Aprendizagem Profunda Aplicada à Classificação do SARS-CoV-2

Karolayne S. Azevedo*, Raquel de M. Barbosa*† e Marcelo A. C. Fernandes*‡
*Laboratório de Aprendizagem de Máquina e Instrumentação Inteligente, nPITI/IMD, UFRN, Natal, RN, Brasil.
†Laboratório de Desenvolvimento de Medicamentos, DFAR, UFRN, Natal, RN, Brasil.
‡Departamento de Engenharia da Computação e Automação, UFRN, Natal, RN, Basil.
Email: *karolayne.azevedo016@ufrn.br,†m.g.barbosafernandes@gmail.com,‡mfernandes@dca.ufrn.br

***Resumo*—Este artigo propõe uma técnica, baseada em aprendizado de máquina, que faz uso de uma rede neural convolucional (*Convolutional Neural Network - CNN*) profunda de uma dimensão (1D), destinada à classificação de genomas virais, capaz de identificar corretamente o vírus SARS-CoV-2, causador da doença COVID-19. Como entrada, foi utilizado amostras genômicas completas de DNAc (DNA complementar) do vírus da família *Coronaviridae*, extraídas a partir do repositório 2019 *Novel Coronavirus Resource* (2019nCoVR). A base de dados utilizada neste trabalho, contém 17.893 amostras de DNAc, cujo tamanho varia entre 26.342 bp e 31.029 bp (*base-pair* - bp). Ao contrário da maioria das abordagens apresentadas na literatura, os resultados obtidos por esta técnica, revelam valores máximos de precisão, sensibilidade, especificidade e F1-score, aplicados numa validação cruzada quíntupla, usada para avaliar o modelo. Os resultados obtidos, mostram-se mais confiáveis se comparados com os trabalhos discutidos no estado da arte, indicando que a ferramenta pode ser aplicada para a classificação de vírus da família *Coronaviridae*.**

***Index Terms*—SARS-CoV-2, COVID-19, classificação viral, aprendizagem profunda.**

I. Introdução

Um vírus em particular vem chamando a atenção do mundo inteiro, o SARS-CoV-2. Pertencente à família de vírus *Coronaviridae*, e é conhecida por conter um dos maiores genomas virais, que varia entre 26.000 bp a 31.700 bp [1]. O SARS-Cov-2 é causador da doença COVID-19 que vem ocasionando a morte de milhares de pessoas no mundo inteiro devido a sua rápida disseminação e alto índice de transmissibilidade [2], [3]. Assim, a classificação rápida e precisa do vírus, pode auxiliar na sua compreensão biológica e molecular, contribuindo para a criação de vacinas e medicamentos, além de auxiliar no diagnóstico, tratamento e prevenção de doenças [4], [5].

A classificação de vírus já é uma tarefa desenvolvida há muito tempo por cientistas do mundo inteiro. Esta atividade, consiste em atribuir uma determinada sequência a um respectivo grupo, a partir de sequências genômicas conhecidas, que compartilham características e traços em comuns [6]. Os métodos convencionais de extração de características do vírus para fazerem classificação viral são baseados em alinhamento de sequências [7], [8]. O alinhamento de sequências, é uma técnica que busca por regiões de similaridade entre sequências biológicas a partir de uma sequência de referência previamente caracterizada, assim por esta razão, as técnicas baseadas em alinhamento genético podem, também, serem utilizadas para identificação viral [6].

Algoritmos como BLAST [9], MALT [10], FASTA [11], ClustalW [12] e USEARCH [13] fazem uso das técnicas baseadas em alinhamento. Entretanto, os algoritmos que fazem uso desses métodos, apresentam limitações como: baixa acurácia e uso de sequências genômicas de tamanho limitado, tendo em vista o alto custo computacional ao utilizar sequências genômicas longas, devido a natureza do problema, como ressalta [7], [14]. Os trabalhos apresentados por [6] e [7] chamam atenção para o fato do uso de métodos baseados em alinhamento, não serem tão satisfatórios quando aplicados a genomas susceptíveis a grande variações genéticas, como é o caso da grande maioria dos vírus. Com intuito de minimizar estes problemas, surgiram os métodos sem alinhamento (*alignment-free* - AF) que se baseiam em recursos da álgebra linear, teoria da informação e mecânica estatística, para calcular a similaridade ou distância entre sequências [6], [7].

Segundo [6], [15] e [16] para obter melhores resultados, a classificação viral baseada em AF, aplicam recursos de inteligência artificial baseada aprendizagem de máquina (*Machine Learning* - ML) para realizar a extração de características nas sequências genômicas utilizadas. Estudos recentes, apontam que algoritmos e técnicas de ML, têm sido amplamente utilizada em pesquisas relacionadas a genômica, incluindo a classificação viral, por oferecer um conjunto de métodos, capazes de identificar padrões altamente complexos de forma automatizada, eficiente e com o mínimo de intervenção humana [17], [18].

Trabalhos embasados na literatura, tem mostrado que técnicas de aprendizagem de máquina baseadas em aprendizagem profunda (*Deep Learning* - DL) apresentam excelentes resultados para aplicações voltadas á sequências genômicas, incluindo problemas de classificação [19], [20]. Os trabalhos de [21] e [18] revelam que, dentre os mais diversos algoritmos de ML, as redes neurais convolucionais (*Convolutional Neural Network* - CNN) vem sendo bastante utilizadas para análises de dados com base em sequências genômicas, por serem capazes de extrair características intrínsecas das sequências, e apresentarem resultados promissores em suas aplicações. Contudo, a maior parte dessas ferramentas e técnicas, fazem uso de sequências genômicas de comprimento limitado, ou são voltadas para outras finalidades como predição de proteínas

[22], [23].

Diante desse contexto, o presente trabalho, tem como objetivo apresentar uma técnica capaz de realizar a classificação do vírus da família *Coronaviridae*. Esta técnica, faz uso de uma CNN que recebe como entrada, sequências genômicas completas de DNAc, codificadas por meio da técnica *one-hot encode*. Assim, a CNN realiza uma classificação binária que identifica a amostra como sendo SARS-CoV-2 ou não.

