Nova Proposta de Representação de Genoma Viral Aplicada na Classificação do SARS-CoV-2 com Aprendizagem Profunda

Luísa C. de Souza*, 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: *luisa.souza.103@ufrn.edu.br,†m.g.barbosafernandes@gmail.com,‡mfernandes@dca.ufrn.br

***Abstract*—Em dezembro de 2019, o primeiro caso de COVID-19 foi descrito em Wuhan, na China, e em abril de 2021, já haviam 136 milhões de casos confirmados. Devido a rápida propagação do vírus, esforços vêm sendo realizados pela comunidade científica para o desenvolvimento de técnicas de classificação viral do SARS-CoV-2. Neste trabalho, foi desenvolvido, utilizando um conjunto de técnicas de Processamento de Sinais Genômicos, uma nova proposta de representação dos dados genéticos de amostras de seis vírus da família Coronaviridae, a qual pertence o vírus SARS-CoV-2. Em seguida, o mapeamento realizado foi empregado a uma arquitetura de aprendizagem profunda para a classificação das amostras virais, obtendo acurácia de 94% e 91% para os tamanhos 64 e 128 da redimensão das sequências, respectivamente, além de obter 100% de sensibilidade para os vetores com tamanho 64.**

***Index Terms*—COVID-19, SARS-CoV-2, Representação genômica, Processamento de sinais genômicos, Aprendizagem profunda.**

I. Introdução

A Organização Mundial de Saúde (OMS) declarou, em 30 de janeiro de 2020 que o surto do COVID-19, doença causada pelo vírus SARS-CoV-2, constituía uma Emergência de Saúde Pública de Interesse Internacional, dada a rápida propagação do vírus, tal que após duas semanas do primeiro caso diagnosticado, outros 1000 pacientes testaram positivo para Coronavírus [1]–[3]. Em abril de 2021, o número total de casos registrados da doença ultrapassou a marca de 136 milhões, com 2,9 milhões de mortes causadas pelo vírus [3].

Devido o alto teor de propagação da doença, é de vital importância o diagnóstico de pacientes infectados para que estes sejam adequadamente tratados e isolados, na tentativa de evitar o contágio de outros indivíduos. Em um experimento realizado por Yang [4] com 213 pacientes infectados pelo Coronavírus, apresentando quadros críticos e moderados da doença, 866 amostras de tratos respiratórios dos pacientes, obtidos por zaragatoa nasofaríngea, expectoração e fluido de lavagem bronco alveolar foram analisados através da técnica de reação em cadeia da polimerase via transcriptase reversa quantitativa (qRT-PCR). Em casos de 0 a 7 dias após o início da doença, os exames das amostras de zaragatoa faríngea obtiveram uma taxa negativa de 40% para casos críticos, e $38{,}7\%$ para casos moderados, com zaragatoa nasal apresentando $26{,}7\%$ e 27%, e por fim, com expectoração obtiveram 11% e $17{,}8\%$, para casos críticos e moderados, respectivamente. Acredita-se que esse resultado falso negativo ocorre devido a mutação típica de vírus de RNA, onde o SARS-CoV-2 tem a taxa de evolução média de aproximadamente 10-4 nucleotídeos substituídos por ano [5]. Outra dificuldade encontrada por [6] e [7], em seus trabalhos de classificação do vírus COVID-19, é a presença de resultados falsos positivos gerados devido a existência de amostras da mesma família presentes na análise realizada nos estudos, onde em função das relações de semelhança genética entre os organismos virais, exemplos de outras espécies são classificados como SARS-CoV-2 [8].

Dessa forma, é fundamental o aprimoramento de técnicas de análise e classificação de genoma viral para auxílio do diagnóstico eficiente. Na bioinformática, a análise de sequências biológicas é realizada através de dois métodos principais. O primeiro método trata-se de técnicas que utilizam alinhamento de sequências, como o BLAST e o FASTA [8], tais algoritmos procuram por correspondentes de bases ou grupos de bases na mesma ordem em duas ou mais sequências. As desvantagens de tais métodos são o alto custo de memória e tempo requeridos [9] além de assumir que as sequências de DNAc (DNA complementar) são linearmente arranjadas, o que não é o caso para sequências virais. O segundo método engloba as técnicas nas quais não é realizado o alinhamento de sequências (*free-alignment*) [10]. Tais técnicas incluem o uso de aprendizagem profunda para classificação de sequências virais. Essa classificação ocorre em duas etapas, a primeira pode ser caracterizada como um mapeamento das sequências biológicas em um espaço de característica, e a segunda etapa consiste no processamento dos dados por uma técnica de aprendizagem de máquina [11], [12].

O DNA contém informações genéticas em suas moléculas que sistematizam o desenvolvimento e funcionamento de organismos vivos e vírus. As técnicas de mapeamento ou representação de sequências de DNA, ou DNAc, transformam os nucleotídeos em informação numérica [13]. As representações numéricas de sequências genéticas podem ser divididas em três categorias, de mapeamento de valor único,

no qual cada nucleotídeo será associado a um valor único no espaço unidimensional, mapeamento de sequências multidimensional, onde cada base nitrogenada será substituída por um vetor contendo um ponto no espaço multidimensional, e por fim, o mapeamento cumulativo, onde um modelo de passeio aleatório acumulará a contribuição de valores consecutivos associados aos nucleotídeos para formar uma curva [14].

