

Predição de séries temporais da COVID19: uma avaliação de redes neurais com células LSTM

Saulo Joel Oliveira Leite
Programa de Pós-Graduação em
Engenharia Elétrica
Universidade Federal do Pará
Belém, Brasil
sauloleite@ufpa.br

Roberto Célio Limão de Oliveira
Programa de Pós-Graduação em
Engenharia Elétrica
Universidade Federal do Pará
Belém, Brasil
limao@ufpa.br

Lídio Mauro Lima de Campos
Faculdade de Computação
Universidade Federal do Pará
Belém, Brasil
limadecampos@gmail.com

Resumo— Neste trabalho será discutido como foram desenvolvidas implementações do modelo de Redes Neurais LSTM, para predição de séries temporais de casos confirmados e mortes por COVID-19, com a biblioteca de Deep Learning Keras. A COVID-19 é a doença causada pelo coronavírus denominado SARS-CoV-2, que tem acarretado num grande número de infectados em nível global. Como se fez necessário o uso de dados fidedignos, a base de dados utilizada para o desenvolvimento desse trabalho, é de domínio público e foi cedida pela Universidade Johns Hopkins. Os dados de casos confirmados e mortes do Brasil, Índia e Estados Unidos da América foram comparados e selecionados para realização de predições. As redes neurais Long Short-Term Memory são capazes de aprender longas sequências de observações para, deste modo, realizarem previsões. Assim sendo, o modelo construído obteve bons resultados para a predição de dados referentes a casos confirmados e mortes da pandemia de COVID-19.

Palavras-chave— COVID-19, Deep Learning, Keras, LSTM.

I. INTRODUÇÃO

A COVID-19 é uma doença que ficou mundialmente conhecida como o novo coronavírus. Existem outros membros da família de coronavírus, como o Sars, o Mers e outros agentes infecciosos responsáveis por resfriado comuns. Essa doença alcançou o posto de uma das maiores pandemias da história por conta da taxa mortalidade considerável, a capacidade de transmissão (inclusive entre assintomáticos) e os diferentes sintomas provocados [1]. A quantidade de pessoas infectadas pelo mundo chegou a 166,197 milhões de pessoas, com 3,444 milhões de mortes, segundo levantamento feito pela universidade norte-americana de Johns Hopkins [2].

Nesta pandemia, a tecnologia se tornou um dos principais pilares de auxílio ao combate ao coronavírus. Atualmente, as soluções utilizadas nesse período podem perdurar por anos, auxiliando neste e em outros desafios que estão por vir. Uma pesquisa realizada pela IBM constatou que quase um terço (31%) dos profissionais de TI do mundo dizem que os negócios em que atuam estão usando inteligência artificial (IA), atualmente, e 43% que aceleraram a implementação da tecnologia devido à pandemia de COVID-19 [3].

Diante deste cenário, é evidente a grande necessidade de pesquisas, que possam trazer propostas de suporte nas predições realizadas com os dados, disponibilizados para estudos. Neste trabalho, será descrito como foi desenvolvido um modelo de predição de casos e mortes por COVID19. Os dados utilizados para a pesquisa são disponibilizados pela Universidade Johns Hopkins e são atualizados diariamente.

Esta instituição tem se comprometido com a fidelidade dos dados, sendo uma das mais determinadas no auxílio do combate à pandemia.

II. FUNDAMENTAÇÃO TEÓRICA

A. Rede Neural Long Short-Term Memory (LSTM)

A rede neural *Long Short-Term Memory*, ou rede LSTM, é uma rede neural recorrente que é treinada usando *Backpropagation Through Time* e supera o problema do gradiente de desaparecimento. Esse gradiente ocorre porque redes neurais convencionais não “armazenam informações”. Dessa forma, a rede LSTM pode ser usada para criar grandes redes recorrentes que, por sua vez, podem ser usadas para resolver problemas de sequência difíceis no aprendizado de máquina [4].

Em vez de neurônios, as redes LSTM têm blocos de memória que são conectados por meio de camadas. Um bloco possui componentes que o tornam mais inteligente do que um neurônio clássico e uma memória para sequências recentes. Um bloco contém portas que gerenciam o estado e a saída do bloco. Um bloco opera em uma sequência de entrada e cada porta dentro de um bloco usa as unidades de ativação sigmoide para controlar se elas são disparadas ou não, tornando a mudança de estado e adição de informações que fluem através do bloco condicional [4].

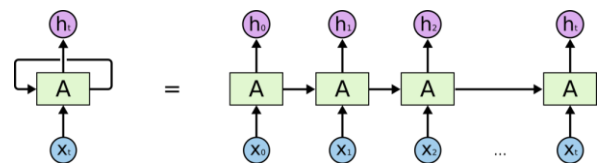


Fig. 1. Funcionamento da Rede Neural LSTM [5].

Como mostra a Fig.1, uma rede neural recorrente (RNR) pode ser imaginada como múltiplas cópias da mesma rede, cada uma passando uma mensagem a um sucessor. É necessário entender que a chave para a LSTM é o estado da célula. Ou seja, a linha horizontal que passa pelo diagrama. O estado da célula é como uma correia transportadora [5].