O artigo está organizado do seguinte modo: Na seção II foi discutido os trabalhos relacionados à pesquisa. Em seguida, a seção III explana a técnica utilizada no pré-processamento dos dados antes de entrar na CNN. A metodologia do trabalho é descrita com detalhes na seção IV. Na seção V é apresentado os resultados. Uma comparação entre os resultados obtidos no trabalho com o estado da arte é feita na seção VI e por fim, na seção VII é feita as considerações finais.

II. Trabalhos Relacionados

O trabalho apresentado por [24] propõem um classificador profundo de genoma viral, conhecido como VGDC, capaz de identificar subtipos virais de diferentes famílias como a dengue, hepatite B e C, HIV-1 e influenza A. Esta arquitetura apresentou F1-score entre $0{,}85$ a 1. Em [25] é apresentada uma arquitetura que reconhece a presença de vírus, por meio de contigs de metagenômicos brutos de diversas amostras humanas. A metodologia proposta foi chamada de ViraMiner e, apesar do uso de duas CNNs a metodologia proposta, obteve uma curva característica de operação do receptor (*Receiver Operating Characteristic Curve - ROC*) de $0{,}923$.

O uso de uma rede neural denominada miRNA, utilizada inicialmente para a detecção de câncer [26], foi utilizada para a classificação viral. A arquitetura apresenta poucas camadas e também foi utilizada para classificar vírus pertencentes à família *Coronaviridae*. Esse modelo apresentou um valor de acurácia de 98%, especificidade de 0,9939 e sensibilidade de 1,00 [20]. Um grande número de sequências genômicas virais de diversos tamanhos (300-500-1000 e 3000 bp) foram analisados por [27] que utilizou como métrica de desempenho a área sob as características operacionais do receptor (*Area Under the Receiver Operating Characteristics - AUROC*), e obteve valores de 0,95, 0,93 0,97 e 0,98 respectivamente. A arquitetura utilizada foi intitulada como DeepVirFinder e faz uso de uma rede neural convolucional de múltiplas camadas.

III. Metodologia

A. Base de Dados

O Centro Nacional de Dados Genômicos (*National Genomics Data Center - NGDC*) fornece acesso aberto e gratuito, a um conjunto de recursos de banco de dados que possui um recurso chamado Recurso de Dados do novo Coronavirus 2019 *New Coronavirus 2019 Data Resource - 2019nCoVR*. O *2019nCoVR* mantêm atualizações diárias e reúne uma coleção abrangente de sequências genômicas e informações clínicas não apenas do SARS-CoV-2, como também de vírus pertencentes à família *coronaviridae* do mundo inteiro e de outros repositórios tradicionais, como do Centro Nacional de Informações para Biotecnologia (*National Center for Biotechnology Information - NCBI*), sendo o repositório escolhido para fazer o *download* do conjunto de dados. Selecionou-se sequências pertencentes à família *coronaviridae* cujo tamanho variam entre 25.000 bp a 35.000 bp abrangendo o tamanho de todos os vírus da família, sem perder nenhuma informação genética importante. O hospedeiro selecionado foi *Homo Sapiens*.

A base de dados construída é formada por 17.893 amostras de sequências genômicas de nove tipos de vírus da família *coronaviridae*, oriundas de 62 países diferentes. A Figura 1 exibe todos os países que apresentam amostras genômicas pertencentes a base dados, observa-se que os Estados Unidos apresenta a maior quantidade de sequências, seguido da Austrália, Índia e China. Das 17.893 amostras, 17.392 pertencem ao vírus SARS-CoV-2 ($97{,}2\%$ do total) , sendo 11.140 amostras provenientes dos Estados Unidos ($62{,}25\%$ do total).

Os dados utilizados para a classificação viral são sequências de DNAc, cujo comprimento varia entre 26.342 bp a 31.029 bp. A Tabela I resume algumas propriedades relacionadas aos subtipos virais presentes na base de dados. O BetaCoronaVirus apresenta o maior comprimento entre todos os subtipos virais, variando entre 31.029 bp e 30.536 bp. Além de apresentar o mesmo comprimento (30.499 bp), o CoronaVirus cya-BetaCov/2019, CoronaVirus cyb-BetaCov/2019 e CoronaVirus cyc-BetaCov/2019 são os vírus que possuem a menor quantidade de amostras da base de dados. Por se tratarem de amostras genômicas longas e por serem vírus muito semelhantes, requerem o uso de um modelo robusto para serem devidamente classificados [24].

Tabela I
Subtipos virais presentes na base de dados criada para este trabalho.

Vírus	Quant. de Amostras	Comp. Mínimo da Sequência	Comp. Máximo da Sequência
BetaCoronaVirus	140	30.536	31.029
CoronaVirus cya-BetaCov/2019	1	30.499	30.499
CoronaVirus cyb-BetaCov/2019	1	30.499	30.499
CoronaVirus cyc-BetaCov/2019	1	30.499	30.499
HCoV-229E	27	26.592	27.307
HCoV-HKU11	18	29.367	29.983
HCoV-NL63	55	27.302	27.832
MERS-CoV	258	29.267	30.150
SARS-CoV-2	17.392	26.342	28.784

B. Balanceamento dos Dados

Como observado na Tabela I, a maior quantidade de amostras contida na base de dados pertence ao vírus SARS-CoV-2, causador da doença COVID-19, seguido do vírus MERS-CoV. Neste contexto, foi necessário realizar o balanceamento dos dados não somente para melhorar o desempenho da rede, mas também evitar problemas como *Overfitting* em virtude da desproporção das amostras dos demais vírus.

