Cadernos Cajuína

Revista Interdisciplinar

e249526

Mineração de dados: o papel da IA durante a pandemia da COVID-19

Data Mining: The Role of AI During the COVID-19 Pandemic

Marcos Vinicius Rossetto[1]

Ivaine Tais Sauthier Sartor[2]

Scheila de Avila e Silva[3]

RESUMO: O presente artigo realiza a descrição da pandemia causada pela COVID-19 e os impactos que a mesma trouxe para o mundo, ressaltando o papel que a mineração de dados e inteligência artificial tiveram durante esse período e os para os próximos possíveis acontecimentos. Para isso, o artigo realiza a revisão da literatura, trazendo a evolução da COVID-19, sua taxinomia e patogênese. Posteriormente, realizamos uma visita a literatura sobre mineração de dados, descrevendo o processo de descoberta de conhecimento em bases dados, como é realizado esses processos e quais são os passos necessários. Dentro desse contexto, os autores descrevem casos de uso de estudos que realizar o uso de algoritmos de inteligência artificial para triar e diagnosticar pacientes e que auxiliaram na análise genômica do vírus e desenvolvimento de novos medicamentos. Desta forma, esse trabalho destaca a importância das tecnologias de IA na resposta a crises sanitárias, sugerindo seu potencial para enfrentar desafios futuros na saúde pública.

PALAVRAS-CHAVE: COVID-19, Mineração de Dados, Inteligência Artificial, Pandemia.

ABSTRACT: This article describes the COVID-19 pandemic and its global impacts, highlighting the role that data mining and artificial intelligence played during this period and for potential future events. To achieve this, the article reviews the literature, presenting the evolution of COVID-19, its taxonomy, and pathogenesis. Subsequently, we review the literature on data mining, describing the knowledge discovery process in databases, how these processes are carried out, and the necessary steps involved. Within this context, the authors describe case studies that utilized artificial intelligence algorithms to screen and diagnose patients, assist in the genomic analysis of the virus,

[1] Graduação em Sistemas de Informação pela Universidade de Caxias do Sul (UCS). Mestre em Biotecnologia pela Universidade de Caxias do Sul (UCS) e Doutorando em Biotecnologia pela Universidade de Caxias do Sul (UCS). ORCID Link: https://orcid.org/0000-0002-6310-5913 E-mail: mvrossetto@ucs.br

[2] Graduação em Licenciatura e Bacharelado Em Ciências Biológicas pela Universidade de Caxias do Sul (UCS). Mestrado em Ciências Biológicas pela Universidade Federal do Rio Grande do Sul (UFRGS) e Doutorado em Genética e Biologia Molecular pela Universidade Federal do Rio Grande do Sul (UFRGS). ORCID Link: https://orcid.org/0000-0002-8775-6622 E-mail: ivaine.sauthier@gmail.com

[3] Graduação em Ciências Biológicas pela Universidade de Caxias do Sul (UCS). Mestrado em Computação Aplicada pela Universidade do Vale do Rio dos Sinos (UNISINOS) e Doutora em Biotecnologia pela Universidade de Caxias do Sul (UCS). E-mail: sasilva6@ucs.br

ISSN: 2448-0916

and develop new medications. Thus, this work emphasizes the importance of AI technologies in responding to health crises, suggesting their potential to address future public health challenges.

KEYWORDS: COVID-19, Data Mining, Artificial Intelligence, Pandemic.

1. INTRODUÇÃO

A pandemia de COVID-19, que começou em dezembro de 2019 em Wuhan, na China, rapidamente se transformou em uma crise sanitária global, afetando milhões de vidas e desafiando sistemas de saúde ao redor do mundo. O surgimento do vírus SARS-CoV-2 e sua rápida disseminação forçaram a Organização Mundial da Saúde (OMS) a declarar oficialmente o estado de pandemia em março de 2020 (Muralidar *et al.*, 2020). Desde então, o impacto foi devastador, com mais de 704 milhões de casos confirmados globalmente até maio de 2024, e o Brasil figurando entre os países mais afetados, com mais de 38,7 milhões de casos e mais de 711 mil mortes registradas (Our World in Data, 2024).

Ao longo da pandemia, a ciência e a tecnologia desempenharam papéis cruciais na resposta à crise. O desenvolvimento acelerado de vacinas e a implementação de novas abordagens terapêuticas trouxeram avanços significativos. Entretanto, a rápida evolução do vírus, com o surgimento de variantes como Gama, Delta e Omicron, trouxe desafios adicionais, exigindo adaptações contínuas nas estratégias de controle e tratamento (Forchette et al., 2021). Nesse cenário, as tecnologias de inteligência artificial (IA) e a mineração de dados emergiram como ferramentas fundamentais para analisar e prever padrões epidemiológicos, gerenciar recursos e apoiar a descoberta de novas soluções terapêuticas.

A aplicação da IA durante a pandemia foi diversificada. Desde a análise de grandes volumes de dados de saúde pública para prever surtos e direcionar a alocação de recursos, até a triagem hospitalar e o diagnóstico de pacientes, a IA proporcionou ganhos consideráveis de eficiência e precisão (Forchette et al., 2021). Além disso, a IA foi essencial na análise genômica do SARS-CoV-2, acelerando o desenvolvimento de vacinas e medicamentos. Esses avanços evidenciam o impacto imediato da IA durante a pandemia e sugerem seu enorme potencial para enfrentar futuras crises sanitárias.

Diante desse contexto, o presente artigo busca responde a seguinte pergunta: De que forma a mineração de dados pode ser aplicada à pandemia de COVID-19 para identificar padrões e fornecer insights que contribuam para o enfrentamento de futuras emergências de saúde pública?

Este estudo tem como objetivo investigar como a mineração de dados foi utilizada durante a pandemia de COVID-19, explorando suas aplicações mais avançadas, como triagem hospitalar, diagnóstico por imagem e análise genômica. Através de uma revisão da literatura, busca-se compreender o estado da arte nessas tecnologias, oferecendo subsídios valiosos tanto para o aprimoramento das respostas a crises atuais quanto para a preparação frente a futuras emergências globais.

2. REVISÃO BIBLIOGRÁFICA

Este capítulo aborda aspectos do surgimento e classificação desse novo agente etiológico, bem como informações genéticas acerca do SARS-CoV-2 e epidemiologia viral, e especialmente sobre a fundamentação e aplicabilidade de métodos de mineração de dados na COVID-19.

2.1. COVID-19

A mais recente epidemia de vírus CoV iniciou em dezembro de 2019, em Wuhan, na China, em um mercado de alimentos que vendia animais vivos, na qual, os morcegos foram apontados como hospedeiros e o agente causador foi classificado como SARS-CoV-2 (Kirtipal *et al.*, 2020). No início de 2020, o Centro de Controle de Doenças e Prevenção Chinês detectou, em pessoas hospitalizadas, o novo CoV, que levou à nomenclatura do vírus como CoV-19. Além da China, a Tailândia anunciou em janeiro de 2020 a confirmação laboratorial do primeiro caso que foi importado de Wuhan (Muralidar *et al.*, 2020). Gradativamente, foram relatados casos em Macau, Estados Unidos, Hong Kong e, casos de disseminação por pessoas que haviam viajado para Wuhan. A rápida disseminação e contágio do vírus levou à OMS (Organização Mundial de Saúde) a decretar, em 11 de março de 2020, a pandemia COVID-19 (Muralidar *et al.*, 2020).

O primeiro caso de COVID-19 no Brasil, foi notificado em 26 de fevereiro de 2020, na cidade de São Paulo e em 4 de setembro de 2020 aproximadamente 125 mil mortes já tinham sido relatadas (Hallal *et al.*, 2020). O Brasil, em agosto de 2021 teve o terceiro maior número de casos cumulativos de COVID-19 no mundo, segundo a Fundação Oswaldo Cruz (Penetra *et al.* 2022). O estado do Amazonas foi mais afetado pela pandemia até o dia 31 de julho de 2021 com 13.531 mortes registradas (Naveca *et al.* 2022). Em março de 2020, início o primeiro pico, atingindo o máximo em maio de 2020, o qual foi associado às linhagens B.1.195 e B.1.1.28. Posteriormente, o segundo pico iniciou-se em dezembro de 2020, com máximo em fevereiro de 2021, no qual foi identificada a linhagem Gama, após ser detectada em japoneses que haviam visitado o estado do

ISSN: 2448-0916

Amazonas (Naveca *et al.* 2022). Os números do Brasil[1], atualizados em maio de 2024, são superiores a 38,791,997 de casos, totalizando 711,964 óbitos, já no mundo os números superam 704 milhões de casos, e um número acumulado de mais de 7 milhões de óbitos[2].