O Processamento de Sinais Genômicos (PSG) se baseia na utilização da teoria, algoritmos e métodos matemáticos do processamento digital de sinais para análise, processamento e uso de dados genômicos [11], [13]–[15]. Propriedades de periodicidade e distribuição ocultas podem ser identificadas por técnicas de PSG. Logo, o uso dessas ferramentas em conjunto com as representações numéricas de sequências de DNA podem oferecer mais informações sobre o perfil genético de organismos, em comparação com métodos de representações convencionais [13]. O proposta apresentada em [14] utilizou técnicas de PSG para conversão de sequências de nucleotídeos para uma representação gráfica afim de ser empregada na classificação de três tipos de genoma funcional realizada por uma arquitetura de aprendizagem profunda. O trabalho proposto em [16], em sua pesquisa, desenvolveu uma nova forma de mapeamento numérico de sequências de DNA utilizando uma representação multidimensional em conjunto com a transformada discreta de Fourier-*Discrete Fourier Transform* (DFT), uma das mais importantes ferramentas da PSG. No trabalho apresentado em [7], foram utilizadas técnicas de PSG para seleção de características, em conjunto com métodos de máquina de aprendizado, para desenvolver um sistema automático de classificação do SARS-Cov-2, SARS-Cov e MERS-CoV.

A utilização de aprendizagem de máquina baseada em redes neurais profundas tem apresentado resultados bastante significativos na área de classificação viral. A técnica proposta em [12] utiliza uma rede neural convolucional -*Convolutional Neural Network* (CNN) profunda para realizar a classificação viral, aplicando o método em vírus da dengue, HIV-1, influenza A, hepatite B e C, e dependendo do tipo viral e do número de subtipos associados, obteve um F1-score de 0.85 a 1.0. Por sua vez, o trabalho apresentado em [17] fez uso de uma rede neural convolucional para classificação de sequências de DNA considerando-as como texto, testou o método em 12 datasets, e para os exemplos de fácil classificação obteve excelentes resultados. Na pesquisa realizada em [18], foi desenvolvido o ViraMiner, um método de identificação viral que contém dois ramos de CNNs projetada para detectar padrões de frequência em contigs metagenômicos, para contigs com 300 bases par (bp), o método atingiu uma área de 0,923 sob a curva -*Receiver operating characteristic* (ROC).

Perante o exposto, o presente trabalho tem como objetivo o desenvolvimento de uma nova estratégia de representação de sequências de DNAc viral, tal como a do SARS-CoV-2, utilizando um conjunto de técnicas de processamento de sinais genômicos, tais como a *Chaos Game Representation* (CGR) e a *Discrete Fourier Transform*, para ser empregada em métodos de aprendizagem profunda para classificação viral. Tal representação de sequências genéticas gera uma nova assinatura viral contendo as informações em um novo espaço de características, e apresentam comprimento consideravelmente menor que a sequência genômica original

II. Dataset

Para este estudo, foi realizado o download de 12467 amostras de sequências de genoma viral disponibilizadas por 67 países através do banco de dados National Genomics Data Center (NGDC). As amostras do dataset são de seis espécies, Severe acute respiratory syndrome-related coronavirus (SARS-CoV-2), Betacoronavirus 1, Middle East respiratory syndrome-related coronavirus (MERS-CoV), Human coronavirus NL63 (HCoV NL63), Human coronavirus 229E (HCoV 229E) e Human coronavirus HKU1 (HCoV HKU1). Pertencentes à família Coronaviridae, do reino Riboviria, dispõem de um comprimento de genoma variando de 26000 a 32000 pares de base (*base-pair* - bp). As sequências formadas por bases de nucleotídeos se apresentam na forma de vetores de caracteres, onde cada letra representa um nucleotídeo específico, guanina (G), adenina (A), timina (T), e citosina (C). A Tabela I exibe um resumo dos dados das amostras utilizadas no trabalho.

TABLE I: Amostras das Sequências Virais

Espécie viral	Informações das sequências		
	Nº de seq.	Comp. seq. mín. (N)	Comp. seq. máx. (N)
SARS-Cov-2	11969	26973	30018
Betacoronarivus 1	140	30536	31029
MERS-CoV	258	29267	30150
HCoV NL63	55	27302	27832
HCoV 229E	27	26592	27307
HCoV HKU1	18	29367	29983

Considerando que no método desenvolvido as assinaturas virais foram classificadas em duas classes, para que haja o balanceamento dos dados, foram criados dois grupos de 400 amostras escolhidas aleatoriamente do SARS-CoV-2 e dos demais vírus pertencentes ao dataset, totalizando 800 amostras de sequências virais.

III. Proposta de representação

A Figura 1 apresenta a técnica de representação proposta, na qual uma sequência de DNAc de tamanho N expressa como

$$\mathbf{s} = [s_1, \ldots, s_i, \ldots, s_N] \tag{1}$$

onde cada i-ésimo elemento s_i representa um dos possíveis nucleotídeos da sequencia de DNAc, ou seja, $s_i \in \{\mathrm{A}, \mathrm{C}, \mathrm{T}, \mathrm{G}\}$. A proposta utiliza duas técnicas de processamento sinais genômicos em cascata objetivando a criação de uma assinatura única para cada i-ésima sequencia de DNAc. As técnicas de processamento são CGR e a DFT, que serão detalhadas nas próximas subseções [19], [20].

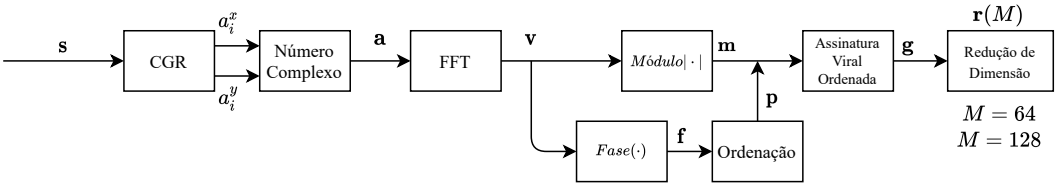


Fig. 1: Esquema da Representação de Sequências Biológicas.