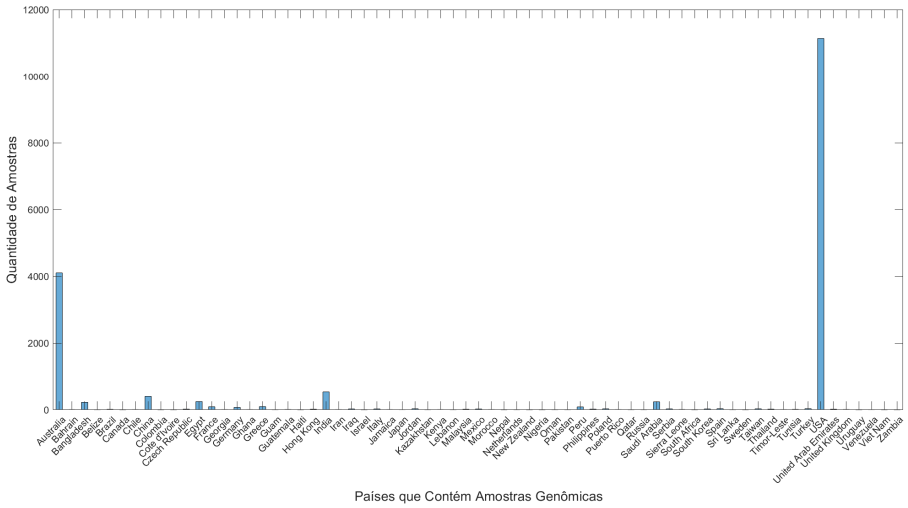

Figura 1. Países que contém amostras genômicas da família *Coronaviridae* na base de dados.

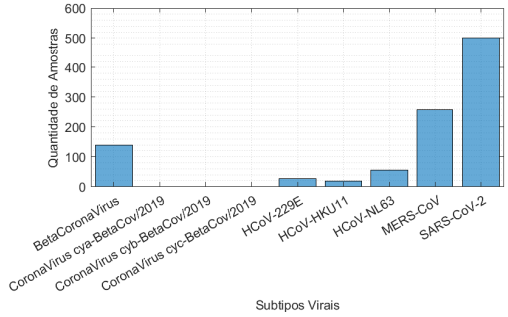

Figura 2. Conjunto de todos os subtipos virais após realizar o balanceamento das amostras.

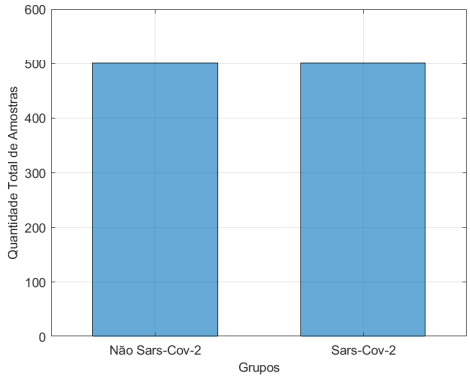

Figura 3. Conjunto de dados após realizar o balanceamento das amostras de acordo com seus grupos.

O conjunto de amostras foi dividido em dois grupos intitulados ”Não SARS-CoV-2”e ”SARS-CoV-2”conforme ilustra a Figura 2. O grupo Não SARS-CoV-2 é formado por oito subtipos virais diferentes do SARS-CoV-2, totalizando 501 amostras. Logo, 501 amostras foram retiradas de todos os países que apresentaram sequências genômicas do vírus SARS-CoV-2, de forma aleatória e uniforme, garantindo diversidade e representatividade de cada subtipo viral nos conjuntos de treinamento e validação, como ilustra a Figura 3. O conjunto de dados utilizado para etapa de treinamento e validação da rede, passa a conter 1.002 amostras pertencentes aos dois grupos, representados pelos rótulos 0 e 1, onde 0 está associado a outras espécies, e 1 está relacionado ao SARS-CoV-2. Parte das amostras genômicas restantes, foram utilizadas para testar a performance da rede.

C. Técnica de Mutação Artificial

Com intuito de investigar a sensibilidade e robustez da arquitetura proposta, a possíveis mutações que o vírus SARS-CoV-2 viria sofrer, foi aplicado a metodologia de mutações artificiais, proposta por [20], no qual os autores chamam de “ruídos“ aplicados em sequências genômicas. Para esse teste, foram utilizadas 10.000 amostras divididas aleatoriamente em dois grupos cada um contendo 5.000. Vale salientar que o método de mutação artificial foi aplicada em apenas um dos grupos.

O processo de mutação artificial é iniciado pela busca do maior comprimento entre as amostras, ou seja, para um conjunto de H amostras, tem-se que $V_{max} = \max\{N_1, \ldots, N_H\}$ onde N_i é tamanho das sequências e V_{max} é comprimento da maior sequência. Após está etapa, é realizado a inserção de zeros, em cada i-ésima sequencia, s_i, onde $N_i < V_{max}$. Cada i-ésima sequência é completada com zeros até atingir o valor de V_{max}, ou seja, a quantidade zeros inseridos é para a i-ésima sequência é $V_{max} - N_i$. Ao final desta etapa, todas as H amostras escolhidas têm o mesmo tamanho, V_{max}. Após esta etapa, define-se uma taxa de mutação artificial, definida aqui como γ. O valor de γ define a porcentagem da quantidade de nucleotídios que serão alterados, N_{mut}, que pode ser expressa

como

$$N_{\text{mut}} = \left\lfloor \frac{\gamma \times V_{max}}{100} \right\rfloor . \quad (1)$$

Após a definição do N_{mut}, é definida de forma aleatória a posição dos N_{mut} nucleotídeos que serão alterados, no qual é armazenada no vetor $\mathbf{k}_{\text{mut}} = [k_1, \ldots, k_{N_{\text{mut}}}]$. A partir do vetor de posições, $\mathbf{k}_{\text{mut}}$, são aplicados dois métodos para alterar os nucleotídeos selecionados para mutação artificial. O primeiro método é aplicado a primeira metade dos nucleotídeos selecionados, ou seja, as posições $\left[k_1, \ldots, k_{N_{\text{mut}}/2}\right]$ e o segundo foi aplicado a segunda metade do vetor de posição $\left[k_{N_{\text{mut}}/2+1}, \ldots, k_{N_{\text{mut}}}\right]$.