Em 31 de maio de 2021, a OMS anunciou as novas variantes advindas da COVID-19, utilizando letras gregas para classificá-las (Alfa, Gama, Beta e Delta). Muitas das cepas tiveram suas mutações ligadas à proteína S do vírus, mudando o comportamento e patogênese escapando da imunidade natural (Forchette *et al.*, 2021). De acordo com o Dashboard da Our World in Data (https://ourworldindata.org), em 18 de junho de 2024, foram administradas globalmente mais de 13.58 bilhões de doses de vacina contra a Covid-19. No entanto, mesmo com a rápida produção, as novas variantes do COVID-19, tiveram modificações estruturais genéticas, que burlaram as vacinas convencionais (Forchette *et al.*, 2021).

2.2. TAXONOMIA DO CORONAVÍRUS

Coronaviridae é um agrupamento monofilético4 da ordem Nidovirales, família Coronaviridae, subfamília Orthochoronaviridae e gênero Betacoronavirus. Os membros desta família são divididos em 4 gêneros: Alphacoronavirus, Betacoronavirus, Gammacoronavirus e Deltacoronavirus (Coronaviridae Study Group of the International Committee on Taxonomy of Viruses, 2020). A classificação taxonômica dos Coronaviridae, de acordo com o Comitê Internacional de Taxonomia de Vírus (ICTV), é apresentada na Figura 1.

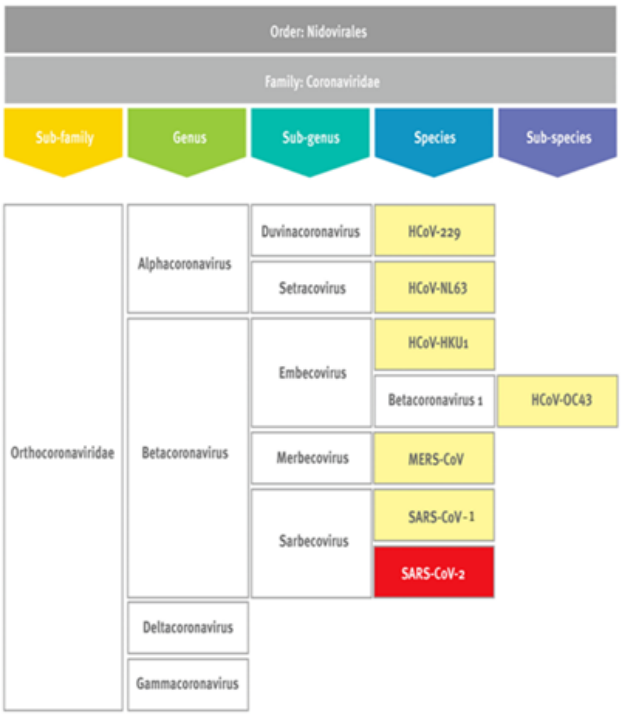

FIGURA 1- TAXONOMIA DO CORONAVÍRUS HUMANO SEGUNDO A ICTV
FONTE: EUROPEAN CENTRE FOR DISEASE PREVENTION AND CONTROL (ECDC). CORONAVIRUSES.

Os Coronavírus foram identificados em diversas espécies reservatórias, como morcegos, camundongos, ratos, galinhas, cães, gatos, cavalos e camelos (Lu *et al.*, 2020, Sharma *et al.* 2021). Algumas espécies causam doenças pandêmicas em mamíferos e aves domésticas e selvagens, provocando mortalidade em criações e perdas econômicas. Esses vírus incluem coronavírus de galinhas (Infectious bronchitis virus - IBV), coronavírus de baleia beluga (Beluga whale CoV - BWCoV-SW1), coronavírus de morcego (CDPHE15 e HKU10), vírus de diarreia epidêmica suína (Porcine epidemic diarrhea virus - PEDV) e síndrome de diarreia aguda súbita em porcos (Swine acute diarrhea syndrome coronavirus - SADS-CoV), conforme compilado por Helmy e colaboradores (2020) e reproduzido na Figura 1.

Os gêneros Alphacoronavirus e Betacoronavirus são conhecidos por infectar humanos. Nesses gêneros, seis espécies de coronavírus podem causar doenças em seres humanos. Os coronavírus 229E, OC43, NL63 e HKU1 são prevalentes e geralmente causam sintomas de resfriado comum. As cepas de coronavírus da síndrome respiratória aguda grave (*Severe acute respiratory syndrome coronavirus* - SARS-CoV) e coronavírus da síndrome respiratória do Oriente

Médio (Middle East respiratory syndrome coronavirus - MERS-CoV) são de origem zoonótica e têm sido associadas a doenças mais graves e às vezes fatais.

Em relação ao SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus), estudos filogenéticos (Wu et al., 2020; Coronaviridae Study Group of the International Committee on Taxonomy of Viruses, 2020) apontam semelhanças com os coronavírus de morcego, com o SARS-CoV e com a MERS-CoV (Zhu et al, 2020; Sharma et al. 2021).

2.3 PATOGÊNESE

A infecção por SARS-CoV-2 têm início nas células epiteliais nasais e brônquicas. Conforme a replicação viral evoluiu para estágios posteriores, o vírus infecta diretamente as células endoteliais em vários órgãos, podendo causar a ruptura da barreira endotelial e lesão difusa das células endoteliais (Varga *et al.,* 2020; Wiersinga *et al.*, 2020).

A infecção é iniciada com a ligação da partícula viral na superfície externa ao receptor da superfície do hospedeiro por meio dos picos de glicoproteína (proteína spike) (Wit *et al.,* 2016; Kadam *et al.*, 2020). Estudos sobre o mecanismo de entrada do vírus SARS-CoV-2 na célula hospedeira indicam que é utilizado o receptor celular ACE2, similar ao mecanismo descrito para o vírus SARS-CoV (Wit *et al.*, 2016; Wan *et al.*, 2020). O receptor ACE2 está presente nas membranas celulares de vários órgãos, incluindo pulmões, artérias, rins, coração e intestinos (Kadam *et al.*, 2020, Varga *et al.*, 2020).

O vírus entra na célula por meio de endocitose mediada por pH e receptor do hospedeiro, processo no qual o nucleocapsídeo viral é entregue no citoplasma. Ao infectar a célula hospedeira, fatores do hospedeiro interagem com o RNA viral nesses locais e participam da síntese do RNA viral. Deste modo, o genoma atua como um mRNA para tradução das poliproteínas replicases necessárias para a replicação viral (WIT et al., 2016; KADAM et al., 2020; VARGA et al., 2020). A representação esquemática da patogênese é mostrada na Figura 2.

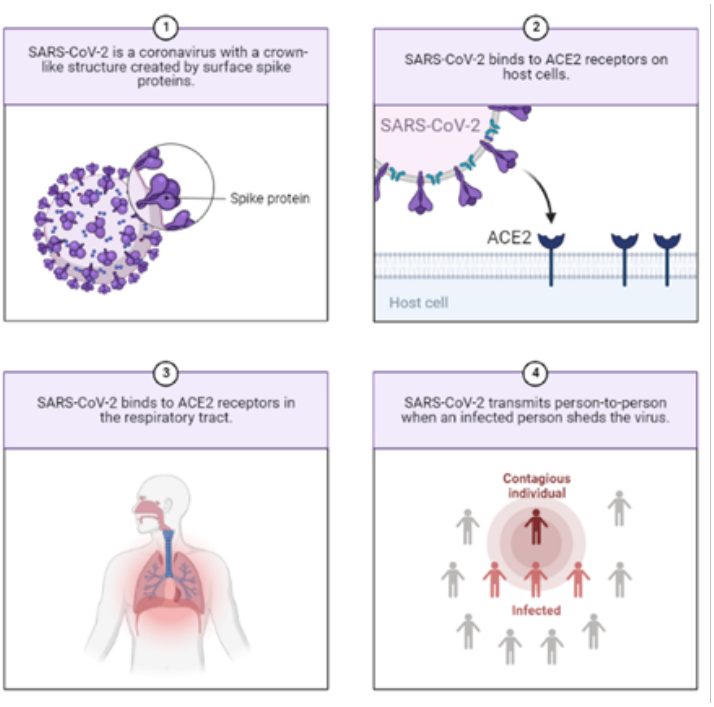

FIGURA 2- PATOGÊNESE CORONAVÍRUS HUMANO

2.4 SINTOMATOLOGIA E EPIDEMIOLOGIA DO VÍRUS

No início da pandemia, por se tratar de uma doença emergente, as formas de transmissão ainda estavam sendo estudadas. Acreditava-se que a doença era transmitida apenas de animais para humanos, depois apenas entre pessoas sintomáticas (Helmy *et al.,* 2020; Wiersinga *et al.*, 2020), até que o primeiro caso de transmissão humano a humano de um portador assintomático foi documentado na Alemanha (Rothe *et al.*, 2020). Entende-se que o SARS-CoV-2 é transmitido entre humanos por contato direto, gotículas de aerossol, via fecal-oral e fômites intermediários de pacientes sintomáticos e/ou assintomáticos durante o período de incubação (Helmy *et al.*, 2020).

