ocorrência de diferentes variantes e momentos de picos — estes bastante distintos — colocando assim o modelo sob uma avaliação mais rigorosa.

Os resultados do modelo são mostrados no próximo capítulo.

Treinamento (θ)	Input (τ)	Output (η)
30	7	1
30	7	5
30	14	1
30	14	5
30	14	10
30	14	15
60	7	1
60	7	5
60	14	1
60	14	5
60	14	10
60	14	15
60	30	15
60	30	21
60	45	15
60	45	21
60	45	30
75	30	15
75	60	30

Tabela 3.1 – *Combinações de tuplas θ, τ e η*

4 RESULTADOS

Neste capítulo, os resultados são apresentados. Primeiramente, as diferentes tuplas de predição apresentadas na Tabela 4.1 são testadas usando a quantidade de casos diários ($x_c(t)$) como característica de entrada ($\phi = 1$), para identificar a melhor combinação da tupla (θ, τ, η) para os dados do Brasil. Em seguida, usando a tupla que retornou o menor erro, são simulados os casos de outros países, bem como testes envolvendo uma segunda característica de entrada (portanto $\phi = 2$), a quantidade diária de vacinas aplicadas ($x_v(t)$).

4.1 Configuração da rede e primeiros testes

Conforme mencionado no capítulo anterior, a rede consiste em três conjuntos de camadas LSTM e dropout seguidas por uma camada densa com a saída. Para prever η dias à frente, a rede é treinada com os últimos θ dias, utilizando os últimos τ dias como entrada da rede. A lista completa de parâmetros de dados e hiperparâmetros da rede pode ser encontrada nas Tabelas 4.2 4.3.

Descrição	Valor	Referência
Característica objetivo da predição	novos casos diários ($x_c(t)$)	-
Outras características de input	novas vacinas diárias ($x_v(t)$)	experimental
Início da série temporal	2020-03-15	-
Final da série temporal	2023-07-31	-
Número de camadas LSTM combinadas com dropout	3	experimental
Número de camadas densas com o output da rede	1	experimental
Taxa de aprendizado	0.001	default
Número de épocas	500	experimental
Tamanho do Batch	32	experimental
Taxa de Early stopping para patience	0.50	default
Taxa de decaimento da taxa de aprendizagem	0.94	default
Steps até o decaimento da taxa de aprendizagem	8000	experimental
Taxa de Dropout	0.3	experimental
Janela de média móvel	14 days	experimental
Regularização do L1	0.001	default
Regularização do L2	0.001	default

Tabela 4.1 – *Parâmetros do modelo LSTM*

Inicialmente, foram realizados testes com dados da COVID-19 no Brasil, abrangendo o período de 15 de março de 2020 a 31 de julho de 2023, com o objetivo de identificar uma configuração ótima para a tupla (θ, τ, η). Nestes testes, a série temporal do número diário de casos foi utilizada como característica de entrada da rede, com $\phi = 1$. As tuplas testadas, juntamente com o MAPE e o coeficiente de determinação R^2, são apresentadas nas Tabelas 4.2 e 4.3, respectivamente. A primeira coluna lista a tupla, seguida pelos dias futuros intermediários para a previsão da rede, denotados por t, que assume valores no intervalo $t = 1, 2, \ldots, \eta$. Embora tenham sido testados valores altos de η, apenas os primeiros quinze valores de t são exibidos nessas tabelas. A última coluna apresenta o MAPE médio (R^2) ao longo dos diferentes valores de t para aquela linha.

Para a camada de entrada η, será fornecida a série de dados reais correspondentes aos

dias anteriores ao período previsto. Ou seja, se $\eta = 14$, um conjunto de 14 dias de dados reais será enviado ao modelo. Na rodada de predição, a saída η corresponderá a todos os valores de $\eta = t$. Assim, para $t = 15$, será retornado um conjunto de 15 elementos, representando a previsão para os próximos 15 dias.

Para a avaliação do desempenho do modelo, os indicadores MAPE e R^2 serão calculados comparando os valores previstos com os dados reais da mesma série temporal.

θ, τ, η	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$	$t=7$	$t=8$	$t=9$	$t=10$	$t=11$	$t=12$	$t=13$	$t=14$	$t=15$	Média
30,14,15	13.42	12.37	13.81	13.52	13.36	13.01	13.23	14.64	15.65	15.44	17.10	17.21	18.37	19.24	18.94	15.29
60,7,5	18.65	19.30	20.76	22.91	25.30	-	-	-	-	-	-	-	-	-	-	21.38
60,14,10	27.49	26.75	26.22	25.34	26.22	26.93	27.86	29.56	31.78	33.51	-	-	-	-	-	28.17
30,14,10	18.00	17.68	17.65	17.99	18.24	18.56	19.01	19.70	20.33	20.82	-	-	-	-	-	18.8
30,7,5	22.56	22.68	22.81	23.11	23.57	-	-	-	-	-	-	-	-	-	-	22.95
60,7,1	38.26	-	-	-	-	-	-	-	-	-	-	-	-	-	-	38.26
60,14,5	37.26	36.79	37.01	37.98	38.89	-	-	-	-	-	-	-	-	-	-	37.59
30,14,5	40.37	38.36	37.04	35.74	34.90	-	-	-	-	-	-	-	-	-	-	37.28
30,7,1	43.52	-	-	-	-	-	-	-	-	-	-	-	-	-	-	43.52
75,30,15	55.58	52.89	50.25	48.08	46.70	45.46	44.81	44.92	45.07	45.81	47.57	49.82	52.65	55.83	58.95	49.63
60,30,21	61.45	59.40	57.28	55.18	53.58	52.13	51.19	50.22	50.05	50.01	50.32	50.13	50.52	50.58	51.08	52.87
60,14,1	78.28	-	-	-	-	-	-	-	-	-	-	-	-	-	-	78.28
60,30,15	66.77	65.48	65.19	63.95	63.90	64.44	64.64	64.49	65.48	65.88	67.56	68.90	70.44	71.96	73.60	66.85
60,45,30	82.44	79.03	75.61	71.51	72.75	72.39	67.20	65.63	64.68	60.69	59.45	60.57	58.81	57.73	57.14	67.04
60,45,21	91.77	92.69	89.67	87.76	86.39	83.09	83.45	81.40	82.69	81.60	79.52	79.23	79.56	78.57	79.62	83.8
60,45,15	100.15	99.67	98.29	94.00	94.64	92.08	92.46	93.29	88.55	91.29	93.09	94.35	94.04	92.70	97.52	94.41
75,60,30	118.55	115.99	114.19	110.40	107.93	105.26	103.56	102.85	102.17	100.26	97.72	97.65	98.13	95.32	96.60	104.44

Tabela 4.2 – *Valores de MAPE para predições utilizando diferentes combinações de tupla* (θ, τ, η)*, com sua correspondência média de MAPE entre diferentes horizontes de predição. A contagem de casos diários para o Brasil utiliza somente uma característica de input* ($\phi = 1$)*.*