O primeiro método altera a posição dos nucleotídeos levando em considerando duplas, ou seja,

$$\begin{gathered}\left[k_1, k_2, \ldots, k_{N_{\text{mut}}/2-1}, k_{N_{\text{mut}}/2}\right] \Rightarrow \\ \left[k_2, k_1, \ldots, k_{N_{\text{mut}}/2}, k_{N_{\text{mut}}/2-1}\right].\end{gathered} \quad (2)$$

Já o segundo método, altera valores dos nucleotídeos para,

$$s_{k_i} = \begin{cases} \text{A} & \text{se } s_{k_i} = \text{T} \\ \text{T} & \text{se } s_{k_i} = \text{A} \\ \text{C} & \text{se } s_{k_i} = \text{G} \\ \text{G} & \text{se } s_{k_i} = \text{C} \\ \text{N} & \text{se } s_{k_i} = \text{T} \end{cases} . \quad (3)$$

D. Arquitetura da CNN

Com base na estratégia apresentada na Seção IV e os valores obtidos da base de dados criada, detalhados na Tabela I, tem-se que o $N_{\text{max}} = 31.029$. Assim, a cada m-ésima amostra a CNN terá como entrada 5 canais de dimensão 31.029×1. Como foi descrito na Seção IV esta estratégia permite que todos as M sequencias virais tenham o mesmo tamanho.

A CNN utilizada neste trabalho é composta por vinte e seis camadas divididas entre camadas convolucionais 1D, responsáveis pela extração de características das sequências genômicas de DNAc, e camadas totalmente conectadas, responsáveis pela classificação dos dados extraídos das camadas superiores gerando um total de $14.545.426$ de parâmetros ao longo de todas as camadas conforme apresentada na Tabela II. A Figura 4 detalha a estrutura da CNN utilizada no classificador viral adequado para a base de dados descrita na Seção III-A.

A CNN é formado por quatro camadas convolucionais, seguidas por uma camada de normalização e pela função de ativação ReLu (*Rectified Linear Unit*). A função *MaxPool* é aplicada após cada camada de ativação com janelas cujo tamanho variam entre 8, 16, 32 e 64. Além das camadas convolucionais a estrutura da CNN utilizada, contém quatro camadas totalmente conectadas com 64, 32, 16 e 2 neurônios respectivamente. A quantidade de neurônio na última camada corresponde à quantidade de classes a serem classificadas, seguida pela função *softmax* que fornecerá como saída a probabilidade de cada sequência pertencer a uma classe.

Tabela II
ARQUITETURA DA CNN UTILIZADA NESTE TRABALHO CONTENDO QUATRO CAMADAS CONVOLUCIONAIS E QUATRO CAMADAS TOTALMENTE CONECTADAS.

Camadas	Descrição	Valores
1	Input (L × 1 × 5)	N = 31030
2	Conv1d (K_1@B_1)	K_1 = 256 and B_1 = 8
3	BatchNorm	-
4	ReLU	-
5	MaxPool1D (P_s)	P_s = 8
6	Conv1D (K_2@B_2)	K_2 = 64 and B_2 = 16
7	BatchNorm	-
8	ReLU	-
9	MaxPool1D (P_s)	P_s = 16
10	Conv1D (K_3@B_3)	K_3 = 32 and B_3 = 8
11	BatchNorm	
12	ReLU	-
13	MaxPool1D (P_s)	P_s = 32
14	Conv1D (K_4@B_4)	K_4 = 32 and B_4 = 64
15	BatchNorm	-
16	ReLU	-
17	MaxPool1D (P_s)	P_s = 64
18	Flatten	-
19	Dense1 (P_1)	P_1 = 64
20	Dropout (a_1)	a_1 = 0.4
21	Dense2 (P_2)	P_2 = 32
22	Dropout (a_2)	a_2 = 0.4
23	Dense3 (P_3)	P_3 = 16
24	Dropout (a_3)	a_3 = 0.4
25	Dense4 (P_4)	P_4 = 2
26	SofMax	2 Classes

E. Treinamento

Para avaliar o modelo proposto, foi utilizado a validação cruzada k-fold, onde k se refere a quantidade de subconjuntos, ou dobras em que seu conjunto de dados será divido. Definiu-se o valor de $k = 5$, assim o conjunto de dados será dividido em 5 subconjuntos, cada fold contendo 201 amostras. No método de validação cruzada, $k-1$-folds são destinadas para o treinamento do modelo (801 amostras) e 1-fold destinado para validação do modelo (201 amostras) totalizando 1.002. O otimizador escolhido para a atualização dos pesos da rede foi o *adam*, cuja taxa de aprendizagem foi de 0,001. É importante ressaltar que o treinamento convergiu em aproximadamente 10 épocas. Os parâmetros utilizados na etapa de treinamento da arquitetura são observados na Tabela III-E.

Tabela III
HIPERPARÂMETROS UTILIZADOS NA FASE DE TREINAMENTO DA ARQUITETURA PROPOSTA

Parâmetros	**Valores**
Mini-Batches	128
MaxEpochs	12
InitialLearnRate	0,001
Otimizador	*adam*

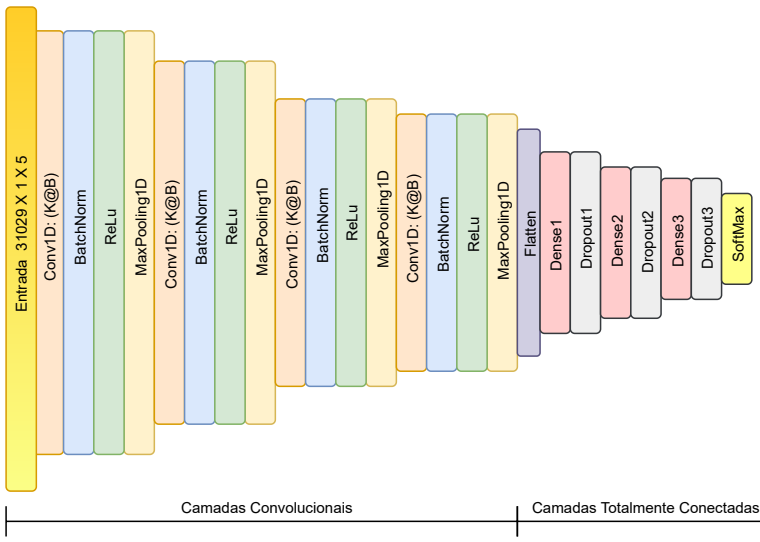

Figura 4. CNN utilizada para a proposta de classificador viral apresentada neste trabalho.