A. *Chaos Game Representation (CGR)*

Proposto em [21], o CGR é uma metodologia capaz de fornecer representações numéricas e gráficas de sequências genéticas através de sistemas de funções iterativas (SFI) [16], [21]. O CGR mapeia a sequência de DNAc caracterizada pelo vetor $\mathbf{s}$ (ver Equação 1) em um espaço bidimensional através dos símbolos a_n^x e a_n^y expressos como

$$a_n^x = \frac{1}{2}s_n^x + \frac{1}{2}a_{n-1}^x, \text{ para } n = 1, \ldots, N \tag{2}$$

e

$$a_n^y = \frac{1}{2}s_n^y + \frac{1}{2}a_{n-1}^y, \text{ para } n = 1, \ldots, N \tag{3}$$

onde

$$s_n^x = \begin{cases} 1 & \text{caso } s_n = \text{A} \\ -1 & \text{caso } s_n = \text{T} \\ -1 & \text{caso } s_n = \text{C} \\ 1 & \text{caso } s_n = \text{G} \end{cases} \tag{4}$$

e

$$s_n^y = \begin{cases} 1 & \text{caso } s_n = \text{A} \\ 1 & \text{caso } s_n = \text{T} \\ -1 & \text{caso } s_n = \text{C} \\ -1 & \text{caso } s_n = \text{G} \end{cases}. \tag{5}$$

Na técnica proposta assume-se como condição inicial ($n = 0$), $a_0^x = 0$ e $a_0^y = 0$ [16], [19]. Assim, cada base associada a um s_n, representará um ponto no espaço bidimensional contendo as coordenadas a_n^x e a_n^y, e esses valores serão relacionados a um número complexo na forma $a_n^x + ja_n^y$, obtendo o vetor $\mathbf{a}$, expresso como

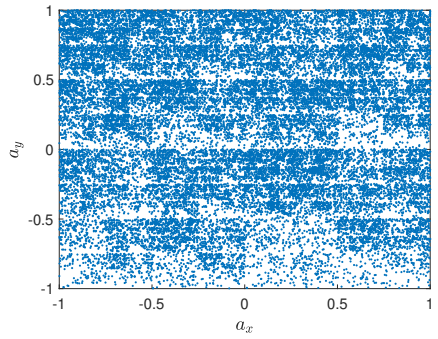
$$\mathbf{a} = [a_1^x + ja_1^y, a_2^x + ja_2^y, \ldots, a_N^x + ja_N^y]. \tag{6}$$

A Figura 2 ilustra dois exemplos de vírus da família Coronaviridae mapeados com CGR, nos quais observa-se que cada vírus possui uma assinatura distinta.

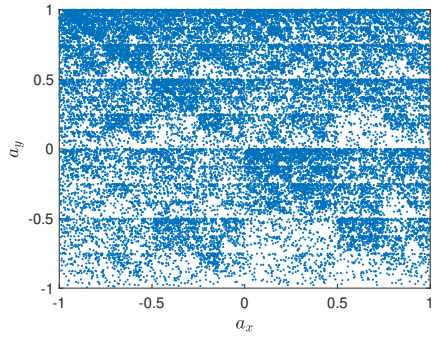
Como apresentado na Figura 1, na próxima etapa da proposta de representação, o vetor $\mathbf{a}$ será empregado na DFT.

B. *DFT e Ordenação de Vetores*

Com base nos trabalhos apresentados em [16], [20], esta proposta faz a utilização da DFT, objetivando gerar uma assinatura no domínio da frequência do sinal genômica, dado que a partir da análise do espectro provido, periodicidades e informações latentes das sequências de nucleotídeos podem ser observadas mais facilmente do que em análises no domínio do tempo [20], [22].



(a) Virus SARS-CoV-2 (GU553363).



(b) Vírus Betacoronavirus-1 (KX538977).

Fig. 2: Exemplo de representação viral utilizando CGR.

Como ilustrado na Figura 1, o vetor de números complexos $\mathbf{a}$ de comprimento N passa por uma DFT gerando o vetor $\mathbf{v}$, que pode ser expresso como

$$\mathbf{v} = [v_1, v_2, \ldots, v_N] \tag{7}$$

onde cada i-ésimo elemento v_i pode ser expresso como

$$v_i = \sum_{n=0}^{N-1} v_n e^{-j\frac{2\pi}{N}in}. \tag{8}$$

Após o cálculo da DFT, em razão de seus dados se apresentarem na forma complexa, é necessário decompor os componentes de módulo e fase do vetor $\mathbf{v}$ gerando os vetores $\mathbf{m}$ e $\mathbf{f}$, respectivamente [23]. O vetor $\mathbf{m}$ pode ser expresso como

$$\mathbf{m} = [m_1, m_2, \ldots, m_N] \tag{9}$$

onde cada i-ésimo elemento m_i é a amplitude em uma determinada frequência e pode ser expresso como

$$m_i = |v_i| \tag{10}$$

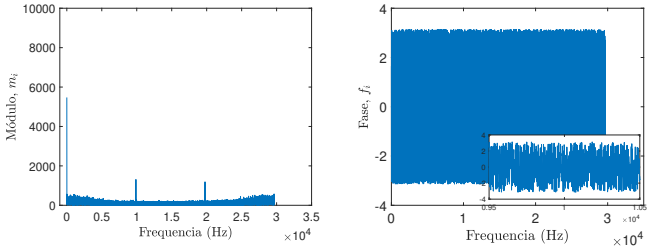
A fase da DFT, representada pelo vetor $\mathbf{f}$ apresenta-se como

$$\mathbf{f} = [f_1, f_2, \ldots, f_N] \tag{11}$$

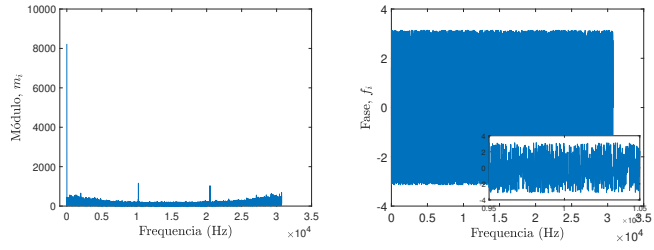
onde cada i-ésimo elemento f_i é a fase da transformada distribuída de $-\pi$ a π sendo expresso como

$$f_i = \angle v_i. \tag{12}$$

A Figura 3 exibe a DFT de duas amostras virais obtidas da CGR como apresentadas anteriormente na Figura 2, onde a primeira imagem de 3a e 3b exibe $\mathbf{m}$, ou seja, o módulo da transformada, e a segunda imagem exibe sua fase $\mathbf{f}$.