θ,τ,η	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$	$t=7$	$t=8$	$t=9$	$t=10$	$t=11$	$t=12$	$t=13$	$t=14$	$t=15$
30,14,15	0.97	0.98	0.98	0.98	0.98	0.98	0.97	0.97	0.96	0.96	0.95	0.95	0.95	0.95	0.95
60,7,5	0.95	0.94	0.92	0.90	0.87	-	-	-	-	-	-	-	-	-	-
60,14,10	0.89	0.89	0.88	0.87	0.85	0.83	0.80	0.77	0.74	0.71	-	-	-	-	-
30,14,10	0.90	0.90	0.90	0.90	0.90	0.89	0.89	0.88	0.88	0.87	-	-	-	-	-
30,7,5	0.86	0.87	0.88	0.88	0.88	-	-	-	-	-	-	-	-	-	-
60,7,1	0.78	-	-	-	-	-	-	-	-	-	-	-	-	-	-
60,14,5	0.77	0.76	0.74	0.73	0.71	-	-	-	-	-	-	-	-	-	-
30,14,5	0.72	0.72	0.73	0.74	0.74	-	-	-	-	-	-	-	-	-	-
30,7,1	0.73	-	-	-	-	-	-	-	-	-	-	-	-	-	-
75,30,15	0.66	0.66	0.66	0.66	0.65	0.64	0.63	0.61	0.60	0.58	0.55	0.52	0.50	0.47	0.44
60,30,21	0.51	0.54	0.55	0.57	0.57	0.58	0.59	0.59	0.59	0.59	0.59	0.57	0.58	0.56	0.55
60,14,1	0.48	-	-	-	-	-	-	-	-	-	-	-	-	-	-
60,30,15	0.43	0.44	0.43	0.44	0.44	0.42	0.42	0.41	0.40	0.39	0.37	0.35	0.33	0.32	0.30
60,45,30	0.19	0.23	0.25	0.29	0.28	0.29	0.34	0.35	0.34	0.38	0.38	0.40	0.39	0.41	0.41
60,45,21	0.04	0.02	0.06	0.07	0.08	0.11	0.12	0.14	0.13	0.13	0.15	0.14	0.14	0.19	0.17
60,45,15	-0.09	-0.03	-0.04	0.01	0.02	0.05	0.01	0.00	0.05	0.05	0.01	0.02	0.01	0.06	
75,60,30	-0.16	-0.13	-0.12	-0.09	-0.08	-0.04	-0.03	-0.02	-0.01	0.01	0.02	0.02	0.03	0.06	0.04

Tabela 4.3 – *Valores de* R^2 *para predições utilizando diferentes combinações de tupla* (θ, τ, η)*. A contagem de casos diários para o Brasil utiliza somente uma característica de input* ($\phi = 1$)*.*

A partir desses resultados iniciais, a tupla ($\theta = 30, \tau = 14, \eta = 15$) apresentou o

melhor desempenho, com um MAPE médio de 15,29% e um R^2 de 0,95 no pior caso. A Figura 4.1 mostra a evolução temporal dos casos diários de COVID-19 no Brasil (em azul) junto com os valores previstos (em verde) e o erro percentual absoluto (APE - em vermelho). Observa-se um bom alinhamento entre os dados e as previsões, apesar de um erro ligeiramente maior próximo ao maior pico de casos. Nesse caso, a tupla $(\theta = 30, \tau = 14, \eta = 15)$ com $\phi = 1$ (utilizando $x_c(t)$ como entrada) foi aplicada.

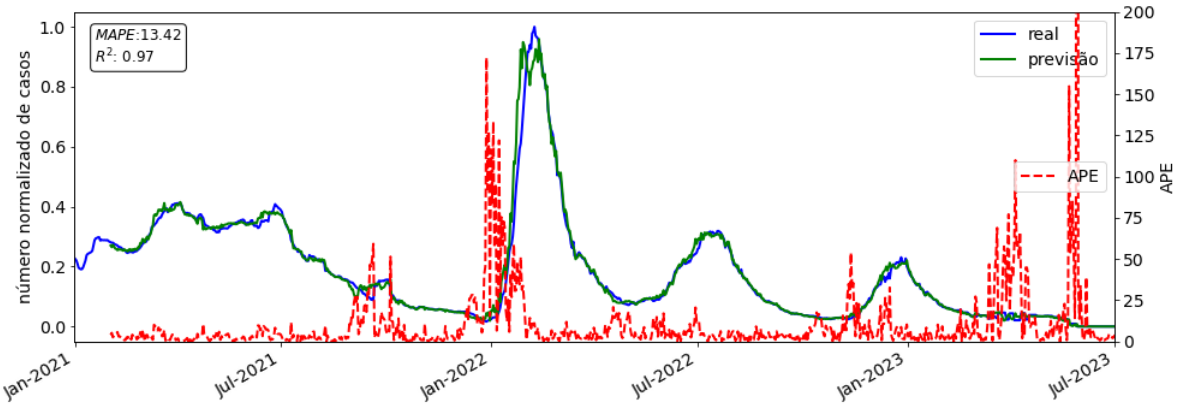
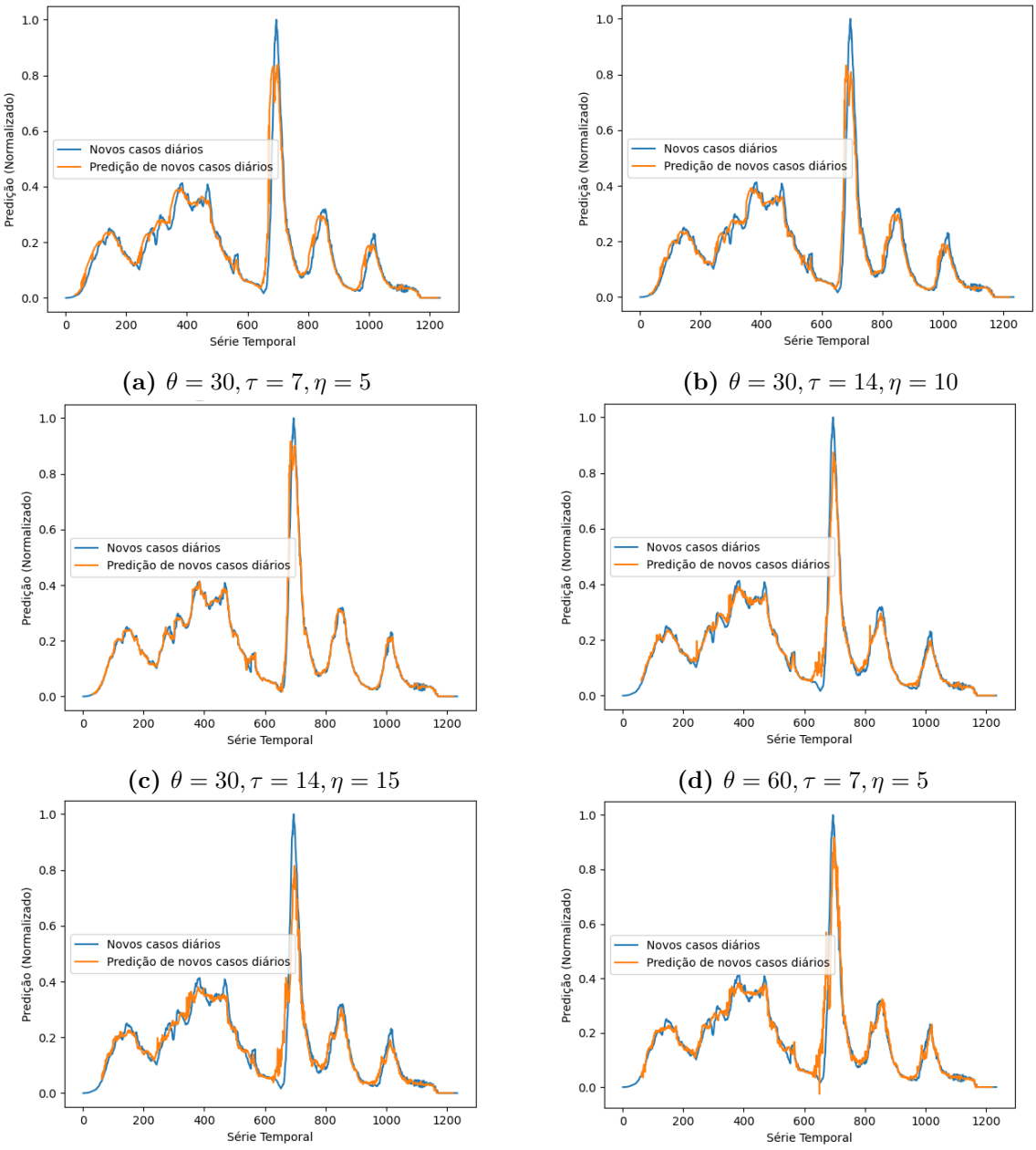