IV. Técnica Proposta

A Figura 5 ilustra o esquema da técnica proposta neste artigo, no qual, a partir de uma base de dados de M amostras de sequências virais de DNAc, cada m-ésima amostra da base, $\mathbf{s}_m$ é mapeada em uma matriz de característica, $\mathbf{S}_m$, e depois processada por uma CNN. A CNN faz uma classificação binária no qual identifica ou não o SARS-CoV-2.

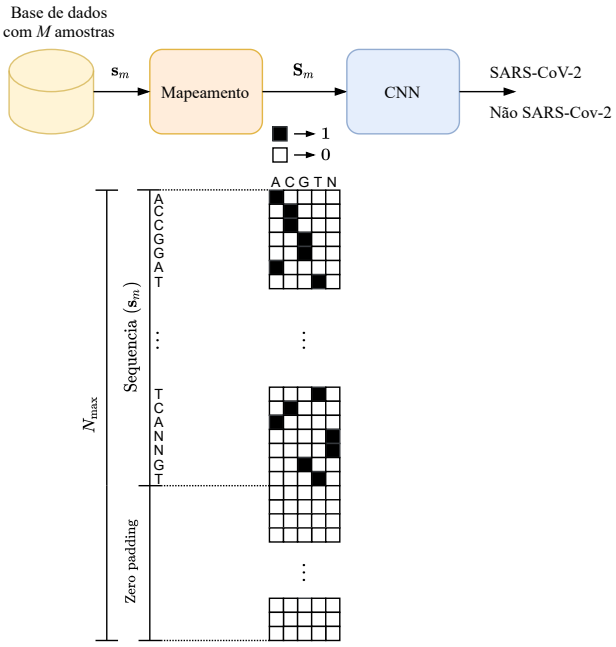

Figura 5. Visão geral da técnica proposta.

Cada m-ésima amostra de sequencia viral de entrada é expressa como

$$\mathbf{s}_m = [s_{1,m}, \ldots, s_{N_m,m}] \tag{4}$$

onde cada i-ésimo elemento de uma m-ésima amostra $s_{i,m}$ representa um possível nucleotídeo de conjunto $S \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ e N_m é o comprimento da m-ésima amostra de sequencia viral. Cada elemento de S corresponde a uma das bases nitrogenadas Adenina (A), Citosina (C), Guanina (G) e Timina (T).

A matriz de característica associada a m-ésima amostra $\mathbf{s}_m$ é construída com técnica de *one-hot encode*, no qual pode ser expressa como

$$\mathbf{A}_m = \begin{bmatrix} a_{1,1,m} & \cdots & a_{1,5,m} \\ \vdots & \ddots & \vdots \\ a_{N_{\max},1,m} & \cdots & a_{N_{\max},5,m} \end{bmatrix} \tag{5}$$

onde

$$a_{i,j,m} = \begin{cases} 1 & \text{para } j = 1 \ \& \ s_{i,m} = \mathrm{A} \\ 1 & \text{para } j = 2 \ \& \ s_{i,m} = \mathrm{C} \\ 1 & \text{para } j = 3 \ \& \ s_{i,m} = \mathrm{G} \\ 1 & \text{para } j = 4 \ \& \ s_{i,m} = \mathrm{T} \\ 0 & \text{para } \forall j \ \& \ s_{i,m} \notin S \end{cases} \tag{6}$$

e $N_{\max}$ é tamanho da maior sequência entre todas as M amostras de sequencias virais, ou seja, $N_{\max} = \max\{N_1, \ldots, N_M\}$. Desta forma, a matriz de característica possui a mesma dimensão ($N_{\max} \times 5$) para todas as M amostras de sequencias virais. Caso o tamanho da m-ésima sequência seja menor que a sequência máxima ($N_m < N_{\max}$), são inseridos $N_{\max} - N_m$ zeros (*zero padding*).

Antes de entrar na CNN, a matriz de característica de cada m-ésima amostra, $\mathbf{A}_m$, é transformada em uma matriz de dimensão $N_{\max} \times 1 \times 5$, expressa como

$$\mathbf{B}_m = \begin{bmatrix} \mathbf{b}_{1,m} & \cdots & \mathbf{b}_{5,m} \end{bmatrix} \tag{7}$$

onde

$$\mathbf{b}_{j,m} = \begin{bmatrix} b_{1,1,j,m} \\ \vdots \\ b_{N_{\max},1,j,m} \end{bmatrix} \tag{8}$$

no qual $b_{i,1,j,m} = a_{i,j,m}$. Esta transformação permite que a CNN processe cada m-ésima sequencia como uma entrada formada por 5 canais de vetores de dimensão $(N_{\max} \times 1)$, $\mathbf{b}_{j,m}$.

V. Resultados e Análises

Os resultados da classificação foram aferidos por meio das métricas de sensibilidade, especificidade, precisão, acurácia e F1-Score exibidos na Tabela . As métricas de desempenho para a validação cruzada k-fold corresponde a média entre todos os valores obtidos em cada fold. O modelo, resulta de valores de desempenho máximo para os dados de treinamento e validação como observado na Tabela IV.