A LSTM tem a capacidade de remover ou adicionar informações ao estado da célula, cuidadosamente reguladas por estruturas chamadas portas. Vale destacar que, quanto mais blocos de células LSTM a rede neural possuir, maior a sua possibilidade de “aprender” com o tempo [5].

Para entender o modelo de forma matemática, as equações a seguir resumem tudo o que foi discutido acima.

$$h_t = \emptyset(b_h + xW_x + h_{t-1}W_h) \quad (1)$$

Para um melhor entendimento da equação (1), os seus componentes serão descritos. O h_t é o vetor de estado oculto, também conhecido como vetor de saída da unidade LSTM. O \emptyset é a função de ativação. O b_h e W_h são os pesos das conexões de entrada. Por fim, tem-se o h_{t-1} que é o estado oculto anterior do vetor de saída [6].

Nota-se que a diferença fundamental entre as equações acima e as das redes neurais clássicas é que são acionadas informações do estado oculto do período anterior. Além disso, observa-se que os parâmetros que fazem a transição da informação entre os estados ocultos, de diferentes períodos, são sempre os mesmos. Isso mostra que redes neurais recorrentes compartilham parâmetros através do tempo. Na prática, define-se a recorrência acima como um loop no código [7].

Caso ainda não tenha ficado compreensível, no exemplo a seguir, as equações da RNR foram escritas de forma desdobrada no tempo. Desse modo, é considerada uma RNR com uma camada oculta, processando quatro períodos e realizando uma previsão de uma variável contínua (problema de regressão), apenas após observar os 4 períodos [6]:

$$H_0 = \emptyset(b_h + xW_x + XW_x) \quad (2)$$

$$H_1 = \emptyset(b_h + xW_x + XW_x + H_0W_h) \quad (3)$$

$$H_2 = \emptyset(b_h + xW_x + XW_x + H_1W_h) \quad (4)$$

$$H_3 = \emptyset(b_h + xW_x + XW_x + H_0W_h) \quad (5)$$

Vale ressaltar que, não é necessário escrever manualmente as equações, já que os frameworks de Deep Learning, como o TensorFlow, realizam as operações computando as operações matemáticas automaticamente, com bases nos dados recebidos [8]. No caso desta pesquisa, são utilizados dados de séries temporais de casos confirmados e mortes por COVID19.

B. COVID-19

Em 31 de dezembro de 2019, a OMS tomou conhecimento de novo vírus, em Wuhan, na República Popular da China. Em 11 de fevereiro de 2020, ela estabeleceu a nomenclatura oficial, quando o vírus foi denominado coronavírus-2 da síndrome respiratória aguda grave (SARS-CoV-2) e a doença infecciosa do coronavírus-19 (COVID-19) [9].

Dentre os coronavírus já identificados, o SARS-CoV-2 é o sétimo identificado a causar doenças em humanos [10], sendo o terceiro a determinar uma epidemia, após a síndrome respiratória aguda grave (SARS-CoV) e a síndrome respiratória do Oriente Médio (MERS-CoV), que cursavam com sintomas graves de vias aéreas e alta taxa de mortalidade (10%-30%). Os outros quatro, considerados endêmicos (HKU1, OC43, 229E e NL63), apresentam sintomas leves semelhantes a um resfriado e são responsáveis por aproximadamente 10% das doenças sazonais das vias aéreas não causadas pela influenza [11].

O número de infectados por COVID-19 é muito variável nos diversos países, conforme as medidas de enfrentamento, o que depende da realização dos testes diagnósticos, do

distanciamento social, da população, do nível de educação e das medidas governamentais [12].

A velocidade de aumento do número de casos e mortes é elevada. O *Centers for Disease Control and Prevention* (CDC) observou que o território dos EUA alcançou rapidamente em dois meses, 1/3 dos casos mundiais, que se reduziu para 1/4 dos casos mundiais, em julho/2020, devido ao crescimento de casos em outros países [12].

O número de mortes também se encontra subestimado. Considerando os dados do Sistema de Informação da Vigilância Epidemiológica (SIVEP) da Gripe no Brasil, havia quase 50.000 mortes, ao final de agosto/2020, por síndrome respiratória aguda grave (SRAG) não especificada, pois não foram testados para COVID-19 [13].

Do número de casos no Brasil, a Região Sudeste sempre apresentou o maior número, seguida pela Região Nordeste, Norte, Sul e Centro-Oeste. O epicentro da doença iniciou-se no Estado de São Paulo, seguido pelos Estados do Amazonas, Ceará e Pernambuco e em setembro de 2020, mantêm-se em São Paulo, Minas Gerais, Rio Grande do Sul e Bahia [14].

Considera-se o período médio de incubação do SARS-CoV-2 de cinco dias, variando de zero a 14 dias. Estudos encontraram uma mediana do período de incubação 5,0 a 6,5 dias, variando de zero a 24 dias [15]. Esse período de incubação longo pode favorecer o aumento do risco de transmissão. A mediana do início dos sintomas ao óbito foi de 14 dias e 97,5% dos pacientes desenvolvem os sintomas em 11,5 dias da infecção [16].