O período estimado de incubação varia de 2 a 14 dias, com tempo médio de 6 dias. No entanto, alguns casos apresentaram períodos de incubação diversos, compreendendo 21, 24 ou até 27 dias (Helmy *et al.*,2020; Wiersinga *et al.*, 2020).

A fisiopatologia e a gravidade da doença COVID-19 variam entre os pacientes e dependem em parte de fatores de risco subjacentes e doenças crônicas. Usualmente, a COVID-19 apresenta-se com sintomas como: febre, tosse seca, dispneia, espirros ou dor de garganta e diarreia (Helmy *et al.*,2020; Wiersinga *et al.*, 2020). Em casos graves, a doença evolui para o desenvolvimento de pneumonia, acidose metabólica, choque séptico e sangramento associado com baixa contagem de

leucócitos e linfócitos; resposta inflamatória acentuada e parâmetros anormais de coagulação (Levi *et al.*, 2020; Wiersinga *et a*l., 2020; Jahanafrooz *et al.,* 2022).

Em relação às medidas preventivas e precauções relacionadas à limitação à exposição ao vírus e redução da propagação, destacam-se: (1) lavar as mãos frequentemente com água e sabão ou desinfetante para as mãos à base de álcool, (2) etiqueta ao tossir ou espirrar, recomendando cobrir a boca, (3) evitar tocar nos olhos, nariz, e boca se as mãos não estiverem limpas, (4) evitar contato próximo com pessoas doentes, (5) evitar compartilhar pratos, copos, roupas de cama e outros utensílios domésticos com pessoas doentes, (6) limpeza e desinfecção de superfícies que são frequentemente tocadas, e (7) ficar em casa longe do trabalho, escola e áreas públicas quando se sentir doente (Helmy *et al.*,2020).

2.5 INTELIGÊNCIA ARTIFICIAL NA SAÚDE

A recente pandemia causada pela COVID-19 ressalta a importância da realização de pesquisas relacionadas aos vírus emergentes e altamente infectantes em um cenário globalizado. Tanto a ciência, quanto a tecnologia estão em processos de evolução contínua, e o grande número de estudos sobre SARS-CoV-2 demonstram a agilidade e a capacidade da comunidade científica mundial. A pandemia contribuiu para estimular este desenvolvimento, assim como o melhor preparo da sociedade globalizada para as próximas pandemias que podem surgir. Para auxiliar nesse aspecto, projetos de análise de dados com abordagens de mineração de dados apresentam aplicabilidade.

2.5.1 Mineração de Dados: Processo e Aplicações

Projetos analíticos de mineração de dados envolvem cinco fases, sendo elas: seleção dos dados, pré-processamento, aplicação de algoritmos de inteligência Artificial (IA), a validação dos resultados, e por fim o conhecimento gerado (Fayyad; Haussler; Stolorz, 1996; Navathe, 2013). Assim, através de algoritmos que retornam padrões e tendências nos dados analisados é possível realizar inferências que são analisadas e validadas por especialistas de domínio (Silva *et a*l, 2016; Carvalho *et al,* 2011). A figura 3 ilustra o fluxo de atividades relacionadas a um projeto de inteligência artificial.

FIGURA 3 - PROCESSOS DO KDD (KNOWLEDGE DISCOVERY IN DATABASES) - DESCOBERTA DO CONHECIMENTO EM BASES DE DADOS. FONTE: DANI ET AL., 2020

A seleção dos dados, envolve a organização e estruturação dos dados, a fim de criar um repositório único, integrando múltiplas fontes que podem ser apresentados de duas maneiras: dados estruturados e não estruturados. Dados estruturados são organizados em estruturas tabeladas, de modo que as linhas armazenam o dado em si e as colunas representam as variáveis. Os dados não estruturados, como o próprio nome diz, não contém uma estrutura padrão, sendo que esses podem ser apresentados por textos, imagens, vídeos e sons (Silva *et al*, 2016; Amaral, 2016).

No pré-processamento é realizada inicialmente a limpeza e transformação dos elementos, com o objetivo de selecionar e filtrar dados ruidosos, inconsistentes e dados faltantes que podem afetar a qualidade do processo de mineração de dados (Navathe, 2013; Silva *et al*, 2016; Goldschmidt & Passos, 2015)

Apesar da complexidade matemática associada aos algoritmos de inteligência artificial, a qualidade dos dados representa papel importante no desempenho dos algoritmos. A base de dados analisada pelo algoritmo será a base para a construção do modelo. A passagem de um conjunto específico de observações para uma regra geral é chamada indução. A carga de raciocínio deste tipo de aprendizagem é maior do que a aprendizagem de demonstração e aprendizagem dedutiva, porque o ambiente não fornece descrições conceituais gerais (como axiomas). Assim, a indução fornece a resposta para o que falta no conhecimento a partir de uma entrada e saída conhecidas, sendo o modelo de aprendizado explorado pelos algoritmos de aprendizagem de máquina (RusselL & Norvig, 2022).

Especificamente, ao se tratar com dados de origem biológica, têm-se uma dificuldade adicional em aplicações de inteligência artificial, principalmente devido à heterogeneidade natural dos sistemas biológicos. Os dados biológicos variam de estruturas químicas ou moleculares únicas a sistemas complexos em uma escala genômica ampla ou até redes metabólicas completas. Assim, variam em complexidade, formato e escala, incluindo: (i) sequências (DNA, RNA e proteínas, geralmente em formato de texto); (ii) estruturas (moléculas biológicas e químicas, como proteínas estruturais e enzimas envolvidas em vias, em formato de imagem); (iii) perfis de expressão gênica (medição da atividade gênica, em formatos numéricos e de imagem); (iv) vias bioquímicas (em formato de texto ou imagem); (v) mapeamento cromossômico (em formato de texto ou imagem); (vi) Polimorfismos de Nucleotídeo Único (em formato de texto); e (vii) dados filogenéticos (em formato de texto ou imagem). Também é importante ressaltar que os dados biológicos apresentam algumas interferências decorrentes da grande variedade de equipamentos e protocolos utilizados em um determinado experimento. Isso é conhecido como "efeito de lote", que é o resultado de variáveis de origem laboratorial, como reagentes, máquinas, erro humano, entre outros fatores que podem interferir em um experimento. Ao lidar com grandes quantidades de dados e dados de alto rendimento, os efeitos de lote podem ser a fonte de resultados enganosos (Dall'alba *et al*, 2022).

Adicionalmente, ressalta-se preocupações éticas. Como ressaltado por Nunes e Marques (2018), a qualidade dos dados fornecidos aos sistemas de inteligência artificial causará impacto nos resultados, pois os dados são coletados de uma sociedade repleta de desigualdades, exclusões e discriminações (Nunes e Marques, 2018). Corroborando com os autores, Goodman e Flaxman (2017), discorrem que o aprendizado de máquina pode validar normas discriminatórias porque, se tais normas foram descobertas em um banco de dados de dados, um algoritmo de classificação exato irá reproduzi-las (Goodman & Flaxman, 2017).

2.5.2 Algoritmos de Aprendizado de máquina

Após o pré-processamento, os algoritmos de inteligência artificial são aplicados nos dados. Trata-se de um processo de escavação de conjuntos de dados com grandes volumes, os quais são analisados e deles extraídos informações, padrões, tendências, entre outros (Dani et al., 2020).É expressiva a quantidade de dados coletados atualmente e disponíveis para análises, entretanto, a escolha do método correto que consiga abranger essa quantidade de informação, nem sempre se traduz em mais conhecimento. Para tanto, a utilização da mineração de dados para extração de

ISSN: 2448-0916

informação e conhecimento é indicada pois consegue lidar com enorme volume de dados, minimizando o ruído proveniente dos mesmos (Dani et al., 2020).