Tabela IV
Resultado das métricas de desempenho para a classificação de ser ou não SARS-Cov-2 a partir da arquitetura proposta neste trabalho.

Métricas	Desempenho
Sensibilidade	1
Especificidade	1
Precisão	1
Acurácia	1
F1-score	1

A Figura 6, apresenta o resultado da classificação média (ser ou não SARS-Cov-2), e constata que para todos os subconjuntos, todas as sequências foram agrupadas corretamente de acordo com sua respectiva classe. A curva ROC para este problema é apresentada na Figura 7 e apresenta o valor de sensibilidade e especificidade igual a 1 de acordo com IV.

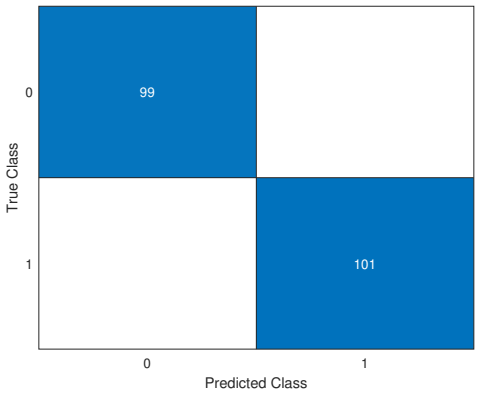

Figura 6. Matriz de confusão da abordagem proposta para o problema de classificação de ser SARS-CoV-2 e Não SARS-Cov-2.

Para testar o desempenho do modelo, foram utilizadas amostras ainda não visualizadas pela rede. Assim, das 16.891 amostras restantes do conjunto de dados, 12.000 amostras foram escolhidas aleatoriamente para fazer parte do conjunto

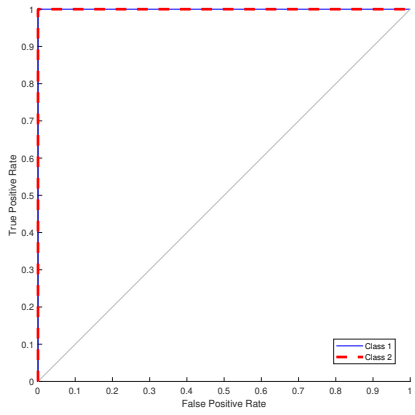

Figura 7. Curva ROC para o problema de classificação de ser SARS-CoV-2 e não SARS-Cov-2.

de teste. Das 12.000 amostras utilizadas, 11.996 foram classificadas de acordo com seu grupo (SARS-Cov-2), e apenas quatro amostras foram classificadas como sendo falso negativo, alcançando 99,99% de sensibilidade, conforme observado na sua matriz de confusão, Figura 8.

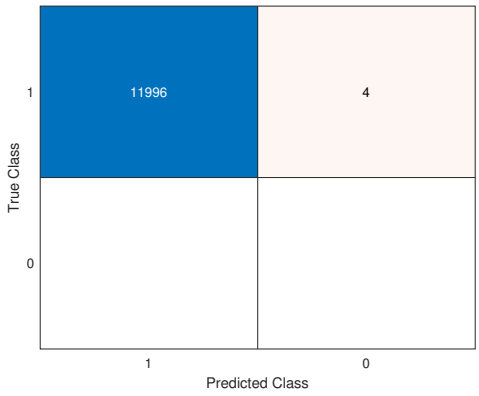

Figura 8. Matriz de confusão da abordagem proposta para o problema de classificação de ser SARS-CoV-2 e não SARS-Cov-2 para 12.000 amostras.

Com intuito de investigar a sensibilidade e robustez da arquitetura proposta, a possíveis mutações que o vírus SARS-CoV-2 viria sofrer, foi aplicado a metodologia de mutações artificiais discutida na Seção III-C. Para esse teste, foram utilizadas 10.000 amostras divididas aleatoriamente em dois grupos cada um contendo 5.000. Dessa forma, o conjunto final utilizado neste procedimento é formado por 5.000 amostras do vírus SARS-Cov-2 que sofreram mutações artificiais e 5.000 amostras genômicas, também do SARS-Cov-2, que não sofreram nenhuma alteração. Vale ressaltar que para este experimento todas as sequências foram previamente rotuladas como SARS-Cov-2 (rótulo 1). É possível visualizar o resultado da classificação para este conjunto de dados, por meio de sua matriz de confusão apresentada na Figura 9.

A matriz de confusão gerada a partir dos resultados da técnica de mutação artificial, discutidas na Seção III-C, con-

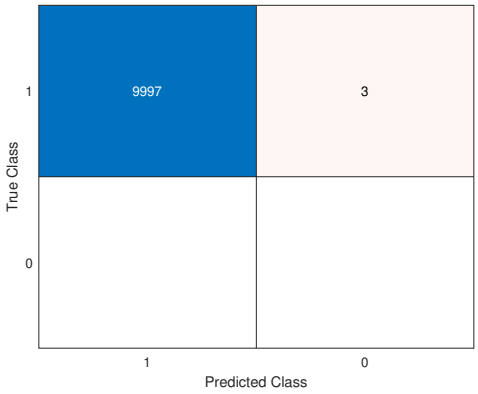

Figura 9. Matriz de confusão da abordagem proposta para o problema de classificação de ser SARS-CoV-2 e não SARS-Cov-2 para 10.000 amostras com 5.000 amostras apresentando mutações artificiais em suas sequencias em suas sequências.

seguiu classificar corretamente 9.973 amostras como sendo SARS-COV-2 e apenas 3 amostras como sendo falso positivas. Mesmo aplicando modificações nas sequências, o modelo é bastante sensível a possíveis mutações que as sequências possam sofrer alcançando um valor de sensibilidade de 99,7%.