(a) Virus SARS-CoV-2 (GU553363).



(b) Vírus Betacoronavirus-1 (KX538977).

Fig. 3: Resposta em módulo e fase da DFT de amostras virais.

Como observado na Figura 3, resposta em módulo das assinaturas, vírus semelhantes apresentam valores máximos de frequência similares, porém em fases diferentes. Portanto, é realizada uma ordenação crescente do vetor $\mathbf{f}$, gerando um vetor de posições da ordenação $\mathbf{p}$, representado como

$$\mathbf{p} = [p_1, p_2, \ldots, p_N] \tag{13}$$

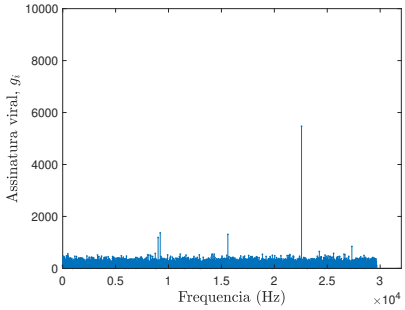
e essas posições são utilizadas para ordenar o vetor módulo $\mathbf{m}$, gerando um novo vetor $\mathbf{g}$, expresso como

$$\mathbf{g} = [g_1, g_2, \ldots, g_N] \tag{14}$$

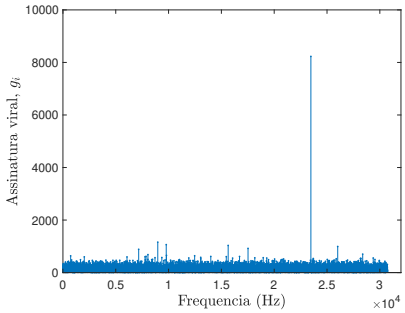
onde cada i-ésimo elemento g_i será o valor de amplitude ordenado de acordo com a posição de sua fase, como visto

$$g_i = m_{p_i}. \tag{15}$$

A partir da ordenação dos vetores, foi obtido um novo vetor $\mathbf{g}$ com a mesma função de módulo do original, porém com posições diferente, relativas a função de suas fases, aumentando assim a diferenciação entre assinaturas de vírus próximos como é exibido na Figura 4, que mostra a nova assinatura viral ordenada das amostras apresentadas na Figura 3.



(a) Virus SARS-CoV-2 (GU553363).



(b) Vírus Betacoronavirus-1 (KX538977).

Fig. 4: Assinaturas virais ordenadas.

Os trabalhos [16], [24]–[26] apresentam estratégias semelhantes a do presente estudo, aplicando a CGR e em seguida calculando a DFT das sequências genéticas, porém, as propostas não fazem uso da ordenada da fase da DFT. As técnicas desenvolvidas por [16], [24] utilizam o espectro de potência da transformada, enquanto [25] optou por utilizar apenas o espectro de amplitude em conjunto com o coeficiente de correlação de Pearson. Na proposta apresentada em [26], foram calculados os valores médios da DFT suavizada. Assim, é importante destacar que a não utilização da informação da fase pode desconsiderar a localização dos valores de máximo local de frequência, focando apenas do valor de sua amplitude. Na Figura 3, é possível observar que as duas amostras virais possuem máximos de frequência em torno dos valores 0 Hz, 1×10^4 Hz e 2×10^4 Hz , e que em torno de 1×10^4 Hz e 2×10^4 Hz, a amplitude para as duas amostras é semelhante, porém, observando a fase em torno do valor de frequência 1×10^4Hz, como exposto na amplificação da imagem no quadrante a direita da Figura 3, as duas amostras apresentam perfis de fase diferentes. Como observado na Figura 4, após a ordenação, os maiores valores de frequência não se encontram mais nas mesmas posições.

C. Redução de Dimensionalidade

Dado que, o vetores da assinatura viral ordenada, $\mathbf{g}$, possuem comprimentos distintos, como observado na Tabela I

que apresenta os valores de mínimo e máximo para N, e que devido o uso da DFT, a quantidade de informação relevante está associada a um número pequeno de valores máximos de frequência [22], foi realizada uma redução na dimensão dos dados até os vetores apresentam os comprimentos 64 e 128 por assinatura. Estes valores de comprimento foram escolhidos após experimentação com diversos tamanhos, pois na classificação realizada pela CNN, apresentaram melhores resultados na caracterização dos dados genéticos.

Para isto, foram selecionados os M maiores valores de $\mathbf{g}$, onde M assume 64 ou 128, gerando o vetor $\mathbf{b}$ e suas posições no vetor original, que formam o vetor $\mathbf{o}$, apresentados como

$$\mathbf{b} = [b_1, b_2, \ldots, b_M] \tag{16}$$

e

$$\mathbf{o} = [o_1, o_2, \ldots, o_M]\,. \tag{17}$$

O vetor de posições $\mathbf{o}$ foi então ordenado de forma crescente e semelhante a ordenação da transformada realizada na seção anterior no vetor $\mathbf{m}$, as novas posições foram empregadas nos maiores valores de módulo apresentadas no vetor $\mathbf{b}$, obtendo o vetor com dimensão reduzida $\mathbf{r}$ com tamanho M, expresso como

$$\mathbf{r} = [r_1, r_2, \ldots, r_M] \tag{18}$$

onde cada elemento r_i foi dado por

$$r_i = b_{o_i(ordenado)}. \tag{19}$$

Dessa forma, cada ponto de $\mathbf{r}$ estará em sua posição relativa aos outros valores de máximo da sequência original $\mathbf{g}$. A Figura 5 expõe o resultado da compressão de duas amostras virais para todos os tamanhos de M.