O objetivo da inteligência artificial é modelar matematicamente tarefas de classificação, análise, previsão e manipulação para construir sistemas inteligentes (Russell, 2010). Para isso, a inteligência artificial utiliza de um conjunto de princípios, modelos e ferramentas em diferentes subdomínios, como: lógica, raciocínio sob incerteza, agentes inteligentes, aprendizado de máquina (Ertel, 2017). O aprendizado de máquina (ou *machine learning*, ML) pode ocorrer de diferentes formas, como: supervisionada, não supervisionado, por reforço ou por transferência.

O aprendizado de máquina supervisionado, se dá a partir do treinamento de um algoritmo baseado em um conjunto de dados com exemplos e respostas esperadas, isto é, o banco de dados possui valores numéricos ou textuais que são utilizados pelos algoritmos para inferir as respostas. Exemplos de uso do ML supervisionado, são os algoritmos de regressão, cujo a resposta alvo é um valor numérico, ou um algoritmo de classificação, cujo a resposta alvo é uma variável qualitativa, como uma classe ou etiqueta (Mueller & Massaron, 2019).

Por outro lado, o aprendizado de máquina não supervisionado, ocorre a partir de dados não rotulados, sem nenhuma resposta associada. Nesse caso, a determinação de padrões se dá com base na similaridade dos dados. Esse tipo de algoritmo tende a estruturar os dados considerando características que representam uma classe ou uma nova série de valores não correlacionados, que podem ser utilizados para realizar o agrupamento de dados, como o agrupamento de clientes em diferentes categorias ou detecção de fraudes/anomalias em operações financeiras (Mueller & Massaron, 2019; Dos Santos *et al*, 2020).

O aprendizado por reforço acontece quando é fornecido ao algoritmo exemplos que não contém rótulos, como também é feito no aprendizado não supervisionado. Entretanto, um exemplo pode ser acompanhado de um retorno positivo ou negativo, dependendo da solução sugerida pelo algoritmo (Mueller & Massaron, 2019). Nesse caso, os erros auxiliam no aprendizado, pois trazem uma penalidade, ensinando ao algoritmo quais são as decisões com menos penalidades, maximizando a recompensa recebida ao longo de sua execução, (Mueller & Massaron, 2019; Dos Santos *et al*, 2020).

A aprendizagem por transferência (Transfer Learning – TL), refere-se a uma coleção de modelos em que o conhecimento adquirido por outro modelo ao resolver um problema. Esse conhecimento adquirido é salvo e depois transferido para outro modelo encarregado de resolver um problema comparável. Essa transferência reduz ou elimina a necessidade de treinamento de

ISSN: 2448-0916

novos modelos, além de permitir o avanço desse conhecimento (Dos Santos *et al.*, 2020). Como é aprofundado por Moraes (2021), a TL, é uma estrutura de aprendizado que relaxa a suposição dos algoritmos de aprendizado assistido por supervisão de que os dados de treinamento e teste devem ser independentes e identicamente distribuídos, ou seja, representados pelas mesmas características e demonstrados pelo mesmo gerador. Esse relaxamento aborda a questão de ter dados insuficientes para treinar alguns algoritmos de aprendizado de máquina, principalmente aqueles que se concentram em aprendizado profundo. Em *Deep Learning*, o sucesso da extração de padrões e características ocultas está diretamente relacionado ao treinamento com grande volume de dados. Nessa situação, a transferência de aprendizado possibilita a reutilização de princípios aprendidos em outro domínio, reduzindo os requisitos de dados e acelerando a convergência do modelo (Tan *et al.*, 2018).

Os principais modelos de aprendizado de máquina usados em cada uma dessas categorias e seus usos pretendidos estão resumidos na tabela 1.

TABELA 1 - CATEGORIA DOS MODELOS DE APRENDIZADO DE MÁQUINA

Categoria	Modelos
Supervisionado	Regressão linear, regressão logística, máquinas de vetores de suporte (SVM - *support vector machine*), classificador Bayesiano, árvores de decisão, análise de discriminantes lineares, k-Nearest Neighbors (KNN), redes neurais.
Não supervisionado	Agrupamento (*clustering*), detecção de anomalias, redes neurais, modelos de variáveis latentes.
Por reforço	Modelos Markovianos, métodos de Monte Carlo, diferenças temporais, redes neurais recorrentes (RNN - *recurrent neural network*), redes neurais convolucionais (CNN - *Convolutional neural network*).
Por Transferência	Inception-v3, ResNet, AlexNet, outros modelos baseados em redes neurais convolucionais (CNN).
Profundo	Redes neurais recorrentes (RNN), redes neurais convolucionais (CNN), modelos generativos (Boltzmann, GAN, *deep belief*).

Fonte: (Dos Santos *et al*, 2020).

2.5.3 Avaliação e Validação dos Algoritmos

A Avaliação e validação dos resultados, são conceitos importantes no campo da estatística e referem-se ao processo de avaliação do desempenho e confiabilidade de modelos e métodos estatísticos.

A Avaliação é o processo de verificação da qualidade ou eficácia de um modelo ou método estatístico. Isso pode ser feito de várias maneiras, como por meio do uso de métricas de avaliação, que são medidas quantitativas usadas para avaliar o desempenho de um modelo. Algumas métricas de avaliação comuns incluem acurácia, sensibilidade, precisão, F1-Score e especificidade. A avaliação também pode ser realizada por meio do uso de visualizações, como plotagens ou gráficos, que podem ajudar a identificar padrões ou tendências nos dados que podem não ser imediatamente aparentes nos dados brutos (Nisbet, Elder & Miner, 2009).

A validação é o processo de verificar se um modelo ou método estatístico é preciso e confiável. Isso pode ser feito por meio do uso de validação cruzada, que é um procedimento de reamostragem que envolve particionar os dados em um conjunto de treinamento e um conjunto de testes, treinando o modelo no conjunto de treinamento e avaliando o modelo no conjunto de testes. A validação também pode ser realizada por meio do uso de fontes de dados externas, como dados históricos ou conjuntos de dados independentes, que podem ser usados para verificar a precisão e a confiabilidade do modelo (Nisbet, Elder & Miner, 2009).

A validação cruzada é um procedimento de reamostragem usado para avaliar o desempenho de um modelo de aprendizado de máquina. Envolve particionar os dados em um conjunto de treinamento, que é usado para treinar o modelo, e um conjunto de teste, que é usado para avaliar o modelo. A validação cruzada é uma técnica importante porque permite avaliar o desempenho do modelo em dados não vistos, o que pode ajudar a identificar o *overfitting* e *underfitting* que pode ter ocorrido durante o treinamento (Kohavi, 1995).

Existem diferentes tipos de validação cruzada, mas um método comum é a validação cruzada k-fold. Na validação cruzada k-fold, os dados são particionados em k subconjuntos de tamanho igual e o modelo é treinado e avaliado k vezes, com um subconjunto diferente de dados usado como conjunto de teste em cada iteração. O desempenho final do modelo é então calculado como a média do desempenho em cada uma das k iterações, conforme ilustrado na Figura 4 (Kohavi, 1995).

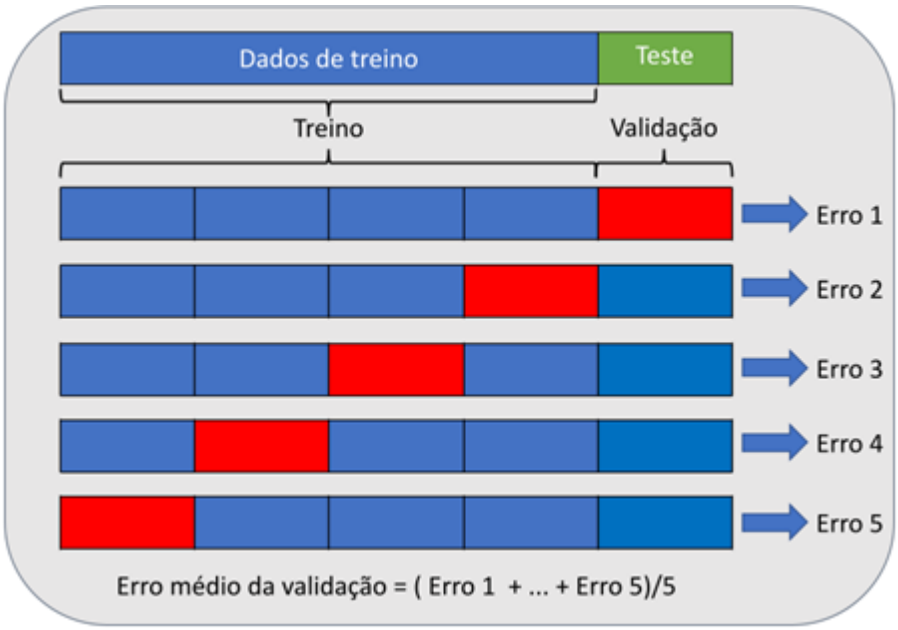

Figura 4- Validação cruzada

Uma matriz de confusão é uma tabela usada para avaliar o desempenho de um modelo de classificação. Ele exibe o número de previsões verdadeiro positivo, verdadeiro negativo, falso positivo e falso negativo feitas pelo modelo e pode ser usado em conjunto com validação cruzada para avaliar a estabilidade e generalização do modelo (Stehman, 1997), ilustrado na Figura 5.