Figura 4.1 – *Evolução temporal dos casos diários observados de COVID-19 no Brasil (azul), valores previstos (verde) e APE (vermelho utilizando a configuração* $(\theta = 30, \tau = 14, \eta = 15)$ *com* $\phi = 1$ *e* $t = 1$*. O modelo captura bem a tendência geral, com pequenas discrepâncias próximas ao pico.*

Alguns intervalos da linha do tempo são responsáveis por aumentar o erro médio, como a queda de casos antes do grande pico da Omicron em Dezembro de 2022 e alguns intervalos com poucos casos próximo ao final do intervalo total considerado nas simulações. Visualmente, as configurações de tuplas $(\theta = 30, \tau = 7, \eta = 5)$, $(\theta = 30, \tau = 14, \eta = 10)$, $(\theta = 30, \tau = 14, \eta = 15)$, $(\theta = 60, \tau = 7, \eta = 5)$, $(\theta = 60, \tau = 14, \eta = 10)$, e $(\theta = 60, \tau = 14, \eta = 15)$ também apresentam bons resultados na maior parte do intervalo, conforme pode ser observado na Figura 4.2.



(a) $\theta = 30, \tau = 7, \eta = 5$

(b) $\theta = 30, \tau = 14, \eta = 10$

(c) $\theta = 30, \tau = 14, \eta = 15$

(d) $\theta = 60, \tau = 7, \eta = 5$

(e) $\theta = 60, \tau = 14, \eta = 10$

(f) $\theta = 60, \tau = 14, \eta = 15$

Figura 4.2 – *Evolução temporal dos casos diários observados de COVID-19 no Brasil (azul) e valores previstos (laranja) para as configurações* $(\theta = 30, \tau = 7, \eta = 5)$, $(\theta = 30, \tau = 14, \eta = 10)$, $(\theta = 30, \tau = 14, \eta = 15)$, $(\theta = 60, \tau = 7, \eta = 5)$, $(\theta = 60, \tau = 14, \eta = 10)$, *e* $(\theta = 60, \tau = 14, \eta = 15)$, *todos com* $t = 1$.

4.2 Resultados para outros países

A análise é expandida para os outros países considerados, também alcançando bons resultados. A tupla ($\theta = 30, \tau = 14, \eta = 15$) é utilizada, com o número de casos diários como entrada única ($\phi = 1$) para Espanha (ESP), Itália (ITA), Brasil (BRA), Grã-Bretanha (GBR) e França (FRA). O MAPE e o R^2 dos resultados são apresentados nas Tabelas 4.4 e 4.5, respectivamente. Observa-se que foram obtidos melhores resultados em relação ao Brasil, sendo os menores MAPE encontrados para Espanha e Itália, em comparação com o Brasil.

País	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$	$t=7$	$t=8$	$t=9$	$t=10$	$t=11$	$t=12$	$t=13$	$t=14$	$t=15$	Média
BRA	13.42	12.37	13.81	13.52	13.36	13.01	13.23	14.64	15.65	15.44	17.10	17.21	18.37	19.24	18.94	15.29
ESP	8.90	8.34	7.28	6.89	6.80	6.57	7.12	7.77	8.52	8.82	9.28	9.51	10.17	10.42	10.94	8.49
FRA	10.89	10.77	9.57	8.79	8.58	8.72	9.55	11.36	12.38	12.84	13.53	14.15	14.68	15.65	17.44	11.93
GBR	7.16	6.56	6.02	5.89	5.89	6.12	6.59	7.46	8.04	8.80	9.67	10.09	10.65	10.97	11.30	8.08
ITA	6.82	6.16	5.56	5.47	5.43	5.65	6.19	6.97	7.92	8.68	9.54	10.19	10.94	11.45	11.90	7.92

Tabela 4.4 – *Valores de MAPE para predições utilizando a configuração de tupla* ($\theta = 30, \tau = 14, \eta = 15$*, com sua correspondência média de MAPE entre diferentes horizontes de predição. A contagem de casos diários para o Brasil utiliza somente uma característica de input* ($\phi = 1$*).*

País	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$	$t=7$	$t=8$	$t=9$	$t=10$	$t=11$	$t=12$	$t=13$	$t=14$	$t=15$
BRA	0.97	0.98	0.98	0.98	0.98	0.98	0.97	0.97	0.96	0.96	0.95	0.95	0.95	0.95	0.95
ESP	0.98	0.98	0.99	0.99	0.99	0.98	0.98	0.97	0.97	0.97	0.96	0.96	0.95	0.95	0.93
FRA	0.98	0.98	0.98	0.98	0.98	0.98	0.97	0.96	0.95	0.94	0.93	0.93	0.93	0.93	0.92
GBR	0.97	0.97	0.97	0.97	0.97	0.96	0.95	0.94	0.94	0.93	0.93	0.94	0.93	0.94	0.94
ITA	0.98	0.99	0.99	0.99	0.99	0.98	0.98	0.97	0.97	0.96	0.96	0.95	0.95	0.95	0.94

Tabela 4.5 – *Valores de* R^2 *para predições utilizando a configuração de tupla* ($\theta = 30, \tau = 14, \eta = 15$*. A contagem de casos diários no Brasil considera apenas uma característica de entrada* ($\phi = 1$).

Esses resultados demonstram a viabilidade de utilizar $t = 15$ para prever o número de casos com a rede proposta. Para alguns países, pode haver um erro maior em comparação a períodos de tempo mais curtos, mas os resultados permanecem satisfatórios. A previsão do número de casos com quinze dias de antecedência pode ser uma característica valiosa para as autoridades de saúde pública, permitindo ajustes oportunos nas políticas de controle durante um surto. As Figuras 4.3 a 4.7 mostram as previsões para cinco países utilizando $t = 15$, a tupla ($\theta = 30, \tau = 14, \eta = 15$) e o número de casos diários como única entrada da rede ($\phi = 1$). Nota-se o forte alinhamento nessas previsões, indicando projeções confiáveis com quinze dias de antecedência.