O quadro clínico dos pacientes infectados por SARS-CoV-2 é muito variável, apresentando-se desde pacientes assintomáticos, passando por quadros leves até graves. Aproximadamente 80% dos casos são leves a moderados com cura espontânea. A frequência dos casos assintomáticos ainda é desconhecida [17].

Os sintomas clínicos mais frequentes são febre (87,9%), tosse (66,7%) e fadiga (38,1%). Outros sintomas incluem dispneia, dor de cabeça, astenia, mialgia, odinofagia, congestão/descarga nasal, anosmia, ageusia, síncope, confusão, conjuntivite, olho seco e erupção cutânea [18]. Uma porcentagem menor de pacientes relata diarreia, vômito e dor abdominal como sintomas relevantes. Em um estudo observou-se que sintomas gastrointestinais estavam presentes em 11,4% [30] e, em outro estudo, manifestações neurológicas foram observadas em 36,4% dos pacientes, além de manifestações cardíacas, como arritmias, e disfunção hepática, em até 50% dos casos [19].

No dia 26 de maio, o Brasil registrou 2.399 mortes, segundo dados reunidos pelo consórcio de veículos de imprensa. Chegando a 454.623 o total de óbitos decorrentes da doença [20]. A seguir, será discutido como foi desenvolvimento deste trabalho.

III. METODOLOGIA

A metodologia está dividida em partes que descrevem como realizou-se a aquisição e preparação do conjunto de dados em algo que pôde ser usado para predição de séries temporais. Contudo, é descrito como implementou-se uma rede neural LSTM, para um problema de predição de séries temporais de casos confirmados e mortes por COVID-19.

Portanto, após seguir essas etapas, discute-se como foi feita a predição.

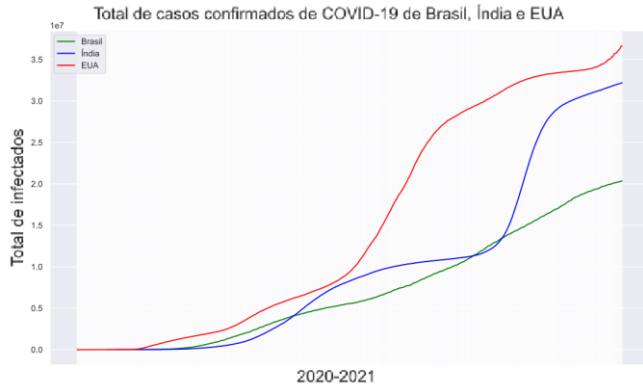


Fig. 2. Casos confirmados de COVID-19, no Brasil, Índia e EUA, de 2020 e 2021.

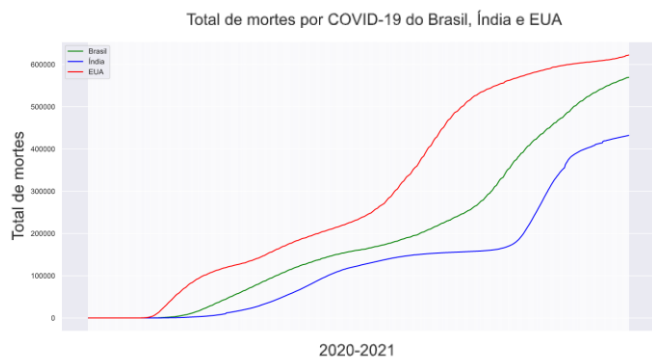


Fig. 3. Mortes COVID-19, no Brasil, Índia e EUA, de 2020 e 2021.

A. Aquisição e preparação de dados

A primeira etapa é realizar a aquisição dos dados. Os dados utilizados são disponibilizados pela Universidade de Johns Hopkins. Essa instituição tem se dedicado em disponibilizar dados relacionados a COVID-19, sendo umas das primeiras a iniciar a distribuição destes [20]. Como esses dados recebidos são de diversos países, o primeiro passo é tratá-los e extrair somente os dados que são necessários para este trabalho. Sendo estes os dados do Brasil, Índia e Estados Unidos da América.

Os dados obtidos são em forma de série temporal. Uma série temporal corresponde a uma coleção de observações feitas sequencialmente ao longo do tempo [8]. No caso deste trabalho, as séries temporais são das quantidades diárias de casos confirmados e mortes por COVID-19 no Brasil, Índia e EUA ao decorrer do período de janeiro de 2020 a junho de 2021. Na Fig. 2 é possível observar, de forma gráfica, os dados obtidos de casos confirmados por COVID-19 no Brasil, Índia e EUA. Entretanto, na Fig. 3 é mostrado o número de mortos na pandemia nesses mesmos países.

Os dados da Tabela I foram retirados da base de dados da Universidade de Johns Hopkins, na data de 11/06/2021. Analisa-se que os Estados Unidos da América é o país com a maior quantidade de casos confirmados e de mortes por COVID-19. Na análise, a Índia ocupa o segundo lugar, com o número de caso confirmados abaixo dos EUA. No entanto, o

Brasil é o segundo país com maior quantidade de mortes entre os três, atrás dos EUA.