		Valores Reais	
		Negativo -	Positivo +
Predito	Negativo -	Verdadeiro Negativo (TN)	Falso Positivo (FP) **Erro tipo I**
	Positivo +	Falso Negativo (FN) **Erro tipo II**	Verdadeiro Positivo (TP)

Figura 5 - Matriz de confusão

As entradas na matriz de confusão podem ser usadas para calcular uma variedade de métricas de avaliação, como precisão, que é o número de previsões positivas verdadeiras feitas pelo modelo dividido pelo número total de previsões positivas feitas pelo modelo (Stehman, 1997),

ISSN: 2448-0916

a seguir, são descritas as principais medidas para avaliação dos modelos. Sendo elas, acurácia, sensibilidade, precisão, F1 score, especificidade e por fio a curva ROC (METZ, 1978):

A acurácia (*Accuracy*), é o grau em que uma classificação binária identifica ou exclui uma classe, ou seja, a proporção de previsões corretas (verdadeiros positivos e verdadeiros negativos) entre o total de observações. Seu cálculo é alcançado através da equação:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

A sensibilidade (Recall), determina a quantidade de positivos reais capturados e rotulados como TP pelo modelo Altas taxas de verdadeiros positivos e baixas taxas de falsos positivos contribuem para esta métrica. Seu cálculo é feito pela equação:

$$\text{Recall} = TP / (TP + FN)$$

A precisão (*Precision*), calcula quantas das classes classificadas como positivas são realmente positivas. O custo de falsos positivos é alto ao calcular a precisão através de sua fórmula, explicada pela equação:

$$\text{Precision} = TP / (TP + FP)$$

O F1 Score, realiza um balanço entre *Precision* e *Recall*. A métrica é usada para qualquer classificação onde há uma diferença considerável entre *Precision* e *Recall*. Em vez de usar a média aritmética, o F1 Score utiliza a média harmônica, sendo uma função de duas variáveis (*Precision* e *Recall*) ao mesmo tempo e penalizando valores extremos. Seu cálculo segue a equação:

$$\text{F1 Score} = 2 \times \text{Recall} \times \text{Precision} / \text{Recall} + \text{Precision}$$

A especificidade (*Specificity*), mensura a taxa de detecção de TNs entre o conjunto de dados. É penalizada por altos números de falsos positivos. A especificidade é calculada através da equação:

$$\text{Specificity} = TN / (TN + FP)$$

Outra derivação de uma matriz de confusão é a curva ROC (*Receiver Operating Characteristic*), a qual é uma medida de desempenho para problemas de classificação sob diferentes limites. Aplicações de aprendizado de máquina, como ANN (redes neurais artificiais), são capazes de prever a adesão a uma determinada classe. A decisão para classificar uma observação é dada por um parâmetro denominado limite de decisão. Além disso, a ROC também pode ser usada para calcular a Área sob a Curva (AUC), que mede o grau de separação entre os limites do mesmo método de classificação, bem como a comparação de desempenho de diferentes técnicas de classificação (por exemplo, regressão linear versus ANN). Quanto maior o AUC, melhor o modelo previu certa classe (JoshI, 2020).

ISSN: 2448-0916

Curvas ROC podem ser calculadas a partir de qualquer métrica derivada de uma matriz de confusão (Peres *et al.*, 2015). Um cálculo comum deriva de duas métricas concorrentes: recall (também conhecido como sensibilidade), que mede a taxa de verdadeiros positivos (TPR); e taxa de falsos positivos (FPR), que é estimada por 1 - especificidade. Esta relação é calculada de tal forma que quanto maior o TPR, menor será o FPR, como descrito na equação abaixo, onde, FPR mede a taxa de falsos positivos em uma escala de 0 a 1(Bewick *et al.*, 2004).:

$$FPR = 1 - TN / (TN + FP)$$

É importante observar que diferentes métricas de avaliação podem ser mais ou menos apropriadas para um determinado problema, dependendo dos custos e benefícios relativos de diferentes tipos de erros. Por exemplo, em uma tarefa de diagnóstico médico, pode ser mais importante minimizar os falsos negativos (falha na detecção de um caso positivo) em detrimento do aumento dos falsos positivos (detectar incorretamente um caso positivo), pois as consequências de diagnósticos perdidos podem ser mais graves. graves do que as de diagnósticos incorretos (Stehman, 1997).

2.6 APLICAÇÕES DA MINERAÇÃO DE DADOS NA COVID-19

No contexto da pandemia de COVID-19, a mineração de dados tem sido utilizada como ferramenta para dar apoio ao diagnóstico da doença, desenvolvimento de modelos de prevenção, pesquisa clínica, pesquisa farmacêutica, epidemiologia do vírus, apoio à tomada de decisão e ações de combate à pandemia. Desta forma, essa seção realiza uma revisão dos trabalhos que utilizam a mineração de dados e a inteligência artificial na pandemia da COVID-19, exemplificando as suas aplicações.

2.6.1 Diagnóstico, triagem hospitalar

A triagem hospitalar desempenha um papel crítico no controle da disseminação da COVID-19, pois esse processo auxilia na identificação precoce de pacientes infectados, contribuindo para o tratamento e alocação eficaz de recursos hospitalares (Liang *et al.*, 2020). Nesse sentido, trabalhos mostram a aplicabilidade da inteligência artificial na triagem e diagnóstico de pacientes.

A COVID-19 levou muitos pacientes ao atendimento médico em um curto período para o tratamento de uma doença não descrita anteriormente, desafiando os serviços de saúde a fornecer intervenções urgentes em circunstâncias difíceis. Desde que os primeiros casos de COVID-19 foram documentados na China no final de dezembro de 2019, o foco epidemiológico foi principalmente estabelecer a gravidade da doença e os resultados entre pacientes graves. Considerando os fatores de risco para a COVID-19, Sartor *et al.* (2022) descreve uma análise sobre a relação entre obesidade e a infecção por COVID-19 em pacientes adultos ambulatoriais. Um total de 1.050 participantes foram triados em dois departamentos de emergência de hospitais (públicos e privados) da região Sul do Brasil. Destes pacientes, 297 foram diagnosticados com COVID-19 por RT-PCR (*Reverse transcription–polymerase chain reaction*) e completaram o acompanhamento de 28 dias, que se deu por contato telefônico nos dias 7, 14 e 28 após a inclusão no estudo. Noventa e cinco (32,0% da amostra) indivíduos apresentavam obesidade e 233 (78,5% da amostra) não apresentavam nenhuma condição médica prévia. Vinte e três participantes (7,7% da amostra) necessitaram de hospitalização durante o período de acompanhamento. Modelos de regressão logística multivariada foram empregados para explorar a associação entre obesidade e outros preditores potenciais de hospitalização. Após o ajuste do modelo, os índices de obesidade (IMC ≥ 30,0 kg/m 2) (OR = 2,69, IC 95% 1,63-4,83, P < 0,001) e idade avançada (OR = 1,05, IC 95% 1,01-1,09, P < 0,001) foram significativamente associados a maiores riscos de hospitalização.