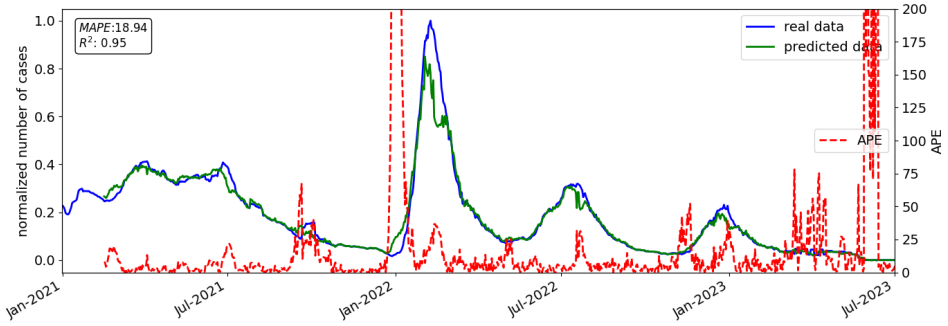


Figura 4.3 – *Evolução temporal dos casos diários observados de COVID-19 no Brasil (azul), valores previstos (verde) e APE (vermelho) usando a configuração* $(\theta = 30, \tau = 14, \eta = 15)$ *com* $\phi = 1$, *e* $t = 15$.

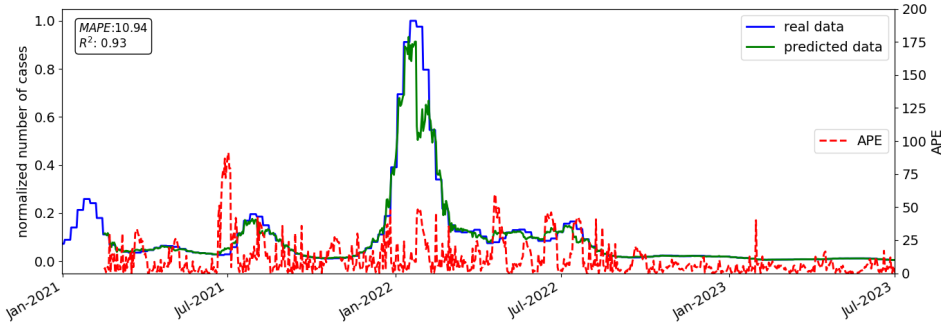


Figura 4.4 – *Evolução temporal dos casos diários observados de COVID-19 na Espanha (azul), valores previstos (verde) e APE (vermelho) usando a configuração* $(\theta = 30, \tau = 14, \eta = 15)$ *com* $\phi = 1$, *e* $t = 15$.

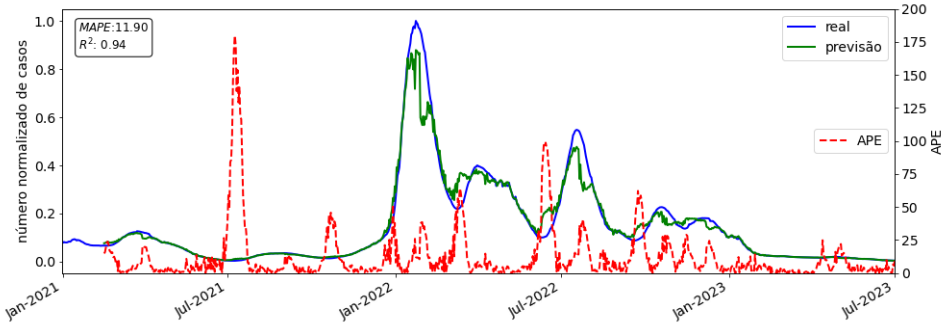


Figura 4.5 – *Evolução temporal dos casos diários observados de COVID-19 na Itália (azul), valores previstos (verde) e APE (vermelho) usando a configuração* $(\theta = 30, \tau = 14, \eta = 15)$ *com* $\phi = 1$, *e* $t = 15$.

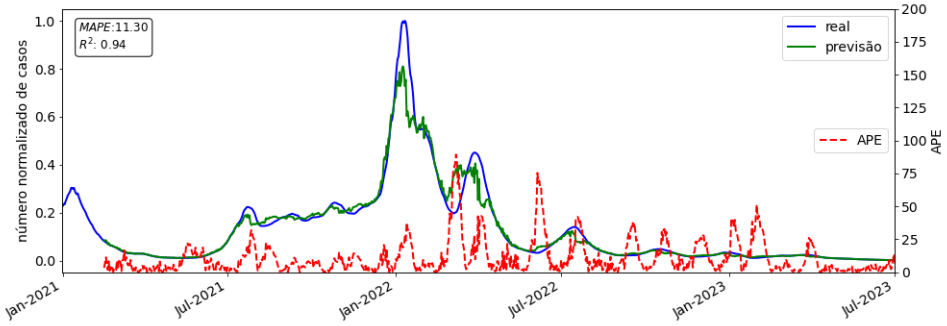


Figura 4.6 – *Evolução temporal dos casos diários observados de COVID-19 na Grã-Bretanha (azul), valores previstos (verde) e APE (vermelho) usando a configuração* $(\theta = 30, \tau = 14, \eta = 15)$ *com* $\phi = 1$, *e* $t = 15$.

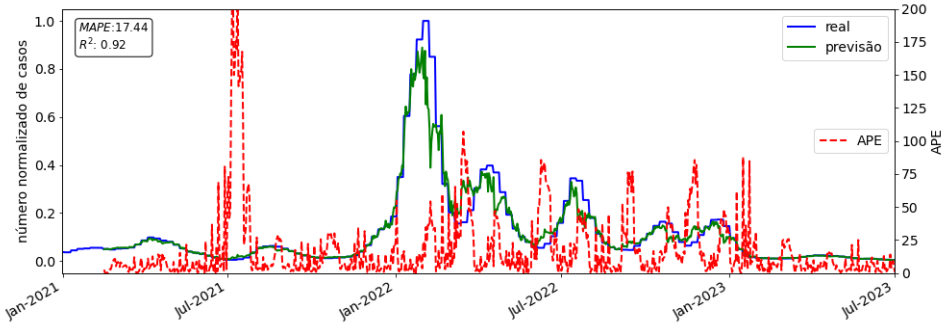


Figura 4.7 – *Evolução temporal dos casos diários observados de COVID-19 na França (azul), valores previstos (verde) e APE (vermelho) usando a configuração* $(\theta = 30, \tau = 14, \eta = 15)$ *com* $\phi = 1$, *e* $t = 15$.

4.3 Vacinação como outro dado de entrada

Adiciona-se mais uma característica como entrada da rede: o número diário de vacinações $(x_v(t))$, sendo agora $\phi = 2$. Entre os países considerados, Brasil, Itália e França forneceram dados que foram possíveis de serem usados no treinamento. O problema de muitos países para os dados da vacinação à epoda da pandemia era uma atualização semanal, quinzenal, ou mesmo mensal dos dados de vacinas aplicadas. Para esses três países, pode-se obter dados (quase) diários. As Tabelas 4.6 e 4.7 apresentam os valores de MAPE e R^2 das previsões em comparação com os números reais de casos diários nesses países.