TABELA I. TOTAL DE CASOS E MORTES POR COVID19

Total de casos confirmados e mortes por COVID19		
<i>Brasil</i>	<i>Índia</i>	<i>EUA</i>
17.296.118 casos confirmados	29.274.823 casos confirmados	33.449.437 casos confirmados
486.272 mortes	370.384 mortes	599.665 mortes

É possível expressar o problema por meio de regressão. Foi desenvolvida uma função para converter uma única coluna de dados em um conjunto de dados de duas colunas: a primeira coluna contendo a contagem de dias, desde 2020 até junho de 2021 e a segunda coluna contendo a contagem de casos confirmados/mortos por COVID19 dos próximos meses, a ser prevista [8]. Assim sendo, seguiu-se para a próxima etapa de desenvolvimento da rede neural LSTM.

B. Desenvolvendo a Rede Neural Long Short-Term Memory

Depois de modelar os dados e estimar a habilidade do modelo no conjunto de dados de treinamento, é necessário ter uma ideia da habilidade do modelo. Para uma classificação normal ou problema de regressão, precisa-se utilizar validação cruzada [7].

Com dados de séries temporais, a sequência de valores é importante. O método utilizado foi dividir o conjunto de dados ordenado em conjuntos de dados de treinamento e teste. Ao desenvolver a implementação de um modelo de predição, uma boa prática é realizar a separação dos dados em conjuntos de dados de treinamento com 67% do total de dados adquiridos, sendo dados utilizados para treinar o modelo e deixando os 33% restantes para testar o modelo [8]. A ação descrita foi realizada neste trabalho. Assim sendo, o modelo foi implementado com 200 épocas.

Contudo, a seguir, será discutido acerca dos hiper parâmetros, que são de suma importância para a realização da predição. Assim sendo, são eles que definem como será o desempenho do modelo com bases nos dados de treino e teste.

C. Hiper parâmetros da Rede Neural LSTM e método de avaliação de desempenho

O otimizador utilizado no código deste trabalho foi o *Adam*. A otimização de *Adam* é um método estocástico de descida gradiente, baseado na estimativa adaptativa, de momentos de primeira e segunda ordem [21]. Esse método foi escolhido, pois é computacionalmente eficiente, tem pouca necessidade de memória, invariante para redimensionamento diagonal de gradientes e é adequado para problemas que são grandes em termos de dados/parâmetros [22].

Além do otimizador, outro método de hiper parâmetro importante utilizado nesta rede neural foi o *Dropout*. Essa é uma técnica de regularização para reduzir o sobre-ajuste em redes neurais artificiais, evitando coadaptações complexas nos dados de treinamento. É uma maneira eficiente de realizar a média do modelo com redes neurais [23].

O método de avaliação de desempenho da rede neural LSTM foi o de erro quadrático médio, também conhecido EQM. Esse tem como principal objetivo encontrar a diferença média de um valor e o seu parâmetro inicial. O seu uso é destinado a compreender um "erro de previsão" [24]. Quanto

menor for o erro quadrático em relação, melhor foi a previsão.

D. Search Grid

Para decidir qual foi o melhor resultado obtido pelo modelo, foi realizada um *Search Grid* (grade de experimentos), modificando os valores dos hiper parâmetros, mencionados anteriormente. Os parâmetros são os de taxa de aprendizagem, valor do *dropout* e número de células da rede neural desenvolvida. A Tabela II, a seguir, mostra quais valores foram experimentados.

TABELA II. SEARCH GRID

Search grid		
Taxa de aprendizagem	Valor do dropout	Número de células LSTM
0.01	0.05	50
0.02	0.1	100
0.03	0.15	200
0.06	0.2	300

Todos os valores da tabela de *search grid* foram testados, para que pudesse ser avaliado qual combinação teve o melhor desempenho. São 64 possibilidades e para garantir que todas essas fossem testadas, foram realizados laços de repetição onde todos os valores da Tabela II foram testados em previsões realizadas.

Contudo, para a avaliação de qual foi a melhor combinação de valores do hiper parâmetros, foi desenvolvida uma condição para escolher a predição que obteve o menor erro quadrático médio. Após 200 épocas, os dados foram armazenados para que pudessem ser avaliados no final da implementação do modelo. Deste modo, escolheu-se a combinação que alcançou os melhores dados de predição.

Portanto, com este código desenvolvido é possível gerar previsões usando o modelo para o conjunto de dados de treino e teste para obter uma indicação visual da habilidade do desse. Além disso, pode-se estimar o desempenho do modelo nas duas etapas, por meio do cálculo do erro médio quadrático. Os dados que são avaliados pelo modelo são dos números diários (série temporal) de casos confirmados e mortes por COVID-19. Os resultados obtidos dos modelos com esses dados serão melhor discutidos na análise dos resultados.

IV. ANÁLISE DOS RESULTADOS

A análise de resultados está dividida em dois tópicos, sendo esses os resultados obtidos com a predição de casos confirmados de COVID-19 e a predição de mortes por COVID-19. Contudo, foram dois modelos implementados um para casos confirmados e outro para mortes confirmadas, obtendo assim seis resultados, dos três países analisados.