Visando uma triagem populacional em larga escala e estratégia de diagnóstico rápida e não invasiva com a finalidade de garantir ou assegurar o conforto do paciente, Karunakaran *et al.* (2022), descrevem a utilização do espalhamento Raman aprimorado por superfície (SERS) como uma modalidade de diagnóstico para a discriminação sem marcação da infecção por SARS-CoV-2 a partir de amostras de saliva. Para isso, os autores analisaram a saliva de pacientes infectados e recuperados de COVID-19 utilizando espectrofotômetro Raman portátil, que misturou a amostra com nanopartículas de ouro (AuNPs - Gold nanoparticles) para avaliar os picos Raman aprimorados. Esses dados foram codificados, segregados e gerenciados com o auxílio de algoritmos de aprendizado de máquina de vetor de suporte (SVM). Os modelos de classificação construídos alcançaram uma Accuracy de 95% e F1-score de 94,73% para pacientes saudáveis e 95,28% de Accuracy para pacientes infectados com COVID-19. Os autores sugerem que a abordagem apresentada é capaz de identificar pacientes infectados por SARS-CoV-2 de pacientes saudáveis e diferenciar os estágios da recuperação do paciente. Assim, os autores concluem que o

ISSN: 2448-0916

espectrofotômetro Raman portátil e a saliva como amostra apresenta-se como uma estratégia de triagem e diagnóstico rápida e não invasiva.

Utilizando dados de biomarcadores imunológicos, Martinez *et al.* (2022) analisou dados de 28 biomarcadores diferentes em duas coortes de pacientes com COVID-19 (totalizando 95 pacientes) com o objetivo de capturar, quantificar e programar sistematicamente como os sinais imunológicos podem estar associados ao resultado clínico de COVID-19. Para isso, os autores utilizaram a medida da concentração longitudinal dos biomarcadores. Esses dados foram utilizados como entrada para diferentes técnicas de inteligência artificial: rede neural artificial, floresta aleatória, árvores de classificação e regressão, k-means e máquinas de vetores de suporte. Os autores apresentam 5 biomarcadores imunológicos relacionados à gravidade da COVID-19 e, sugerem que a triagem de tais biomarcadores pode ajudar na compreensão da resposta imune subjacente a doenças inflamatórias. Uma vez que a concentração diferenciada de biomarcadores imunológicos desempenha um papel fundamental na regulação da resposta do hospedeiro contra patógenos. Em relação às medidas de performance associadas aos resultados, a metodologia empregada apresentou valores 94% de acurácia, 96,6% de precisão, 91,6% de recall e 95% de especificidade nos dados do teste. A validade desses valores foi corroborada por meio da utilização do modelo para prever 83% e 87% (recuperados e falecidos) de dados não vistos.

Em relação ao diagnóstico baseado em imagens, dos Santos e colaboradores (2020), descrevem uma abordagem consolidada dentro de diversas áreas do conhecimento. Em uma revisão sistemática, os autores realizaram a busca de artigos na área de saúde empregando o diagnóstico baseado em imagens de raio-X e tomografias computadorizadas para o diagnóstico de infecções respiratórias e outros desfechos associados à COVID-19. Os autores encontraram artigos que nos quais foram realizadas análises de imagens de raio-X e tomografia computadorizada utilizando *deep learning*, sendo que, a maioria dos artigos encontrado pelos autores, utilizaram Redes Neurais Convolucionais para realizar a identificação de padrões específicos nas imagens, como: sinais de tumores, regiões com coloração diferente, desvios de tecidos e artérias etc. Dos Santos e colaboradores (2020), destacam os trabalhos de Shiri *et al.* (2021) e Syed *et al.* (2021) e Pu *et al.* (2020). com este objetivo apresentado anteriormente.

A proposição de um protocolo para o diagnóstico clínico de pacientes COVID-19 positivo, baseado em resultados de tomografia computadorizada (TC) usando a abordagem de redes neurais com aprendizagem profunda foi proposto por Shiri et al. (2021). O objetivo dos autores era comparar os resultados oriundos de TC de dose completa de radiação com resultados

ISSN: 2448-0916

de TC de baixa dose de radiação. A construção do modelo empregou 800, 170 e 171 pares de imagens para as etapas de treinamento, teste e conjunto de validação externa, respectivamente. Os resultados foram avaliados por meio de métricas quantitativas de avaliação: raiz do erro quadrático médio (RMSE - *Root-mean-square deviation*), índice de similaridade estrutural (SSIM - *Structural similarity index measure*) e a relação sinal-ruído de pico (PSNR - *Peak signal-to-noise ratio*). A avaliação clínica foi realizada por especialistas que atribuíram nota de 1 a 5. A pontuação geral atribuída pelos radiologistas, mostrou uma taxa de aceitação de 4,72 ± 0,57 em para imagens de TC alta dose de radiação, enquanto as imagens de TC de dose ultrabaixa foram avaliadas em 2,78 ± 0,9. As imagens de TC previstas que utilizaram o algoritmo de aprendizado profundo alcançaram uma pontuação de 4,42 ± 0,8.

Os autores discutem que, apesar da qualidade inferior de imagens de baixa dose quando comparadas às imagens de TC de dose total, características como infiltrado nodular, consolidação e características de pavimentação irregular, obtiveram pontuações altas, quase semelhantes às imagens de TC de dose completa. Adicionalmente, eles sugerem que esse modelo é capaz de prever imagens padrão de TC de dose completa com qualidade aceitável para o diagnóstico clínico de pacientes positivos para COVID-19 com redução substancial da dose de radiação. Os autores concluem que, os algoritmos de aprendizado profundo não conseguiram recuperar a estrutura/densidade correta da lesão para vários pacientes considerados discrepantes e, como tal, pesquisas e desenvolvimentos adicionais são necessários para lidar com essas limitações.

A análise de raios-x com redes neurais de aprendizado profundo foi realizada por Rangarajan *et al.* (2021), com base no argumento que a radiografia de tórax (RX), apesar de estar entre as investigações mais comuns realizadas em todo o mundo, não foi considerado um exame sensível ou específico para mudanças relacionadas à COVID-19. Assim, os autores desenvolveram e validaram um algoritmo que poderia diferenciar o RX de pacientes positivos para COVID-19 de pacientes negativos para COVID-19. Foram analisadas as radiografias de 487 pacientes classificadas em 4 categorias pela opinião consensual de 2 radiologistas: normal, COVID clássico, indeterminado e não-COVID. As radiografias classificadas como "normais" e "indeterminados" foram submetidas à análise por IA e a categorização final fornecida com o auxílio da classificação da rede neural. A acurácia dos radiologistas melhorou de 65,9% para 81,9% e a sensibilidade melhorou de 17,5% para 71,75% quando foi fornecida com assistência da IA. A inteligência artificial, mostrou 92% de precisão na classificação de RX "normal" em COVID ou não-COVID. Os autores advogam que a implantação dessa ferramenta pode melhorar a sensibilidade e a

ISSN: 2448-0916

precisão dos radiologistas na identificação de pacientes positivos para COVID na prática clínica de rotina. Adicionalmente, os autores relatam que essa abordagem de IA teve um desempenho superior quando o CXR estava completamente normal.

Em relação ao diagnóstico infantil, Fazolo *et al.* (2021), aplicaram análise de componentes principais e agrupamento hierárquico para avaliar a diferença nas respostas imunes celulares ou humorais de pacientes pediátricos e adultos com COVID-19 para verificar se esses fatores contribuem para gravidade da doença. Os autores apresentam no trabalho uma caracterização detalhada do plasma e das células mononucleares do sangue periférico (PBMCs - *Peripheral blood mononuclear cells*) de pacientes adultos e pediátricos com COVID-19 por citometria de fluxo, definindo 78 subconjuntos de células imunes. Adicionalmente, esses dados foram analisados utilizando métodos multivariados (PCA - Principal *component analysis* e agrupamento hierárquico). Assim, os autores sugerem que as crianças produzem uma resposta imune forte, porém diferenciada, quando comparadas aos adultos, o que se associa à manifestação leve da COVID-19 pediátrica. Deste modo, os autores apontam que não apenas o sistema imunológico adaptativo, mas também o inato, possuem características que permitem que as crianças montem uma resposta imune que controle a infecção. Além disso, o mRNA viral é mais abundante em pacientes pediátricos, sustentando a hipótese que as crianças infectadas com SARS-CoV-2 contribuem para a transmissão, mas são menos suscetíveis aos sintomas de COVID-19 devido a respostas fortes e diferenciadas ao vírus.

Em resumo, a triagem e o diagnóstico rápidos são ferramentas para o combate ao COVID-19, uma vez que auxiliam na identificação e isolamento de indivíduos infectados precocemente, reduzindo assim o número de novos casos e a propagação do vírus.