País	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$	$t=7$	$t=8$	$t=9$	$t=10$	$t=11$	$t=12$	$t=13$	$t=14$	$t=15$	Média
FRA	15.29	14.27	13.58	12.63	12.64	13.41	14.31	14.06	13.81	13.73	14.06	14.68	15.54	16.64	16.70	14.36
ITA	10.37	9.91	9.75	9.70	9.36	9.24	9.38	9.68	10.06	10.40	10.85	11.20	11.37	11.83	12.05	10.34

Tabela 4.6 – *Valores de MAPE para previsões usando a configuração de tupla* $(\theta = 30, \tau = 14, \eta = 15)$, *com a média do MAPE em diferentes horizontes de previsão para o Brasil, Itália e França. A contagem diária de casos e vacinas é usada como entradas* $(\phi = 2)$.

País	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$	$t=7$	$t=8$	$t=9$	$\eta=10$	$t=11$	$t=12$	$t=13$	$t=14$	$t=15$
ITA	0.98	0.98	0.98	0.98	0.98	0.98	0.97	0.97	0.97	0.97	0.96	0.96	0.96	0.96	0.95
FRA	0.96	0.96	0.96	0.96	0.96	0.94	0.95	0.95	0.95	0.95	0.94	0.94	0.94	0.94	0.945

Tabela 4.7 – *Valores de R^2 para previsões usando a configuração de tupla ($\theta = 30, \tau = 14, \eta = 15$), em diferentes horizontes de previsão para o Brasil, Itália e França. A contagem diária de casos e vacinas é usada como entradas ($\phi = 2$).*

Embora os erros sejam ligeiramente maiores do que nas simulações com apenas uma característica como entrada da rede, a inclusão dos dados de vacinação ainda produz bons resultados. A principal melhoria é observada nas previsões ao redor dos picos no número de casos, como mostrado nas Figuras 4.8 e 4.9. Nessas figuras, observa-se que, embora o modelo preveja um aumento antecipado nos casos antes do principal pico logo após janeiro de 2022, ele prevê o momento do pico com maior precisão do que as simulações que não incluíram os dados de vacinação.

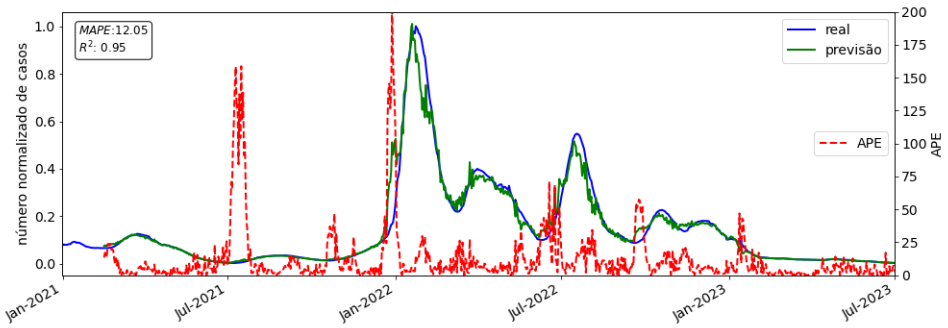


Figura 4.8 – *Evolução temporal dos casos diários observados de COVID-19 na Itália (azul), valores previstos (verde) e APE (vermelho) usando a configuração ($\theta = 30, \tau = 14, \eta = 15$) com $\phi = 2$, e $t = 15$.*

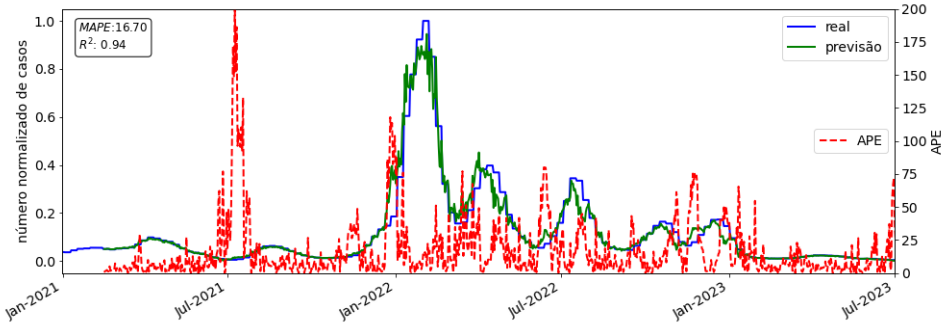


Figura 4.9 – *Evolução temporal dos casos diários observados de COVID-19 na França (azul), valores previstos (verde) e APE (vermelho) usando a configuração ($\theta = 30, \tau = 14, \eta = 15$) com $\phi = 2$, e $t = 15$.*

No próximo capítulo, os resultados são discutidos e a conclusão do trabalho é apresentada.

5 CONCLUSÕES

Este trabalho foi motivado por um problema crítico durante a pandemia de COVID-19: a predição de surtos em uma série temporal contínua, marcada por múltiplas e distintas ondas de contágio causadas por múltiplos fatores como os sócio-econômicos e as novas variantes do vírus. Para enfrentar esse desafio, desenvolveu-se e testou-se um modelo baseado em redes LSTM multi-camadas (*Long Short-Term Memory*) com o objetivo de prever o número diário de casos de uma doença ao longo de diferentes estágios de uma epidemia, usando dados da COVID-19 para validar o modelo.

O modelo obteve bons resultados para a COVID-19 e se mostrou uma abordagem promissora para a previsão do número diário de casos de uma epidemia. Com um bom desempenho preditivo de até quinze dias, o erro percentual absoluto foi menor durante períodos de relativa estabilidade no número de casos. Mesmo durante períodos de aumentos ou reduções súbitas no número de casos, como observado durante o surgimento da variante Ômicron no início de 2022, o modelo retornou resultados satisfatórios. Apresentou-se valores elevados do erro quando o número real de casos era próximo de zero, situação em que pequenas variações na previsão podem fazer com que o erro percentual absoluto atinja valores altos.

As redes LSTM têm sido amplamente utilizadas para prever dados epidemiológicos devido à sua capacidade de lidar eficazmente com sequências temporais. Durante a pandemia de COVID-19, a previsão do número de casos tornou-se uma aplicação crítica, com modelos sendo treinados para prever a contagem de casos em horizontes de alguns dias, considerando dados de algumas semanas (SAID et al., 2021; LUO et al., 2021; WANG et al., 2020). Além da COVID-19, modelos LSTM também foram aplicados a outros surtos epidêmicos. Por exemplo, dados em nível macro foram utilizados para prever a disseminação de doenças como dengue e influenza, com alguns estudos integrando fontes de dados alternativas, como postagens em mídias sociais, ao longo de períodos de aproximadamente um ano (SHAHID; ZAMEER; MUNEEB, 2020b). Para doenças específicas, como HIV na China (WANG et al., 2019) e COVID-19 no Catar (SAID et al., 2021), China (JIN et al., 2022), e em vários outros países (SHAHID; ZAMEER; MUNEEB, 2020a; MA et al., 2021), os modelos LSTM demonstraram precisão preditiva com valores de MAPE em torno de 10%. Horizontes de tempo mais curtos, como previsões de um dia à frente, geralmente apresentaram melhores resultados, com valores de MAPE variando entre 2% e 5%. Esses resultados foram alcançados em estudos como os de Yan et al. (YAN et al., 2020), Luo et al. (LUO et al., 2021) e Wang et al. (WANG et al., 2020), que aplicaram estruturas baseadas em LSTM a países das Américas e outras regiões.