Após a redimensão dos vetores das assinaturas virais, a técnica de representação das sequências de DNAc está finalizada, com tal representação podendo então ser analisada por técnicas de aprendizagem profunda.

IV. Arquitetura da Rede Neural Profunda

Seguindo as propostas da literatura [12], [27], [28], esse trabalho empregou técnicas de processamento de sinais genômicos para representação das sequências genéticas em conjunto com uma Rede Neural Convolucional para realizar a classificação destas em duas classes, SARS-CoV-2 ou outras espécies.

A arquitetura da Rede Neural Profunda utilizada se trata de um modelo unidimensional de rede convolucional, onde o comprimento das assinaturas virais influenciou na escolha de alguns parâmetros, tais como o tamanho do *Input*, a quantidade de camadas e o tamanho dos filtros T_n com $n = 1, 2, 3, 4$, onde o maior valor que pôde ser empregado foi igual a 4. Os classificadores forneceram saídas discretas, caracterizadas pelos valores 1 e 0. Para alcançar a arquitetura final da CNN, foram realizadas várias execuções de treinamento e validação, e a que apresentou os melhores resultados de performance é a exposta na Figura 6, e descrita na Tabela II.

TABLE II: Arquitetura da Rede Neural Convolucional detalhada.

Camada	Descrição	Valores
	Input	
1	$(M \times 1 \times 1)$	$M = 64$ ou 128
2	Conv1D	$T_1 = 4$ e $Q_1 = 16$
3	BachNorm	–
4	ReLu	–
5	MaxPool1D	$S_1 = 2$
6	Conv1D	$T_2 = 4$ e $Q_2 = 8$
7	BachNorm	–
8	ReLu	–
9	MaxPool1D	$S_2 = 2$
10	Conv1D	$T_3 = 2$ e $Q_3 = 2$
11	BachNorm	–
12	ReLu	–
13	MaxPool1D	$S_3 = 2$
14	Conv1D	$T_4 = 2$ e $Q_4 = 2$
15	BachNorm	–
16	ReLu	–
17	MaxPool1D	$S_4 = 2$
18	FC1	$P_1 = 256$
19	Dropout	$\alpha_1 = 0.6$
20	FC2	$P_2 = 128$
21	Dropout	$\alpha_2 = 0.6$
22	FC3	$P_3 = 64$
23	Dropout	$\alpha_3 = 0.6$
24	FC4	$P_4 = 2$
25	SoftMax	2 classes

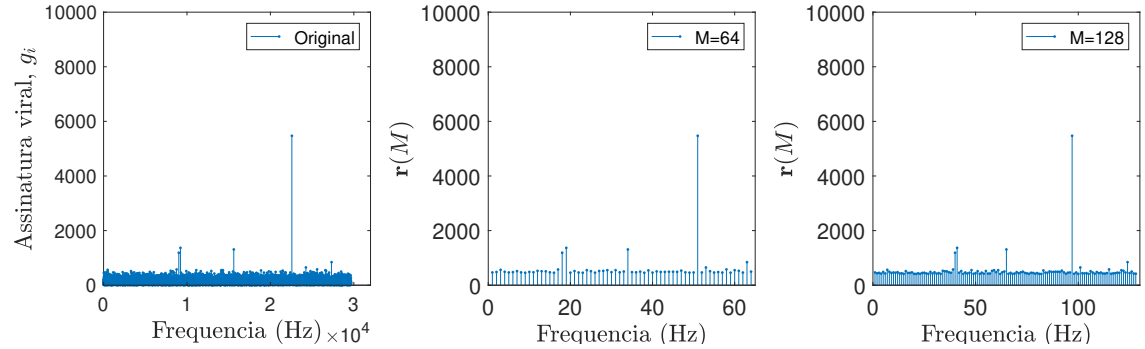
V. Resultados e Discussão

Os algoritmos deste trabalho foram implementados no Matlab 2020.a em um notebook com as configurações Intel Core i5-7200U com CPU 2.50 GHz e RAM de 8 GB. Para o treinamento da rede de aprendizado profundo, as 800 amostras foram divididas em dois conjuntos para cada tamanho de M, com 80% das amostras sendo usadas para o treinamento, e as 20% de amostras restantes, utilizadas para a validação. Para reunir o conjunto de teste, foram escolhidas no dataset aleatoriamente 100 novas amostras ainda não conhecidas pela rede. Dentro de cada conjunto de treinamento, validação e teste, 50% das amostras pertencem à classe SARS-CoV-2, e 50% representam outras espécies. A Tabela III detalha a quantidade de cada espécie presente em cada conjunto de dados.

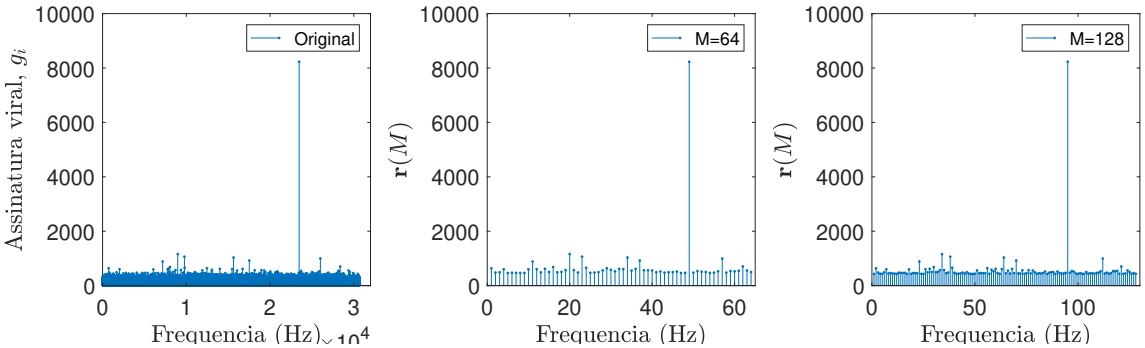
TABLE III: Dados das Sequências Virais Para os Conjuntos de Treinamento e Teste.