2.6.2 Análise genômica e pesquisa de fármacos

Uma das aplicações da inteligência artificial em genômica é a identificação de potenciais alvos de medicamentos. Ao analisar dados genômicos e dados relacionados às doenças, os algoritmos de IA podem identificar genes, vias metabólicas e dados sobre os pacientes que estão envolvidos no desenvolvimento de uma doença específica. Essa informação, pode então, ser usada para identificar compostos que podem ser eficazes no tratamento da doença.

Como descrito por Ahmed e Jeon (2022), um dos principais objetivos das análises genômicas é desenvolver medicamentos genômicos, fomentando a descrição e interpretação de doenças e medicamentos associados a modificações genéticas. Corroborando com esse objetivo, foram realizadas investidas a fim de combinar dados de sequências de nível populacional e dados

ISSN: 2448-0916

genômicos com conhecimento fenotípico, relatórios clínicos e outros tipos de conjuntos de dados multi-ômicos (por exemplo, transcriptômica, proteômica e metabolômica) (Raza, 2023). Dessa forma, foram criadas demandas científicas e computacionais complexas para os pesquisadores, especialistas da saúde e serviços clínicos (Ahmed e Jeon, 2022). Surgindo, assim, a necessidade de técnicas computacionais que possibilitem a avaliação de dados heterogêneos e de alta dimensão e sistemas que forneçam soluções experimentais e analíticas mais fáceis, baratas, extensíveis e abrangentes (Ahmed & jeon, 2022).

Dando suporte às necessidades descritas acima, o emprego da mineração de dados e de tecnologias correlacionadas, abrangendo análise de dados, aprendizado de máquina e aprendizado profundo podem acelerar o processo de identificação de insights e ainda levar a uma melhor resposta as demandas científicas e computacionais. A inteligência artificial tem se dedicado a desenvolvimentos incrementais significativos na análise do genoma clínico, como fenotipagem em síndromes raras e câncer. Embora, principalmente, o trabalho baseado em inteligência artificial no campo da genômica esteja na fase de pesquisa, a demanda de técnicas de aprendizado de máquina e aprendizado profundo para análise de genômica funcional está aumentando. Agora, as melhorias na computação, na inteligência artificial e no aumento dos conjuntos de dados biomédicos permitem avanços nos campos de serviço existentes. Simultaneamente, esses avanços na pesquisa de acesso aberto e nas ferramentas de código aberto tornam próspero o uso da inteligência artificial em vários tipos de estudos de genoma (Ahmed & Jeon, 2022).

Fundamentado nas necessidades descritas acima e no avanço da inteligência artificial na análise da sequência do genoma, os pesquisadores Ahmed e Jeon (2022), tem como objetivo em seu trabalho apresentar um sistema baseado em inteligência artificial para análise da sequência do genoma do SARS-CoV e vírus semelhantes, incluindo SARS, MERS e Ebola. Realizamos análises comparativas para estudar os padrões básicos da sequência do genoma destes vírus e utilizamos ainda um algoritmo de aprendizado de máquina para classificação (Ahmed & Jeon, 2022).

Para análise comparativa das sequências do genoma, os autores utilizaram o Biopython. A análise da sequência do genoma frequentemente se tornou uma ferramenta essencial para o estudo de surtos de doenças. O SARS-CoV e outros vírus, incluindo genomas de SARS, MERS e Ebola, foram utilizados neste estudo. Os autores iniciaram as suas análises lendo a sequência de DNA; ao fazer isso, extraíram as informações de nucleotídeos ou o comprimento da sequência do genoma, o comprimento da sequência do genoma do SARS-CoV é 29.903, a sequência do genoma

da SARS é 2975, a sequência do genoma do MERS é 30.119 e o comprimento da sequência do genoma do Ebola é 18.959 (Ahmed & Jeon, 2022).

Para a classificação de diferentes sequências genômicas, Ahmed e Jeon (2022), utilizaram o SVM linear. O conjunto de dados era formado por sequências genômicas de todos os quatro tipos de vírus, de antemão, realizaram o pré-processamento, atribuindo os rótulos de classe manualmente e extraímos recursos úteis fornecidos aos classificadores SVM. Posteriormente, os dados foram divididos aleatoriamente, as sequências de codificação do genoma coletadas em amostras de treinamento e teste na proporção de 80% e 20%, respectivamente (Ahmed e Jeon, 2022). Por fim, os pesquisadores calcularam a *Accuracy, Recall, Pracision* e *F1-score*, o classificador SVM alcançou bons resultados *Accuracy* de classificação para todos os tipos de sequências genômicas, incluindo COVID-19 com 97%, SARS com 96%, MERS e Ebola com 95%, respectivamente. O P*recision*, *Recall* e *F1-Score* é 96, 77%, 96% para COVID-19, 96%, 74% e 96% para SARS, 95%, 74% e 95% para MERS e 95%, 74% e 95 % para o Ebola, respectivamente (Ahmed & Jeon, 2022).

Os resultados da técnica aplicada pelos autores fornecem subsídio para compreendermos a origem, o comportamento e a estrutura do vírus, o que pode auxiliar nos desenvolvimentos de vacinas, medicamentos antivirais e estratégias preventivas eficientes.

Kumar *et al* (2023), contextualiza que vivemos em um mundo no qual são gerados uma quantidade massiva de dados em quase todos os setores. Para que possamos aproveitar de forma mais eficiente esses dados, a IA nos fornece recursos para realizarmos o manuseio desses dados. Contextualizando a IA dentro das ciências farmacêuticas, a IA fornece a melhor abordagem para um sistema de saúde melhor. Para que isso se torne realidade, a IA e ML necessitam de uma grande quantidade de dados, e a maioria dos setores farmacêuticos e de saúde possui dados extensos. Hemingway *et al* (2018), comenta que a inteligência artificial e o machine learning auxiliam a utilizar os recursos nas melhores opções para médicos, consumidores, seguradoras e reguladores. Para tomar essas decisões, os dados são gerados de diferentes fontes, como domínio acadêmico, organizações e pesquisa e desenvolvimento, unidades industriais, farmácias e comunitárias. Adicionalmente, o uso da IA e ML, já estão sendo utilizadas nas indústrias farmacêuticas em diversos setores, como a descoberta de novos fármacos, avaliação de compostos ativos, realocação de fármacos, diagnóstico de doenças, ensaios clínicos e registro eletrônico de saúde (KUMAR *et al.*, 2023).

ISSN: 2448-0916

O desenvolvimento de uma nova molécula demora aproximadamente 13,5 anos para chegar ao ponto de aprovação. Em relação aos custos de pesquisa e desenvolvimento, são estimados US$2,6 bilhões (DIMASI *et al.*, 2018). Nesse contexto, o desenvolvimento da IA tem um grande impacto na descoberta e no desenvolvimento de novos fármacos, trazendo vantagens como: diminuição no tempo em protocolos que tomariam muito tempo sem a utilização da IA, através da melhor utilização dos recursos disponíveis (Kumar *et al.*, 2023).

Se a mente e a máquina trabalharem juntas, os processos de design de drogas, síntese química e análise de testes biológicos seriam mais eficientes. Há quatro etapas principais no processo de construção de um modelo de IA para a descoberta de medicamentos. A primeira é definir o problema, que deve ser específico ou abrangente. Em seguida, escolher o algoritmo de IA adequado e definir os valores iniciais para hiper parâmetros para fornecer a arquitetura de IA adequada. Em terceiro lugar, os dados de entrada devem ser preparados de forma satisfatória em termos de qualidade, quantidade, característica, por fim, o modelo é selecionado para treinamento e avaliação. Isso inclui algoritmos de treinamento, estratégias de otimização, mecanismos de avaliação e algoritmos de ajuste de hiper parâmetros (Kumar *et al.*, 2023).