A. Predição de casos confirmados de COVID-19

Após a realização da implementação dos modelos, inicialmente as previsões foram realizadas para os casos confirmados de COVID-19 no Brasil, Índia e EUA. Estes três foram selecionados para que pudesse haver uma comparação de desempenho do modelo em diferentes países. Abaixo o

gráfico mostra o desempenho do modelo com base nos dados de casos confirmados no Brasil.

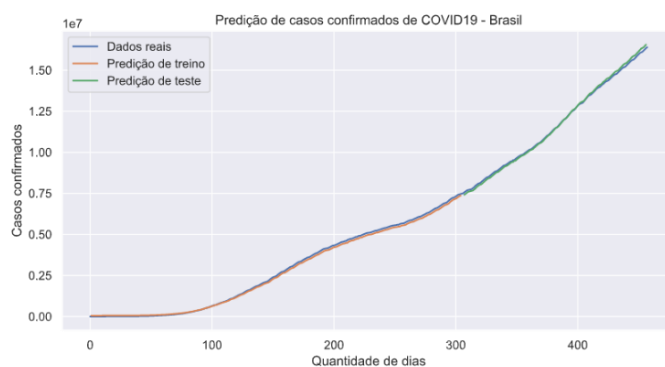


Fig. 4. Predição de casos confirmados de COVID-19 no Brasil.

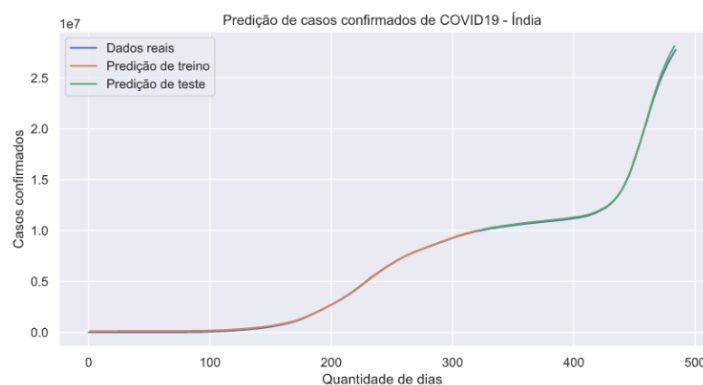


Fig. 5. Predição de casos confirmados de COVID-19 na Índia.



Fig. 6. Predição de casos confirmados de COVID-19 nos EUA.

É possível observar nas figuras que houve um grande aumento de casos no ano de 2021. Em cada uma das figuras a linha azul representa os dados reais, a linha laranja a predição de treino e a linha verde a predição de teste. Pode-se concluir que os dados de treinos e de testes preditos ficaram bem parecidos com os dados reais. Portanto, o modelo teve um bom desempenho.

A Fig 4 demonstra os resultados dos dados de casos confirmados do Brasil. Nesta predição, os hiper parâmetros que obtiveram um menor EQM desse resultado são: 0.01 de taxa de aprendizagem, 0.05 de *dropout* e 300 blocos de células LSTM.

A Fig. 5 demonstra a predição do modelo com os dados da Índia. Conclui-se que os hiper parâmetros que alcançaram um melhor resultado para Índia foram: 0.02 de taxa de aprendizagem, 0.15 de *dropout* e 300 blocos de células LSTM.

Por fim, na Fig 6, nos EUA, os hiper parâmetros que obtiveram melhor desempenho foram: 0.02 de taxa de aprendizagem, 0.2 de *dropout* e 300 blocos de células LSTM.

Conforme demonstrado nos gráficos de predições de casos confirmados do Brasil, Índia e EUA, é possível observar que as predições realizadas pelos modelos estão bem próximas dos dados reais obtidos. A Tabela III descreve os erros médios quadráticos de cada um dos modelos, na etapa de Treino e de Teste.

TABELA III. RESULTADOS OBTIDOS DE CASOS CONFIRMADOS DE COVID

Erro médio quadrático de predição de casos confirmados			
Scores	<i>Brasil</i>	<i>Índia</i>	<i>EUA</i>
Treino	21.524 casos	45.826 casos	91.433 casos
Teste	165.919 casos	331.856 casos	382.391 casos

Contudo, para uma melhor visualização e compreensão, a Tabela VI expõem os valores dos hiper parâmetros do modelo. Pode-se observar que conforme os formatos dos gráficos dos dados mudam, os hiper parâmetros também mudam.

TABELA IV. HIPER PARÂMETROS DO MODELO DE CASOS CONFIRMADOS DE COVID

Hiper parâmetros	<i>Brasil</i>	<i>Índia</i>	<i>EUA</i>
Taxa de aprendizagem	0.01	0.06	0.02
Valor de Dropout	0.05	0.15	0.2
Número de Células LSTM	300	300	300

Os valores do erro médio quadrático parecem ser significativos, mas demonstram uma eficiência de cerca de 99% do modelo. Haja vista, que os gráficos mostram um excelente desempenho. Para comparação, os valores totais podem ser conferidos na Tabela I, mencionada anteriormente. Assim sendo, há uma alta taxa de assertividade, tendo em vista que são milhões de casos confirmados.