Espécie viral	Classe	Quantidade de sequências	
		Treinamento	Teste
SARS-Cov-2	1	400	50
Betacoronarivus 1	0	102	13
MERS-CoV	0	236	18
HCoV NL63	0	36	10
HCoV 229E	0	14	5
HCoV HKU1	0	12	4

A rede foi treinada durante 50 épocas e utilizou o otimizador RMSProp com taxa de aprendizagem de 0.001 para minimizar a função de perda (loss function), além disso, o tamanho de batch escolhido para o treinamento da rede foi igual a 512.



(a) Virus SARS-CoV-2 (GU553363).



(b) Vírus Betacoronavirus-1 (KX538977).

Fig. 5: Redução de dimensão de amostras virais para todos os M.

Após o treinamento para $M = 64$ e $M = 128$, respectivamente, a rede foi testada, obtendo os resultados expostos nas Figuras 7 e 8 que mostram as matrizes de confusão para os dois tamanhos de dimensão.

A partir das matrizes de confusão, é feita a analise do desempenho da rede de aprendizado profunda na classificação do vírus do COVID-19, realizada utilizando a representação com a fase ordenada. Na Tabela IV são expostas as medidas de performance acurácia, sensibilidade, especificidade, precisão e F1-score, para cada tamanho de sequência após a redimensão.

TABLE IV: Comparação de performance da classificação da Rede de Aprendizado para os tamanhos da redimensão $M = 64$ e $M = 128$.

M	Métricas de Performance				
	ACC	SEN	ESP	PRE	F1-Score
64	94%	100%	88%	89,3%	94,34%
128	91%	90%	92%	91,8%	90,90%

É possível observar que para a menor dimensão, foram obtidos os maiores valores de acurácia, F1-score e sensibilidade dado que todas as amostras do vírus SARS-CoV-2 foram corretamente detectadas. Contudo, a maior dimensão apresenta melhor especificidade pois a rede detectou menos falso negativos para este comprimento de vetor.

A curva ROC se trata de um gráfico que apresenta o desempenho de um classificador, sendo produzida plotando no eixo y a taxa de verdadeiro positivo, ou seja, a métrica de performance sensibilidade, e no eixo x a taxa de falso positivo que representa 1 - especificidade, para os valores de teste [29]. A partir dela obtemos a métrica de performance (*Area Under Curve*) (AUC) da curva ROC. Nas Figuras 9 e 10 são exibidas as curvas ROC para as duas dimensões de vetores, 64 e 128, respectivamente.

Essa medida irá avaliar os efeitos de treinamento do classificador para o dataset, de modo que quanto mais alto o valor de AUC, melhor o seu funcionamento [30]. Consequentemente, para os vetores de comprimento 64, a rede de aprendizado profundo apresentou uma melhor classificação dos dados virais com AUC de 0,9400. Porém, como ambas apresentaram AUC superior a 0,900, pode-se concluir que o método proposto tem a habilidade de representar as sequências de DNAc mesmo após diminuição significativa da dimensão. A tabela V expõe uma comparação dos resultados obtidos pela proposta do estudo e outros trabalhos previamente mencionados.

Enquanto todos os trabalhos expostos na tabela V executam a classificação de sequências de DNA, apenas os trabalhos [7], [12], [18] realizam classificação viral, enquanto [17] e [28] concentram suas pesquisas na classificação de proteínas. Pela comparação dos resultados obtidos, é possível afirmar que o método proposto apresenta desempenho semelhante ou superior a técnicas de representação consolidadas na literatura.

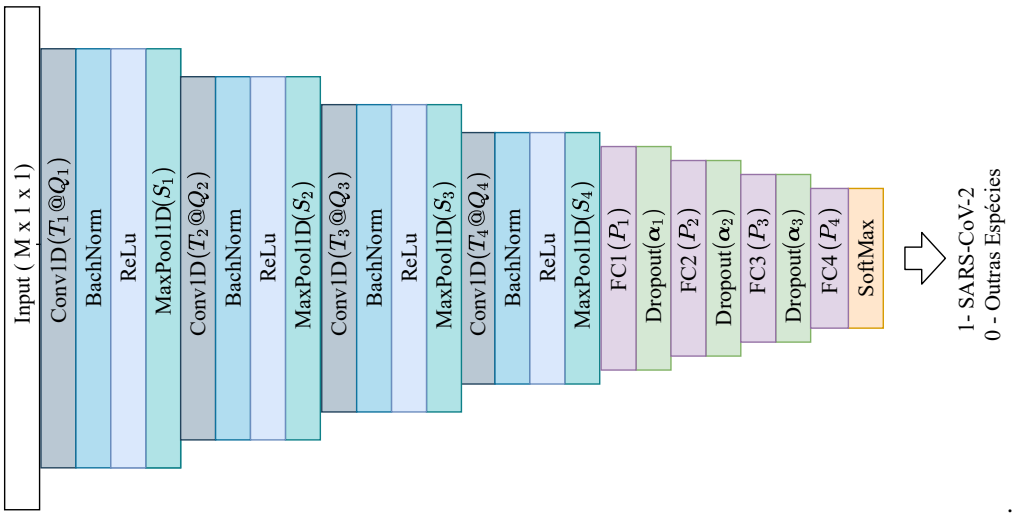


Fig. 6: Arquitetura da Rede neural Convolucional utilizada para classificação do SARS-CoV-2.