O modelo proposto utilizou a arquitetura LSTM multi-camadas, configurada com três camadas principais e ajuste de hiperparâmetros, que incluiu taxa de aprendizado, regularização e dropout. A configuração da tupla temporal ($\theta = 30, \tau = 14, \eta = 15$)

mostrou-se a mais eficiente, resultando em valores de MAPE médio de 15,29% e coeficiente de determinação R^2 de 0,97 durante os testes realizados com dados diários de COVID-19 no Brasil entre março de 2020 e julho de 2023. Apesar dos resultados satisfatórios, foi observado um aumento no erro preditivo em períodos de maior volatilidade, como os picos epidêmicos associados à introdução de novas variantes. Esses achados corroboram estudos da literatura, como os de Shahid, Zameer e Muneeb (2020a), que demonstraram a robustez das redes LSTM para prever surtos em séries temporais não estacionárias, e Moura et al. (2022), que destacaram a relevância da adaptação dinâmica dos modelos em contextos pandêmicos. Ademais, os dados de entrada, como novos casos diários e novas vacinações, foram cruciais para garantir uma modelagem que capturasse a complexidade das ondas epidêmicas.

O objetivo deste trabalho foi fornecer um modelo robusto baseada em uma rede LSTM que pudesse ser ajustada para um horizonte temporal amplo (15 dias) em um único país e, em seguida, aplicada a outros países para prever o número diário de casos de COVID-19. Essa estrutura apresentou bons resultados em comparação com a literatura, pois, com um MAPE no intervalo de 10% a 18%, foi possível prever o número diário de casos com quinze dias de antecedência, proporcionando uma janela valiosa para planejar e controlar a propagação da doença. O uso da evolução temporal da vacinação como outra entrada para a rede LSTM não melhorou os resultados de MAPE, mas a linha do tempo prevista aproximou-se mais da linha do tempo real nos picos da doença, permitindo uma melhor previsão de quando um pico ou mudança de tendência pode ocorrer.

Uma desvantagem do modelo proposta é seu custo computacional. A abordagem de janela deslizante iterativa, combinada com os requisitos de treinamento das redes LSTM, exige recursos computacionais significativos, especialmente para horizontes temporais mais longos e conjuntos de dados maiores. Essa limitação pode dificultar sua escalabilidade para aplicações em tempo real ou análises extensas entre países. Além disso, a precisão das previsões depende fortemente da qualidade e confiabilidade dos dados de entrada. Conjuntos de dados de COVID-19 mais consistentes e abrangentes, incluindo relatórios precisos de casos, taxas de testagem e dados de vacinação, poderiam melhorar ainda mais o desempenho do modelo, particularmente durante períodos de rápidas mudanças na dinâmica da doença. Esforços futuros devem se concentrar em otimizar a eficiência computacional do modelo e em aproveitar dados de maior qualidade para alcançar resultados ainda mais confiáveis.

Como perspectivas para trabalhos futuros, sugere-se a expansão do modelo para incluir variáveis complementares, como indicadores socioeconômicos, ambientais e dados de mobilidade, que podem somar aos fatores que influenciam a propagação de doenças. A adaptação para séries temporais mais curtas, especialmente em cenários com dados escassos, pode ampliar sua aplicabilidade em regiões menos desenvolvidas. Além disso, explorar arquiteturas emergentes, como Transformers (WOLF et al., 2020), ou a integração

de técnicas híbridas combinando LSTM com algoritmos de otimização, pode trazer ganhos adicionais de desempenho, especialmente em cenários com alta variabilidade ou incerteza.

REFERÊNCIAS BIBLIOGRÁFICAS

ABADI, M. et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. 2015. <https://www.tensorflow.org/>. Citado na pág. 23.

ABDEL-NASSER, M.; MAHMOUD, K. Accurate photovoltaic power forecasting models using deep lstm-rnn. *Neural Computing and Applications*, v. 31, 2019. Citado na pág. 13, 19.

ABU-RADDAD, L. J. et al. Effect of mrna vaccine boosters against sars-cov-2 omicron infection in qatar. *New England Journal of Medicine*, v. 386, 2022. Citado na pág. 11.

ACEMOGLU, D. et al. Optimal targeted lockdowns in a multigroup sir model. *American Economic Review: Insights*, v. 3, 2021. Citado na pág. 12, 14, 17.

ALASSAFI, M. O.; JARRAH, M.; ALOTAIBI, R. Time series predicting of covid-19 based on deep learning. *Neurocomputing*, v. 468, 2022. Citado na pág. 20, 21.

ALBRECHT, D. Vaccination, politics and covid-19 impacts. *BMC Public Health*, v. 22, 2022. Citado na pág. 11.

AUNG, N. N. et al. A novel bidirectional lstm deep learning approach for covid-19 forecasting. *Scientific Reports*, v. 13, 2023. Citado na pág. 21.

BAO, W.; YUE, J.; RAO, Y. A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *PLoS ONE*, v. 12, 2017. Citado na pág. 13, 19.

BERNOULLI, D. Essai d'une nouvelle analyse de la mortalité causée par la petite vérole, et des avantages de l'inoculation pour la prévenir. *Mémoires de Mathématique et de Physique, tirés des registres de l'Académie Royale des Sciences de l'année*, 1760. Citado na pág. 16.

BOUKTIF, S. et al. Optimal deep learning lstm model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches. *Energies*, v. 11, 2018. Citado na pág. 13, 19.

CAMPILLO-FUNOLLET, E. et al. Predicting and forecasting the impact of local outbreaks of covid-19: Use of seir-d quantitative epidemiological modelling for healthcare demand and capacity. *International Journal of Epidemiology*, v. 50, 2021. Citado na pág. 11.

CHEMALI, E. et al. Long short-term memory networks for accurate state-of-charge estimation of li-ion batteries. *IEEE Transactions on Industrial Electronics*, v. 65, 2018. Citado na pág. 13, 19.

CHOLLET, F. et al. *Keras*. 2015. <https://keras.io>. Citado na pág. 23.

DIAGNE, M. L. et al. A mathematical model of covid-19 with vaccination and treatment. *Computational and Mathematical Methods in Medicine*, v. 2021, 2021. Citado na pág. 12, 14, 17.

DIN, R. u. et al. Study of global dynamics of covid-19 via a new mathematical model. *Results in Physics*, v. 19, 2020. Citado na pág. 12, 14, 17.

ELSHEIKH, A. H. et al. Deep learning-based forecasting model for covid-19 outbreak in saudi arabia. *Process Safety and Environmental Protection*, v. 149, 2021. Citado na pág. 19.

FOKAS, A. S.; DIKAIOS, N.; KASTIS, G. A. Mathematical models and deep learning for predicting the number of individuals reported to be infected with sars-cov-2. *Journal of The Royal Society Interface*, v. 17, 2020. Citado na pág. 12.