B. Predição de mortes por COVID-19

O processo de predição de mortes teve o mesmo modelo implementado, com o mesmo método de análise de quais são os hiper parâmetros que mais se adequam à predição realizada, para casos confirmados de COVID19. No entanto, como o número de mortes é incontestavelmente menor, do que o de casos confirmados, os valores dos resultados dos erros quadráticos médios, desta rede neural LSTM, foram menores, em comparação a rede para casos confirmados, analisada anteriormente. Houve diferenças entre os hiper parâmetros dos dois modelos. As Fig. 7, Fig. 8 e Fig. 9 mostram visualmente como foi o desempenho do modelo. Por fim, a Tabela V e Tabela VI mostram os resultados e hiper parâmetros do modelo.



Fig. 7. Predição de de mortes por COVID-19 no Brasil.



Fig. 8. Predição de de mortes por COVID-19 nos EUA.

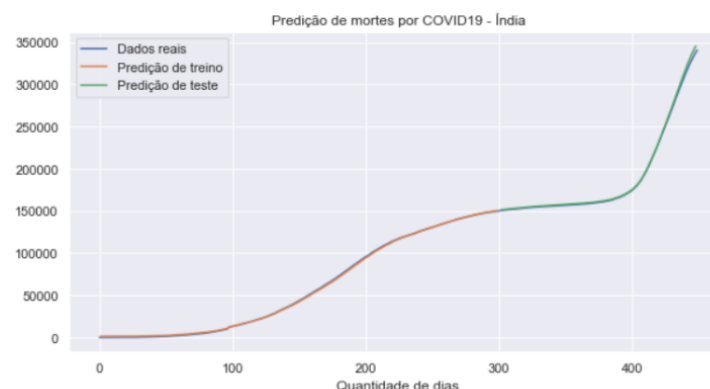


Fig. 9. Predição de de mortes por COVID-19 na Índia.

TABELA V. RESULTADOS OBTIDOS DE MORTES POR COVID

Erro médio quadrático de predição de mortes por COVID19			
Scores	<i>Brasil</i>	<i>Índia</i>	<i>EUA</i>
Treino	1389 mortes	557 mortes	1558 mortes
Teste	5423 mortes	3912 mortes	3758 mortes

TABELA VI. HIPER PARÂMETROS DO MODELO DE MORTES POR COVID

Hiper parâmetros	<i>Brasil</i>	<i>Índia</i>	<i>EUA</i>
Taxa de aprendizagem	0.03	0.01	0.01
Valor de Dropout	0.15	0.05	0.1
Número de Células LSTM	300	300	300

Como mostra a Tabela V, os valores obtidos pelo cálculo do erro médio quadrático são valores bem menores do que a predição anterior, por conta da quantidade menor de mortes em relação a casos confirmados. Todavia, o modelo obteve um excelente desempenho, com cerca 99% de eficiência.

Conforme foi analisado anteriormente, os hiper parâmetros que obtiveram o melhor desempenho mudam conforme o formato do gráfico. A Tabela VI mostra que cada país teve um hiper parâmetro diferente. Vale ressaltar que, os hiper parâmetros mudaram com relação aos que foram selecionados nas predições do modelo de casos confirmados, mesmo sendo dos mesmos países. Assim sendo, a implementação da Search Grid (grade de experimentos) mostrou-se indispensável.

Destaca-se que, tanto os resultados obtidos pelo modelo de casos confirmados, quanto o de mortes por COVID-19 apresentam um nível aceitável de erro quadrático médio e uma boa concordância entre si, implicando que o valor do EQM é uma boa estimativa do desvio padrão dos erros de previsão observados no conjunto de teste.

Portanto, é notório que visualmente os gráficos demonstram bom desempenho dos resultados obtidos, por meio da implementação do modelo. Contudo, os resultados obtidos por meio do modelo de casos confirmados e mortes por COVID19 comprovam a relevância desse modelo desenvolvido, para auxiliar na predição de dados da pandemia.

C. Comparando dados com outro método matemático para previsão da COVID19

Não foi encontrado em acervo público nenhum artigo que realizasse predições de casos confirmados e mortes por COVID-19 do começo de 2020 até junho de 2021, dos países analisados neste trabalho. A maioria dos artigos contam com predições de apenas alguns meses ao longo de 2020.

Apesar disso, como forma de avaliar o desempenho do modelo desenvolvido para este trabalho, foi realizada uma comparação dos resultados obtidos com os de outro artigo que segue a mesma premissa de previsão de casos de COVID-19.

No entanto, os resultados comparativos escolhidos são de um outro modelo preditivo, usando modelos de previsão baseados em redes neurais artificiais (RNA) para prever o surto de COVID-19. Contudo, os autores utilizaram o mesmo método de validação que este trabalho, o erro quadrático médio [25]. Os resultados desta análise e o período de dados usados estão nas Fig. 10, Fig. 11 e Fig. 12.