TABLE V: Comparação de performance do método proposto com outros trabalhos da literatura.

Referências	Representação	Métricas de Performance					
		ACC	SEN	ESP	PRE	F1-Score	AUC
[7]	EIIP [31]	98,33%	N/A	N/A	N/A	98,36%	N/A
[12]	Codificação ASCII [32]	97,7% - 100%	84,1% - 100%	98,1% - 100%	84,4% - 100%	83,5% - 100%	N/A
[17]	*One-hot Encoding* [33]	71,5% - 99,6%	N/A	N/A	N/A	N/A	N/A
[18]	*One-hot Encoding* [33]	N/A	N/A	N/A	N/A	N/A	0,9230
[28]	Voss [34]	83% -84%	N/A	N/A	N/A	84% -85%	N/A
Este trabalho	CGR e DFT ordenada	94%	100%	88%	89,3%	94,34%	0,9400

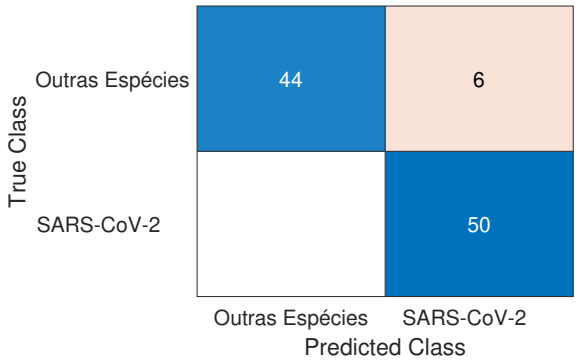


Fig. 7: Matriz de Confusão para o tamanho 64.

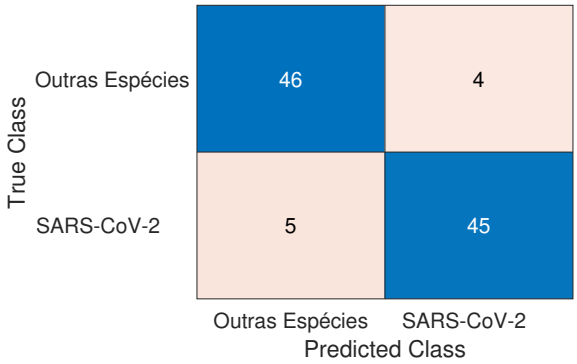


Fig. 8: Matriz de Confusão para o tamanho 128.

VI. Conclusão

Neste trabalho foi proposto uma nova representação de sequências de DNAc, baseada na utilização de técnicas de processamento de sinais genômicos, aplicada em sequências virais da família Coronaviridae, para a classificação do vírus COVID-19. Comparado com outros métodos que utilizam a transformada de Fourier no pré processamento das amostras de dados genéticos, o presente método utiliza a informação de fase em conjunto com a informação de amplitude dos sinais, para aumentar a diferenciação entre as amostras. Em adição a isto, a redução de dimensão dos vetores de assinatura viral possibilitam uma classificação viral com menos custo, sem perda de performance obtendo acurácia de 94% e 91%, e AUC de 0,9400% e 0,9100% para comprimento de vetor igual a 64 e 128, respectivamente, além da diminuição da necessidade de memória.

Agradecimentos

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

References

[1] A. Spinelli and G. Pellino, "Covid-19 pandemic: perspectives on an unfolding crisis," *Journal of British Surgery*, vol. 107, no. 7, pp. 785–787, 2020.

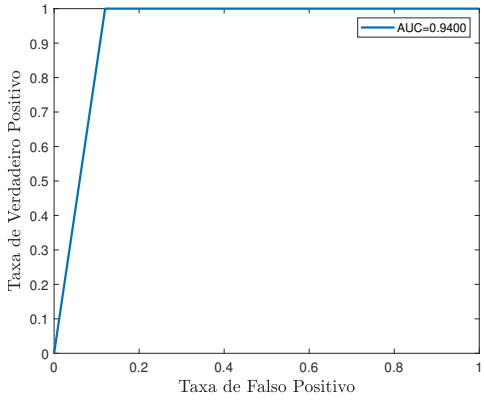


Fig. 9: Curva ROC para o tamanho 64.

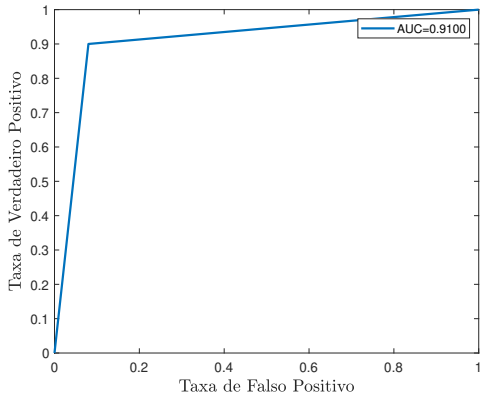


Fig. 10: Curva ROC para o tamanho 128.

[2] W. H. Organization, "Origin of sars-cov-2, 26 march 2020," Technical documents, 2020.

[3] World Health Organisation, "Coronavirus disease (covid-19)," 2020, https://www.who.int/emergencies/diseases/novel-coronavirus-2019, Last accessed on 2021-01-11.

[4] Y. Yang, M. Yang, C. Shen, F. Wang, J. Yuan, J. Li, M. Zhang, Z. Wang, L. Xing, J. Wei *et al.*, "Laboratory diagnosis and monitoring the viral shedding of 2019-ncov infections," *MedRxiv*, 2020.