Outro experimento foi feito para esse mesmo conjunto de dados no qual, as amostras que passaram por modificações foram previamente rotuladas como sendo Não SARS (rótulo 0) propositalmente. Para esse cenário o modelo caracterizou apenas duas amostras como falso negativo e 4.998 amostras como falso positivas, mesmo rotulando-as como Não SARS, o modelo consegue identificar a qual grupo as sequências realmente pertencem. A Figura 10 mostra a matriz de classificação obtida para este conjunto de dados.

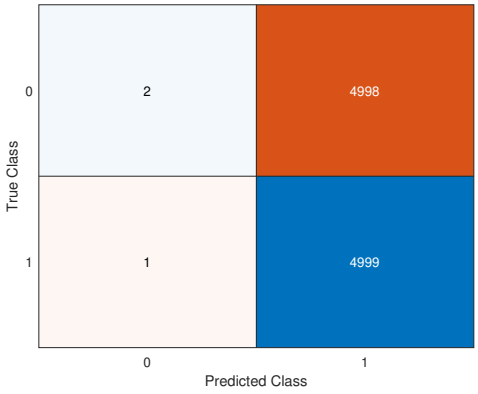

Figura 10. Matriz de confusão da abordagem proposta para o problema de classificação de ser SARS-CoV-2 e não SARS-Cov-2 para 10.000 amostras com 5.000 amostras apresentando modificações em suas sequências e rotuladas como não SARS.

VI. Comparação com estado da arte

As Tabelas V e IV sintetizam um conjunto de abordagens e resultados, descritos na Seção II, que utilizaram CNNs para identificar genomas virais. Características como: quantidade de camadas e o tamanho das sequências genômicas, serão apresentadas na Tabela V.

Tabela V
Comparação entre arquiteturas proposta com trabalhos relacionados.

Referências	Codificação	Camadas	Tamanho da Sequência
Fabijańska e Grabowski [24]	ASCII	30	3.257-24.751 bp
Ren {et al.} [27]	*One-Hot Encoded*	6	150-3.000 bp
Tampuu{et al.} [25]	*One-Hot Encoded*	2 CNN's cada qual com 7 camadas	300 bp
Lopez-Rincon{et al.} [20]	Atribuiu valores de 0 a 1 aos canais	10	31.029
Arquitetura Proposta	*One-Hot Encoded*	26	31.029

A maior parte dos trabalhos apresentados na Tabela V fazem uso de sequências genômicas incompletas para o processamento de suas redes. Ao aplicar sequências mais longas, [24] e [25], tiveram uma redução considerável no desempenho de seus modelos, implicando no uso de uma rede mais robusta, [24], [25], ou até mesmo, na redução dos tamanhos das sequências genômicas de entrada de seus modelos, [27] [25], a fim de obter melhorias no desempenho de suas arquiteturas.

Com relação a [20], apesar de fazer uso das sequências genômicas completas e apresentar uma quantidade de camadas menor, o autor faz uso de um conjunto de dados pequeno para o treinamento e validação de seu modelo, podendo acarretar em problemas de generalização e consequentemente no desempenho de sua rede, ao apresentar novas amostras. A Tabela VI, compara os resultados do desempenho da arquitetura proposta com os resultados disponíveis dos modelos da Tabela V.

Embora apresente uma arquitetura com muitas camadas, observou-se a variação dos valores de desempenho da arquitetura VGDC á medida que aumentava o tamanho das sequências genômicas utilizadas na rede. Embora utilize dois ramos convolucionais, a ferramenta ViraMiner alcançou 92,3% e 32% dos valores de sensibilidade e precisão, mesmo fazendo uso de sequências relativamente curtas.

A arquitetura DeepVirFinder forneceu apenas os valores da AUROC obtidos no seu modelo, atingindo o valor máximo de 96,68% para amostras de comprimento 3000 bp. Apesar de ter obtido o valor de sensibilidade de 100% e precisão de 98%. O trabalho apresentado por [20] obteve o valor da AUROC de 92%. Os resultados obtidos no modelo proposto é superior para todas as arquiteturas e métricas de desempenho apresentadas na Tabela VI indicando o alto desempenho e robusteza do modelo.

VII. Conclusão

A classificação molecular de vírus utilizando as Redes Neurais Convolucionais, vem se mostrando bastante promissora nos últimos anos. O SARS-CoV-2 é um vírus que apresenta uma alta taxa de transmissividade, ocasionando a infecção de milhares de pessoas no mundo inteiro. O artigo realiza uma discussão a respeito do novo coronavírus e técnicas que auxiliam na sua classificação. Nesse sentido, a pesquisa propõem o uso de uma CNN de múltiplas camadas, capaz de identificar a

Tabela VI
COMPARAÇÃO DAS MÉTRICAS DE DESEMPENHO DA ARQUITETURA PROPOSTA COM TRABALHOS RELACIONADOS.

Ref	Acurácia	Precisão	Sensibilidade	Especificidade	F1 Score	AUROC
Fabijańska e Grabowski [24]	0,99-1	0,83-1	0,84-1	0,99-1	0,83-1	
Ren *et al.* [27]	-	-	-	-	-	0,8635, 0,9210, 0,9496, 0,9668
Tampuu *et al.* [25]	0,90	0,90	0,32	-	-	0,923
Lopez-Rincon *et al.* [20]	0,985	0,98	1	0,9939	0,9797	0,92
Este trabalho	**1**	**1**	**1**	**1**	**1**	**1**

presença do SARS-CoV-2, a partir de amostras genômicas de DNAc do vírus, extraídas do 2019nCoVR, alcançando valores máximos em todas as métricas de avaliação. Os resultados obtidos, mostram-se mais confiáveis se comparados com os trabalhos discutidos no estado da arte, indicando que a ferramenta pode ser aplicada para a classificação de vírus da família *Coronaridae*.

AGRADECIMENTOS

Os autores agradecem à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo suporte e financiamento.

REFERÊNCIAS

[1] H. Wang, X. Li, T. Li, S. Zhang, L. Wang, X. Wu, and J. Liu, "The genetic sequence, origin, and diagnosis of sars-cov-2," *European Journal of Clinical Microbiology & Infectious Diseases*, pp. 1–7, 2020.