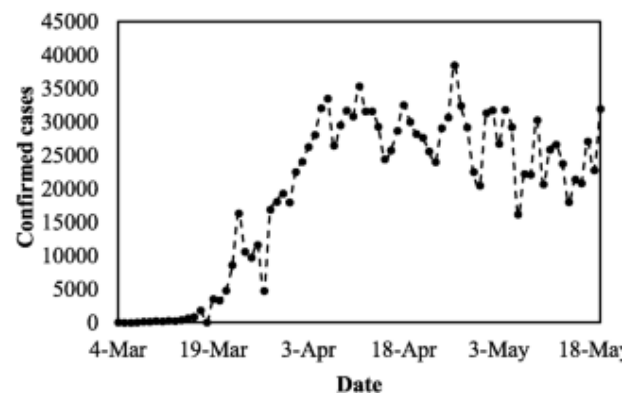
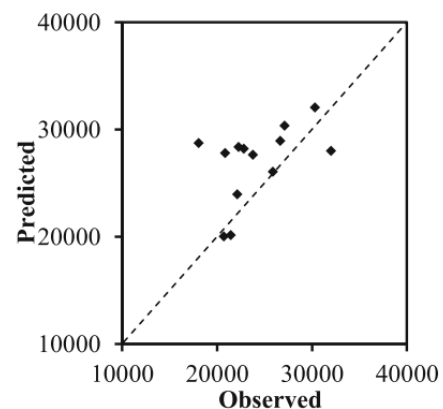


Fig. 10. Análise descritiva dos casos confirmados de COVID-19 nos EUA [25].

Ranking the ANN-based models for estimating confirmed cases of the test data

Models	Ranks			
	Italy	South Africa	USA	Overall
1st model	5	4	8	11
2nd model	1	14	3	5
3th model	14	7	13	12
4th model	12	5	2	4
5th model	7	11	1	2
6th model	8	12	12	13
7th model	13	3	11	6
8th model	10	10	5	8
9th model	9	6	6	9
10th model	11	12	13	14
11th model	4	8	9	10
12th model	3	1	7	2
13th model	1	9	10	7
14th model	5	2	4	1

Fig. 11. Classificação dos modelos baseados em ANN para estimar casos confirmados dos dados de teste [25].



(h) 5th model for USA (from 7 May to 19 May 2020)

Fig. 12. Casos confirmados diários previstos de COVID-19 [25].

A análise descritiva dos dados coletados pelos autores foi realizada em Excel que, segundo eles, oferece recursos robustos para a análise de dados e implementação de métodos numéricos. Conforme mostrado na Fig. 10, os dados dos EUA manifestam flutuações significativas nos casos confirmados diariamente. A ampla gama e variações consideráveis de dados exigem uma ferramenta poderosa para prever o surto de COVID-19 [25].

Ao analisar as Fig. 11 e Fig. 12, onde são mostradas as previsões no intervalo de 45 dias, de 4 de março a 18 de abril e o desempenho do modelo, pode-se observar que algumas previsões possuem um erro quadrático considerável. Os autores discursam em seu trabalho que isso indica que, considerando apenas uma métrica, essa pode não ser suficiente para avaliar o desempenho dos resultados de previsão. No entanto, várias métricas podem fornecer uma melhor perspectiva da precisão dos resultados [25].

Assim sendo, como mencionado anteriormente, o uso de ferramentas como Excel, limita os ajustes de parâmetros para a previsão [25]. O modelo desenvolvido por neste trabalho foi desenvolvido em Python, com o uso da biblioteca Keras. Nesse é possível ajustar hiper parâmetros para evitar os erros de flutuações e dados incongruentes. Desse modo fidelizando a previsão dos dados.

V. CONSIDERAÇÕES FINAIS

Resgatando o objetivo do presente trabalho, destaca-se que a principal contribuição que este trabalho pretende trazer consiste na apresentação de um modelo de Redes Neurais LSTM, com foco na previsão de séries temporais de casos confirmados e mortes por COVID19. Esse voltado para o auxílio à comunidade acadêmica, no combate contra a pandemia.

Pôde-se observar que o modelo conseguiu atingir o objetivo final de prever os dados da pandemia nos países Brasil, Índia e EUA. É importante ressaltar que o modelo pode ser mais uma ferramenta que pode ser utilizada para auxiliar no combate da COVID19. Deste modo, conseguindo maximizar o bem-estar da sociedade que vem sofrendo em decorrência da pandemia.

A implementação desenvolvida neste trabalho pode servir como modelo para outros países. O que se faz necessário é a aquisição e tratamento dos dados, para que possam ser aplicados no modelo de Redes Neurais LSTM. O *Search Grid* adequará os melhores hiper parâmetros com base nos dados recebidos, independente do país.

Portanto, como sugestões para trabalhos futuros, fica a possibilidade de avaliar o desempenho na previsão de dados de outros países em torno do mundo, como países da Europa e do oriente, por exemplo. Desse modo, visando aprimorar o resultado e atingir maior assertividade nas previsões, trazendo desta forma soluções tecnológicas cada vez melhores para a sociedade.