FUDOLIG, M.; HOWARD, R. The local stability of a modified multi-strain sir model for emerging viral strains. *PLoS ONE*, v. 15, 2020. Citado na pág. 12, 14, 16.

GLOVER, F. W.; KOCHENBERGER, G. A. (Ed.). *Handbook of Metaheuristics*. [S.l.]: Kluwer / Springer, 2003. v. 57. (International Series in Operations Research & Management Science, v. 57). Citado na pág. 12.

GOPAL, K.; LEE, L. S.; SEOW, H. V. Parameter estimation of compartmental epidemiological model using harmony search algorithm and its variants. *Applied Sciences (Switzerland)*, v. 11, 2021. Citado na pág. 12.

GRAVES, A.; SCHMIDHUBER, J. Framewise phoneme classification with bidirectional lstm and other neural network architectures. In: . [S.l.: s.n.], 2005. v. 18. Citado na pág. 13, 19.

GUAN, W. et al. Clinical characteristics of coronavirus disease 2019 in china. *New England Journal of Medicine*, v. 382, 2020. Citado na pág. 11.

HAAS, E. J. et al. Impact and effectiveness of mrna bnt162b2 vaccine against sars-cov-2 infections and covid-19 cases, hospitalisations, and deaths following a nationwide vaccination campaign in israel: an observational study using national surveillance data. *The Lancet*, v. 397, 2021. Citado na pág. 11.

HOUDT, G. V.; MOSQUERA, C.; NáPOLES, G. A review on the long short-term memory model. *Artificial Intelligence Review*, v. 53, 2020. Citado na pág. 18.

HUANG, B. et al. *Characteristics of the Coronavirus Disease 2019 and related Therapeutic Options*. 2020. Citado na pág. 11.

JIN, Y. et al. Prediction of covid-19 trends based on lstm in a dynamic epidemiological environment. *Computers, Materials & Continua*, v. 70, 2022. Citado na pág. 20, 36.

KERMACK, W. O.; MCKENDRICK, A. G. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, v. 115, n. 772, 1927. Citado na pág. 16.

KHYAR, O.; ALLALI, K. Global dynamics of a multi-strain seir epidemic model with general incidence rates: application to covid-19 pandemic. *Nonlinear Dynamics*, v. 102, 2020. Citado na pág. 12, 14, 16.

KUMAR, S. et al. Forecasting the spread of covid-19 using lstm network. *BMC Bioinformatics*, v. 22, 2021. Citado na pág. 20, 21.

KUNIYA, T. Recurrent epidemic waves in a delayed epidemic model with quarantine. *Journal of Biological Dynamics*, v. 16, 2022. Citado na pág. 12, 14, 17.

KUNIYA, T. Hopf bifurcation in an sir epidemic model with psychological effect and distributed time delay. In: _____. [S.l.: s.n.], 2023. Citado na pág. 12, 14, 17.

KWUIMY, C. A. K. et al. Nonlinear dynamic analysis of an epidemiological model for covid-19 including public behavior and government action. *Nonlinear Dynamics*, v. 101, 2020. Citado na pág. 11.

LUO, J. et al. Time series prediction of covid-19 transmission in america using lstm and xgboost algorithms. *Results in Physics*, v. 27, 2021. Citado na pág. 36.

MA, R. et al. The prediction and analysis of covid-19 epidemic trend by combining lstm and markov method. *Scientific Reports*, v. 11, 2021. Citado na pág. 20, 36.

MA, X. et al. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transportation Research Part C: Emerging Technologies*, v. 54, 2015. Citado na pág. 13, 19.

MARINOV, T. T.; MARINOVA, R. S. Dynamics of covid-19 using inverse problem for coefficient identification in sir epidemic models. *Chaos, Solitons & Fractals: X*, v. 5, 2020. Citado na pág. 12.

MATHIEU, E. et al. Coronavirus pandemic (covid-19). *Our World in Data*, 2020. Https://ourworldindata.org/coronavirus. Citado na pág. 15, 22.

MONTEIRO, L. H. A.; FANTI, V. C.; TESSARO, A. S. On the spread of sars-cov-2 under quarantine: A study based on probabilistic cellular automaton. *Ecological Complexity*, v. 44, 2020. Citado na pág. 11, 12, 14, 16.

MONTEIRO, L. H. A.; GANDINI, D. M.; SCHIMIT, P. H. T. The influence of immune individuals in disease spread evaluated by cellular automaton and genetic algorithm. *Computer Methods and Programs in Biomedicine*, v. 196, 2020. Citado na pág. 12.

MOURA, E. C. et al. Covid-19: temporal evolution and immunization in the three epidemiological waves, brazil, 2020-2022. *Revista de saude publica*, v. 56, 2022. Citado na pág. 11, 12, 14, 17, 37.

NYBERG, T. et al. Comparative analysis of the risks of hospitalisation and death associated with sars-cov-2 omicron (b.1.1.529) and delta (b.1.617.2) variants in england: a cohort study. *The Lancet*, v. 399, 2022. Citado na pág. 11.

OH, S. L. et al. Automated diagnosis of arrhythmia using combination of cnn and lstm techniques with variable length heart beats. *Computers in Biology and Medicine*, v. 102, 2018. Citado na pág. 13, 19.

ORDóñEZ, F. J.; ROGGEN, D. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors (Switzerland)*, v. 16, 2016. Citado na pág. 13, 19.

PANWAR, V. S.; UDUMAN, P. S. S.; GóMEZ-AGUILAR, J. F. Mathematical modeling of coronavirus disease covid-19 dynamics using cf and abc non-singular fractional derivatives. *Chaos, Solitons and Fractals*, v. 145, 2021. Citado na pág. 12, 14.

PEIRLINCK, M. et al. Outbreak dynamics of covid-19 in china and the united states. *Biomechanics and Modeling in Mechanobiology*, v. 19, 2020. Citado na pág. 11.

PEIRLINCK, M. et al. Visualizing the invisible: The effect of asymptomatic transmission on the outbreak dynamics of covid-19. *Computer Methods in Applied Mechanics and Engineering*, v. 372, 2020. Citado na pág. 11.

PEREIRA, F. H.; SCHIMIT, P. H. T.; BEZERRA, F. E. A deep learning based surrogate model for the parameter identification problem in probabilistic cellular automaton epidemic models. *Computer Methods and Programs in Biomedicine*, v. 205, 2021. Citado na pág. 12.

PUGA, G. F.; MONTEIRO, L. H. A. The co-circulation of two infectious diseases and the impact of vaccination against one of them. *Ecological Complexity*, v. 47, 2021. Citado na pág. 11.

RAMBAUT, A. et al. A dynamic nomenclature proposal for sars-cov-2 lineages to assist genomic epidemiology. *Nature Microbiology*, v. 5, 2020. Citado na pág. 11.

ROSS, R. An application of the theory of probabilities to the study of a priori pathometry.—part i. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, v. 92, n. 638, 1916. Citado na pág. 16.

RăDULESCU, A.; WILLIAMS, C.; CAVANAGH, K. Management strategies in a seir-type model of covid 19 community spread. *Scientific Reports*, v. 10, 2020. Citado na pág. 11.