[5] R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, W. Wang, H. Song, B. Huang, N. Zhu *et al.*, "Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding," *The lancet*, vol. 395, no. 10224, pp. 565–574, 2020.

[6] A. Abbas, M. M. Abdelsamea, and M. M. Gaber, "Classification of covid-19 in chest x-ray images using detrac deep convolutional neural network," *Applied Intelligence*, vol. 51, no. 2, pp. 854–864, 2021.

[7] S. M. Naeem, M. S. Mabrouk, S. Y. Marzouk, and M. A. Eldosoky, "A diagnostic genomic signal processing (gsp)-based system for automatic feature analysis and detection of covid-19," *Briefings in bioinformatics*, vol. 22, no. 2, pp. 1197–1205, 2021.

[8] A. Lopez-Rincon, A. Tonda, L. Mendoza-Maldonado, D. G. Mulders, R. Molenkamp, C. A. Perez-Romero, E. Claassen, J. Garssen, and A. D. Kraneveld, "Classification and specific primer design for accurate detection of sars-cov-2 using deep learning," *Scientific reports*, vol. 11, no. 1, pp. 1–11, 2021.

[9] S. Pei, R. Dong, R. L. He, and S. S.-T. Yau, "Large-scale genome comparison based on cumulative fourier power and phase spectra: central moment and covariance vector," *Computational and structural biotechnology journal*, vol. 17, pp. 982–994, 2019.

[10] 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.

[11] J. A. Morales, R. Saldaña, M. H. Santana-Castolo, C. E. Torres-Cerna, E. Borrayo, A. P. Mendizabal-Ruiz, H. A. Vélez-Pérez, and G. Mendizabal-Ruiz, "Deep learning for the classification of genomic signals," *Mathematical Problems in Engineering*, vol. 2020, 2020.

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

[13] H. K. Kwan and S. B. Arniker, "Numerical representation of dna sequences," in *2009 IEEE International Conference on Electro/Information Technology*. IEEE, 2009, pp. 307–310.

[14] G. Mendizabal-Ruiz, I. Román-Godínez, S. Torres-Ramos, R. A. Salido-Ruiz, and J. A. Morales, "On dna numerical representations for genomic similarity computation," *PloS one*, vol. 12, no. 3, p. e0173288, 2017.

[15] D. Anastassiou, "Genomic signal processing," *IEEE signal processing magazine*, vol. 18, no. 4, pp. 8–20, 2001.

[16] T. Hoang, C. Yin, and S. S.-T. Yau, "Numerical encoding of dna sequences by chaos game representation with application in similarity comparison," *Genomics*, vol. 108, no. 3-4, pp. 134–142, 2016.

[17] N. G. Nguyen, V. A. Tran, D. L. Ngo, D. Phan, F. R. Lumbanraja, M. R. Faisal, B. Abapihi, M. Kubo, K. Satou *et al.*, "Dna sequence classification by convolutional neural network," *Journal of Biomedical Science and Engineering*, vol. 9, no. 05, p. 280, 2016.

[18] 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, no. 9, p. e0222271, 2019.

[19] R. d. M. Barbosa and M. A. Fernandes, "Chaos game representation dataset of sars-cov-2 genome," *Data in brief*, vol. 30, p. 105618, 2020.

[20] C. Yin and S. S.-T. Yau, "An improved model for whole genome phylogenetic analysis by fourier transform," *Journal of theoretical biology*, vol. 382, pp. 99–110, 2015.

[21] H. J. Jeffrey, "Chaos game representation of gene structure," *Nucleic acids research*, vol. 18, no. 8, pp. 2163–2170, 1990.

[22] K. Sedlar, H. Skutkova, M. Vitek, and I. Provaznik, "Set of rules for genomic signal downsampling," *Computers in biology and medicine*, vol. 69, pp. 308–314, 2016.

[23] A. Oppenheim, A. Willsky, and I. Young, "Signals and systems, prentice-hall, inc," *Englewood Cliffs, NJ*, 1983.

[24] A. R. Marcal, "Evaluation of chaos game representation for comparison of dna sequences," in *International Workshop on Combinatorial Image Analysis*. Springer, 2018, pp. 179–188.

[25] 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.

[26] I. Messaoudi, A. Elloumi-Oueslati, and Z. Lachiri, "Building specific signals from frequency chaos game and revealing periodicities using a smoothed fourier analysis," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 11, no. 5, pp. 863–877, 2014.

[27] 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.

[28] J. A. Morales, R. Saldaña, M. H. Santana-Castolo, C. E. Torres-Cerna, E. Borrayo, A. P. Mendizabal-Ruiz, H. A. Vélez-Pérez, and G. Mendizabal-Ruiz, "Deep learning for the classification of genomic signals," *Mathematical Problems in Engineering*, vol. 2020, 2020.

[29] Z. H. Hoo, J. Candlish, and D. Teare, "What is an roc curve?" 2017.

[30] K. Zheng, L. Wang, and Z.-H. You, "Cgmda: an approach to predict and validate microrna-disease associations by utilizing chaos game representation and lightgbm," *IEEE Access*, vol. 7, pp. 133 314–133 323, 2019.

[31] A. S. Nair and S. P. Sreenadhan, "A coding measure scheme employing electron-ion interaction pseudopotential (eiip)," *Bioinformation*, vol. 1, no. 6, p. 197, 2006.

[32] S. Goel *et al.*, "A compression algorithm for dna that uses ascii values," in *2014 IEEE International Advance Computing Conference (IACC)*. IEEE, 2014, pp. 739–743.

[33] J. T. Hancock and T. M. Khoshgoftaar, "Survey on categorical data for neural networks," *Journal of Big Data*, vol. 7, no. 1, pp. 1–41, 2020.

[34] R. F. Voss, "Evolution of long-range fractal correlations and 1/f noise in dna base sequences," *Physical review letters*, vol. 68, no. 25, p. 3805, 1992.