[2] H. S. Maghdid, K. Z. Ghafoor, A. S. Sadiq, K. Curran, and K. Rabie, "A novel ai-enabled framework to diagnose coronavirus covid 19 using smartphone embedded sensors: Design study," *arXiv preprint arXiv:2003.07434*, 2020.

[3] M. E. H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. A. Emadi, M. B. I. Reaz, and M. T. Islam, "Can ai help in screening viral and covid-19 pneumonia?" *IEEE Access*, vol. 8, pp. 132 665–132 676, 2020.

[4] Y. Toyoshima, K. Nemoto, S. Matsumoto, Y. Nakamura, and K. Kiyotani, "Sars-cov-2 genomic variations associated with mortality rate of covid-19," *Journal of human genetics*, vol. 65, no. 12, pp. 1075–1082, 2020.

[5] M. A. Remita, A. Halioui, B. Daigle, G. Kiani, A. B. Diallo *et al.*, "A machine learning approach for viral genome classification," *BMC bioinformatics*, vol. 18, no. 1, pp. 1–11, 2017.

[6] D. Lebatteux, A. M. Remita, and A. B. Diallo, "Toward an alignment-free method for feature extraction and accurate classification of viral sequences," *Journal of Computational Biology*, vol. 26, no. 6, pp. 519–535, 2019.

[7] A. Zielezinski, S. Vinga, J. Almeida, and W. M. Karlowski, "Alignment-free sequence comparison: benefits, applications, and tools," *Genome biology*, vol. 18, no. 1, pp. 1–17, 2017.

[8] S. Nooij, D. Schmitz, H. Vennema, A. Kroneman, and M. P. Koopmans, "Overview of virus metagenomic classification methods and their biological applications," *Frontiers in microbiology*, vol. 9, p. 749, 2018.

[9] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped blast and psi-blast: a new generation of protein database search programs," *Nucleic acids research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[10] Å. J. Vågene, A. Herbig, M. G. Campana, N. M. R. García, C. Warinner, S. Sabin, M. A. Spyrou, A. A. Valtueña, D. Huson, N. Tuross *et al.*, "Salmonella enterica genomes from victims of a major sixteenth-century epidemic in mexico," *Nature ecology & evolution*, vol. 2, no. 3, pp. 520–528, 2018.

[11] D. J. Lipman and W. R. Pearson, "Rapid and sensitive protein similarity searches," *Science*, vol. 227, no. 4693, pp. 1435–1441, 1985.

[12] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic acids research*, vol. 22, no. 22, pp. 4673–4680, 1994.

[13] R. C. Edgar, "Search and clustering orders of magnitude faster than blast," *Bioinformatics*, vol. 26, no. 19, pp. 2460–2461, 2010.

[14] G. S. Randhawa, M. P. Soltysiak, H. El Roz, C. P. de Souza, K. A. Hill, and L. Kari, "Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: Covid-19 case study," *Plos one*, vol. 15, no. 4, p. e0232391, 2020.

[15] G. S. Randhawa, K. A. Hill, and L. Kari, "Ml-dsp: Machine learning with digital signal processing for ultrafast, accurate, and scalable genome classification at all taxonomic levels," *BMC genomics*, vol. 20, no. 1, p. 267, 2019.

[16] J. Ren, N. A. Ahlgren, Y. Y. Lu, J. A. Fuhrman, and F. Sun, "Virfinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data," *Microbiome*, vol. 5, no. 1, pp. 1–20, 2017.

[17] T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," *Future healthcare journal*, vol. 6, no. 2, p. 94, 2019.

[18] M. S. Mottaqi, F. Mohammadipanah, and H. Sajedi, "Contribution of machine learning approaches in response to sars-cov-2 infection," *Informatics in Medicine Unlocked*, p. 100526, 2021.

[19] Y. Park and M. Kellis, "Deep learning for regulatory genomics," *Nature biotechnology*, vol. 33, no. 8, pp. 825–826, 2015.

[20] A. Lopez-Rincon, A. Tonda, L. Mendoza-Maldonado, E. Claassen, J. Garssen, and A. D. Kraneveld, "Accurate identification of sars-cov-2 from viral genome sequences using deep learning," *bioRxiv*, 2020.

[21] S. Lalmuanawma, J. Hussain, and L. Chhakchhuak, "Applications of machine learning and artificial intelligence for covid-19 (sars-cov-2) pandemic: A review," *Chaos, Solitons & Fractals*, p. 110059, 2020.

[22] G. Eraslan, Ž. Avsec, J. Gagneur, and F. J. Theis, "Deep learning: new computational modelling techniques for genomics," *Nature Reviews Genetics*, vol. 20, no. 7, pp. 389–403, 2019.

[23] J. Zou, M. Huss, A. Abid, P. Mohammadi, A. Torkamani, and A. Telenti, "A primer on deep learning in genomics," *Nature genetics*, vol. 51, no. 1, pp. 12–18, 2019.

[24] A. Fabijańska and S. Grabowski, "Viral genome deep classifier," *IEEE Access*, vol. 7, pp. 81 297–81 307, 2019.

[25] A. Tampuu, Z. Bzhalava, J. Dillner, and R. Vicente, "Viraminer: Deep learning on raw dna sequences for identifying viral genomes in human samples," *PLOS ONE*, vol. 14, p. e0222271, 09 2019.

[26] A. Lopez-Rincon, A. Tonda, M. Elati, O. Schwander, B. Piwowarski, and P. Gallinari, "Evolutionary optimization of convolutional neural networks for cancer mirna biomarkers classification," *Applied Soft Computing*, vol. 65, pp. 91–100, 2018.

[27] J. Ren, K. Song, C. Deng, N. A. Ahlgren, J. A. Fuhrman, Y. Li, X. Xie, R. Poplin, and F. Sun, "Identifying viruses from metagenomic data using deep learning," *Quantitative Biology*, pp. 1–14, 2020.