SAGHEER, A.; KOTB, M. Time series forecasting of petroleum production using deep lstm recurrent networks. *Neurocomputing*, v. 323, 2019. Citado na pág. 13, 19.

SAID, A. B. et al. Predicting covid-19 cases using bidirectional lstm on multivariate time series. *Environmental Science and Pollution Research*, 2021. Citado na pág. 20, 36.

SCHIMIT, P. H. T. A model based on cellular automata to estimate the social isolation impact on covid-19 spreading in brazil. *Computer Methods and Programs in Biomedicine*, v. 200, 2021. Citado na pág. 11, 12, 14, 17.

SEZER, O. B.; GUDELEK, M. U.; OZBAYOGLU, A. M. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied Soft Computing Journal*, v. 90, 2020. Citado na pág. 19.

SHAHID, F.; ZAMEER, A.; MUNEEB, M. Predictions for covid-19 with deep learning models of lstm, gru, and bi-lstm. *Chaos, Solitons & Fractals*, v. 140, 2020. Citado na pág. 20, 36, 37.

SHAHID, F.; ZAMEER, A.; MUNEEB, M. Predictions for covid-19 with deep learning models of lstm, gru and bi-lstm. *Chaos, Solitons and Fractals*, v. 140, 2020. Citado na pág. 20, 36.

SIMSKE, S. J. *Meta-Algorithmics: Patterns for Robust, Low-Cost, High-Quality Systems*. [S.l.]: IEEE Press, 2013. Citado na pág. 12.

SINGANAYAGAM, A. et al. Community transmission and viral load kinetics of the sars-cov-2 delta (b.1.617.2) variant in vaccinated and unvaccinated individuals in the uk: a prospective, longitudinal, cohort study. *The Lancet Infectious Diseases*, v. 22, 2022. Citado na pág. 11.

SLAVOV, S. N. et al. Dynamics of sars-cov-2 variants of concern in vaccination model city in the state of sao paulo, brazil. *Viruses*, v. 14, 2022. Citado na pág. 11.

ULLAH, A. et al. Action recognition in video sequences using deep bi-directional lstm with cnn features. *IEEE Access*, v. 6, 2017. Citado na pág. 13, 19.

VADYALA, S. R. et al. Prediction of the number of covid-19 confirmed cases based on k-means-lstm. *Array*, v. 11, 2021. Citado na pág. 20.

VOLZ, E. et al. Assessing transmissibility of sars-cov-2 lineage b.1.1.7 in england. *Nature*, v. 593, p. 266–269, 2021. Citado na pág. 11.

VRIES, G. de et al. *A Course in Mathematical Biology*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2006. Citado na pág. 12.

WANG, G. et al. Application of a long short-term memory neural network: A burgeoning method of deep learning in forecasting hiv incidence in guangxi, china. *Epidemiology and Infection*, v. 147, 2019. Citado na pág. 13, 19, 36.

WANG, G. et al. Application of a long short-term memory neural network: A burgeoning method of deep learning in forecasting hiv incidence in guangxi, china. *Epidemiology and Infection*, v. 147, 2020. Citado na pág. 19, 36.

WANG, Q. et al. Scene classification with recurrent attention of vhr remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, v. 57, 2019. Citado na pág. 19.

WOLF, T. et al. Transformers: State-of-the-art natural language processing. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, 2020. p. 38–45. Citado na pág. 37.

XIANG, H.; LIU, B. Solving the inverse problem of an sis epidemic reaction-diffusion model by optimal control methods. *Computers and Mathematics with Applications*, v. 70, 2015. Citado na pág. 12.

XU, S. et al. A safety study evaluating non-covid-19 mortality risk following covid-19 vaccination. *Vaccine*, v. 41, 2023. Citado na pág. 12, 14.

YAN, B. et al. An improved method for the fitting and prediction of the number of covid-19 confirmed cases based on lstm. *Computers, Materials & Continua*, v. 64, 2020. Citado na pág. 20, 36.

YILDIRIM, A novel wavelet sequences based on deep bidirectional lstm network model for ecg signal classification. *Computers in Biology and Medicine*, v. 96, 2018. Citado na pág. 13, 19.

YUDISTIRA, N. et al. Learning where to look for covid-19 growth: Multivariate analysis of covid-19 cases over time using explainable convolution–lstm. *Applied Soft Computing*, v. 109, 2021. Citado na pág. 20.

ZHA, W. et al. Forecasting monthly gas field production based on the cnn-lstm model. *Energy*, v. 260, 2022. Citado na pág. 13, 19.

ZHANG, J. et al. Developing a long short-term memory (lstm) based model for predicting water table depth in agricultural areas. *Journal of Hydrology*, v. 561, 2018. Citado na pág. 13, 19.

ZHANG, Y. et al. Long short-term memory recurrent neural network for remaining useful life prediction of lithium-ion batteries. *IEEE Transactions on Vehicular Technology*, v. 67, 2018. Citado na pág. 13, 19.

ZHAO, J.; MAO, X.; CHEN, L. Speech emotion recognition using deep 1d & 2d cnn lstm networks. *Biomedical Signal Processing and Control*, v. 47, 2019. Citado na pág. 13, 19.

ZHOU, L. et al. Improved lstm-based deep learning model for covid-19 prediction using optimized approach. *Engineering Applications of Artificial Intelligence*, v. 122, 2023. Citado na pág. 20.

APÊNDICES

Tabela 1 – *Disponibilidade de dados para o Brasil no conjunto de dados do OWID COVID com percentual de preenchimento acima de 50%*

Variável	Dados não nulos	Total de dados	% de preenchimento
new cases	1427	1428	99.93
new vaccinations	636	1428	44.54
new tests	265	1428	18.56
new deaths	1428	1428	100.00
male smokers	1428	1428	100.00
gdp per capita	1428	1428	100.00
cardiovasc death rate	1428	1428	100.00
diabetes prevalence	1428	1428	100.00
female smokers	1428	1428	100.00
aged 70 older	1428	1428	100.00
hospital beds per thousand	1428	1428	100.00
life expectancy	1428	1428	100.00
human development index	1428	1428	100.00
population	1428	1428	100.00
aged 65 older	1428	1428	100.00
extreme poverty	1428	1428	100.00
new deaths per million	1428	1428	100.00
median age	1428	1428	100.00
population density	1428	1428	100.00
new cases per million	1427	1428	99.93
new deaths smoothed	1423	1428	99.65
new deaths smoothed per million	1423	1428	99.65
new cases smoothed per million	1422	1428	99.58
new cases smoothed	1422	1428	99.58
total cases	1373	1428	96.15
total cases per million	1373	1428	96.15

Continua na próxima página

Tabela 1 – *Disponibilidade de dados para o Brasil no conjunto de dados do OWID COVID com percentual de preenchimento acima de 50%*

Variável	Dados não nulos	Total de dados	% de preenchimento
total deaths per million	1352	1428	94.68
total deaths	1352	1428	94.68
stringency index	1094	1428	76.61
reproduction rate	1025	1428	71.78
new vaccinations smoothed	794	1428	55.60
new people vaccinated smoothed per hundred	794	1428	55.60
new people vaccinated smoothed	794	1428	55.60
new vaccinations smoothed per million	794	1428	55.60

Tabela 1 – *Variáveis*