REFERÊNCIAS

- [1] Veja. (2021). *Nanogrids, microgrids, and Big Data: The future of the power grid* [Online]. Disponível: <https://saude.abril.com.br/bem-estar/covid-19-as-principais-perguntas-e-respostas>
- [2] Isto é. (2021). *Coronavírus: número de infectados pela Covid-19 no mundo chega a 166,1 milhões* [Online]. Disponível: <https://istoe.com.br/coronavirus-numero-de-infectados-pela-covid-19-no-mundo-chega-a-1661-milhoes>
- [3] CIO. (2021). *Pandemia fez 43% dos negócios globais acelerarem uso de IA* [Online]. Disponível: <https://cio.com.br/noticias/pandemia-fez-43-dos-negocios-globais-acelerarem-uso-de-ia>
- [4] Sara Filipa. (2020). *Compreendendo LSTM Networks* [Online]. Disponível: <https://www.cetax.com.br/blog/compreendendo-lstm-networks>
- [5] Jose Júnior. (2019). *Redes Neurais Recorrentes — LSTM* [Online]. Disponível: <https://medium.com/@web2ajax/redes-neurais-recorrentes-lstm-b90b720dc3f6>
- [6] Gers, FA; Schmidhuber, J. (2001). "Redes recorrentes LSTM aprendem linguagens simples livres de contexto e sensíveis ao contexto"(PDF). *IEEE Transactions on Neural Networks*. 12(6): 1333–1340. Disponível: <ftp://ftp.idsia.ch/pub/juergen/L-IIEEE.pdf>
- [7] Jason Brownlee. (2020). *Time Series Forecasting with the Long Short-Term Memory Network in Python* [Online]. Disponível: <https://machinelearningmastery.com/time-series-forecasting-long-short-term-memory-network-python>
- [8] Jason Brownlee. (2020). *Time Series Prediction with LSTM Recurrent Neural Networks in Python with Keras* [Online]. Disponível: <https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras>
- [9] World Health Organization. (2021). *Naming the coronavirus disease (COVID-19) and the virus that causes it* [Online]. Disponível: <https://www.who.int>
- [10] L. Wang et al., "Review of the 2019 novel coronavirus (SARS-CoV-2) based on current evidence", *Int J Antimicrob Agents*, vol. 1 Ago. 2020. Disponível: <https://doi.org/10.1016/j.ijantimicag.2020.105948>
- [11] T. Lam et al., "Identifying SARS-CoV- 2-related coronaviruses in Malayan pangolins", em *Nature*, 2020, pp. 282-285
- [12] World Health Organization. (2021) *WHO Coronavirus Disease (COVID-19) Dashboard* [Online]. Disponível: <https://covid19.who.int/>
- [13] openData SUS. (2021). *Banco de Dados de Síndrome Respiratória Aguda Grave - incluindo dados da COVID-19* [Online]. Disponível: <https://opendatasus.saude.gov.br/dataset/bd-srag-2021>
- [14] Ministério da Saúde. (2021) *Painel de casos de doença pelo coronavírus 2019 (COVID-19) no Brasil pelo Ministério da Saúde* [Online]. Disponível: <https://covid.saude.gov.br/>
- [15] W. Wang et al., "Updated understanding of the outbreak of 2019 novel coronavirus (2019 - nCoV) in Wuhan, China", em *J Med Virol*, 2020, pp. 441-447.
- [16] J. Backer, D. Klinkenberg, J. Wallinga, "Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20-28 January 2020", em *Euro Surveill*, 2020, pp. 25.
- [17] F. Stumpf, A. Titzmann et al., "SARS- CoV-2 Infection in Pregnancy – a Review of the Current Literature and Possible Impact on Maternal and Neonatal Outcome", em *Geburtshilfe Frauenheilkd*, 2020; pp. 380-390.
- [18] WHO (World Health Organization).(2021). *Report of the WHO- China Joint Mission on Coronavirus Disease 2019 (COVID-19)* [Online]. Disponível: <https://www.who.int/emergencies/diseases/novel-coronavirus-2019>
- [19] S. Wong, R. Lui, J. Sung, "Covid - 19 and the digestive system", em *J Gastroenterol Hepatol*, 2020, pp. 744-748.
- [20] G1. (2021). *Brasil registra 2.399 mortes por Covid em 24 horas* [Online]. Disponível: <https://g1.globo.com/jornal-nacional/noticia/2021/05/26/brasil-registra-2399-mortes-por-covid-em-24-horas.ghtml/>
- [21] Keras. (2021). *Adam*. Disponível: <https://keras.io/api/optimizers/adam/>
- [22] D. Kingma et al., "Adam: A Method for Stochastic Optimization", *Conrell University*, vol. 1, Jan. 2017. Disponível: <https://arxiv.org/abs/1412.6980>
- [23] Keras. (2021). *Dropout layer*. Disponível: https://keras.io/api/layers/regularization_layers/dropout/
- [24] Mais retorno. (2020). *EQM (Erro Quadrático Médio)*. Disponível: <https://maisretorno.com/porta/termos/e/eqm-erro-quadratico-medio>
- [25] Niazkar, H.R., Niazkar, M. Application of artificial neural networks to predict the COVID-19 outbreak. *glob health res policy*, *Global Health Research and Policy*, Vol. 1, nov 2020. Disponível: <https://doi.org/10.1186/s41256-020-00175-y>