Vários modelos de algoritmos comuns podem ser usados no primeiro e segundo estágios do modelo de construção. Esses modelos incluem floresta aleatória, rede neural artificial, máquina de Boltzmann profunda, rede de crença profunda, rede adversária generativa, auto codificador variacional, auto codificador adversário, aprendizagem simbólica e meta-aprendizagem. Devido à sua utilização em relações estrutura-atividade quantitativa e triagem virtual, a rede neural artificial tornou-se um dos modelos de dados não lineares mais poderosos. Isso aumentou significativamente nos últimos vinte anos. Ao mesmo tempo, a técnica de rede adversária generativa ajudou na química medicinal, criando modelos moleculares e na bioquímica, criando peptídeos e proteínas, reduzindo a dimensão dos dados de células únicas no desenvolvimento pré-clínico (Kumar *et al.*, 2023)

Kumar *et al* (2023), concluíram que o uso da inteligência artificial no sistema de saúde aumentou gradualmente, abrangendo uma gama de aplicações nos campos da farmacologia. Ressaltando que as tecnologias de IA são utilizadas em todas as etapas do processo de desenvolvimento de fármacos, o que evita riscos associados aos ensaios pré-clínicos e clínicos, também reduzindo significativamente os custos. Outro ponto destacado pelos autores, é o potencial de melhorar o atendimento ao paciente, auxiliando no diagnóstico de doenças. Ressaltando algumas limitações da IA, os autores descrevem sobre a alto custo e violação de

ISSN: 2448-0916

segurança em relação à privacidade dos dados. Também comentam sobre a questão de que a IA não pode compensar o estudo in vivo no processo de descoberta de medicamentos. Evidenciando que experimentos in vivo são necessários no processo de desenvolvimento de medicamentos, para confirmar a segurança e eficácia dos fármacos (Kumar *et al.,* 2023).

2.7 POSSIVEL UTILIZAÇÃO DA IA NO CONTEXTO DO BRASIL

A pandemia de COVID-19 trouxe uma série de desafios para o sistema de saúde brasileiro, especialmente na Atenção Primária à Saúde (APS) e na Estratégia Saúde da Família (ESF) (De Lira et al., 2024; Simolini et al., 2024). Dentro desses contextos, podemos visualizar a utilização da inteligência artificial, podendo exercer um papel fundamental na superação desses desafios, ajudando a otimizar processos, melhorar a promoção de saúde e enfrentar a disseminação de desinformação.

Conforme discutido por De Lira e colaboradores (2024), durante a crise, as equipes da ESF foram essenciais nas iniciativas de promoção da saúde, focando em medidas como vacinação, higiene e distanciamento social. Sob essa ótica, a IA pode ampliar o alcance dessas ações, facilitando o acesso a informações e adaptando o conteúdo de saúde para públicos diferentes. Uma das utilizações que podemos descrever é que, algoritmos de IA poderiam ter sido utilizados para criar campanhas de conscientização mais eficazes, ajustadas às características específicas das comunidades atendidas, o que aumenta o impacto dessas medidas preventivas.

A pandemia também forçou uma rápida reorganização do trabalho dos profissionais de saúde, que precisaram lidar com uma grande quantidade de casos suspeitos e confirmados, ao mesmo tempo que enfrentavam a falta de Equipamentos de Proteção Individual (EPIs) (De Lira et al, 2024). Dentro desses cenários, também é possível identificar uma possível utilização da IA, onde a utilização de algoritmos de triagem de pacientes, identificando rapidamente aqueles que têm maior risco de complicações graves. Além disso, sistemas inteligentes podem otimizar a distribuição de recursos, garantindo que as áreas mais necessitadas recebam EPIs e suprimentos de forma eficiente.

Simolini e colaboradores (2024), discorrem que, além dos desafios físicos, a saúde mental também foi seriamente afetada durante a pandemia, especialmente entre os estudantes universitários, que enfrentaram altos níveis de ansiedade, depressão e estresse. Nesse âmbito, a IA pode ser utilizada como ferramenta para identificar estudantes em risco, analisando dados de saúde mental em grande escala e permitindo intervenções mais personalizadas e eficazes, melhorando

ISSN: 2448-0916

assim a saúde mental dos estudantes e realizando um monitoramento ativo desse público, possibilitando que sejam realizadas interversões que possibilitem evitar o adoecimento psíquico.

Nesta seção, realizamos uma breve discussão de como a IA pode ser, demonstrando uma parte do potencial da sua utilização, especialmente na triagem de pacientes, promoção da saúde e promoção de saúde mental. Ainda existem muitas oportunidades para expandir o uso dessas tecnologias, principalmente na personalização dos cuidados e no suporte à saúde mental, preparando o país para lidar melhor com crises de saúde futuras.

3. CONSIDERAÇÕES FINAIS

A inteligência artificial tem revolucionado a forma como realizamos pesquisas. Ela pode auxiliar no desenvolvimento de novos fármacos, na triagem de pacientes em hospitais e na projeção de casos, contribuindo assim para a redução dos custos de saúde e indicando os melhores caminhos a serem seguidos. Embora a IA tenha o potencial de melhorar o atendimento ao paciente e auxiliar no diagnóstico, ainda enfrentamos limitações significativas relacionadas ao seu uso, tanto na qualidade dos dados utilizados para o treinamento quanto nos custos envolvidos.

Durante a pandemia, a IA se consolidou como uma ferramenta indispensável não apenas na pesquisa e desenvolvimento farmacêutico, mas também em diversas outras áreas. Ela foi crucial na análise rápida de grandes volumes de dados epidemiológicos, ajudando a prever surtos e a alocar recursos de saúde de maneira mais eficaz. Além disso, a tecnologia facilitou o desenvolvimento acelerado de vacinas e tratamentos, otimizando ensaios clínicos e identificando potenciais candidatos a medicamentos em tempo recorde. A inteligência artificial também melhorou a triagem de pacientes, permitindo diagnósticos mais rápidos e precisos, e auxiliou na gestão de cadeias de suprimentos, garantindo a distribuição eficiente de equipamentos médicos e vacinas. Assim, a IA ofereceu soluções inovadoras e eficientes para os desafios contemporâneos, demonstrando seu valor inestimável em tempos de crise.

No entanto, é importante reconhecer algumas limitações desta pesquisa. Apesar de a revisão da literatura ter abrangido um amplo conjunto de estudos, a diversidade e a complexidade dos métodos de IA aplicados à COVID-19 indicam que ainda existem lacunas consideráveis a serem preenchidas, sobretudo no que diz respeito à aplicação prática e à adaptação desses modelos para diferentes contextos geográficos e demográficos. Além disso, muitos dos estudos disponíveis se baseiam em dados de alta qualidade e fácil acesso, o que não reflete a realidade de muitos países

ISSN: 2448-0916

ou regiões com infraestrutura precária. A variedade dos dados usados nesses modelos, especialmente os biomédicos, também impõe desafios, já que eles variam significativamente em termos de qualidade e formato.

Como sugestões para estudos futuros trazemos a necessidade de criar modelos mais robustos e flexíveis, que possam lidar com dados mais diversos e de menor qualidade, especialmente em áreas com infraestrutura limitada. A inclusão de novas fontes de dados, como informações sobre saúde mental e condições socioeconômicas, pode aumentar a precisão e relevância dos modelos de IA na previsão e gestão de futuros surtos. Além disso, é fundamental continuar aprimorando algoritmos de aprendizado profundo e outras técnicas avançadas, com o intuito de minimizar os vieses nos dados e garantir que os modelos sejam mais justos e aplicáveis em diferentes realidades.

REFERÊNCIAS

AHMED, I.; JEON, G. Enabling Artificial Intelligence for Genome Sequence Analysis of COVID-19 and Alike Viruses. **Interdisciplinary Sciences: Computational Life Sciences**, 6 ago. 2021.

AHOUZ, F.; GOLABPOUR, A. Predicting the incidence of COVID-19 using data mining. **BMC Public Health**, v. 21, n. 1, 7 jun. 2021.

AMOUTZIAS, G. D. *et al.* The Remarkable Evolutionary Plasticity of Coronaviruses by Mutation and Recombination: Insights for the COVID-19 Pandemic and the Future Evolutionary Paths of SARS-CoV-2. **Viruses**, v. 14, n. 1, p. 78, 2 jan. 2022.

BEWICK, V.; CHEEK, L.; BALL, J. Statistics review 13: Receiver operating characteristic curves. **Critical Care**, v. 8, n. 6, p. 508, 2004.

CAUDAI, C. *et al.* AI applications in functional genomics. **Computational and Structural Biotechnology Journal**, v. 19, p. 5762–5790, 2021.

DALL'ALBA, G. et al. A Survey of Biological Data in a Big Data Perspective. **Big Data**, 7 abr. 2022.

DE LIRA, J. M. et al. Promoção da Saúde na Pandemia: impacto da Covid-19 na prática da Enfermagem na ESF. **Cadernos Cajuína,** v. 9, n. 4, p. e249424–e249424, 8 ago. 2024.

DE WIT, E. *et al.* SARS and MERS: recent insights into emerging coronaviruses. **Nature Reviews Microbiology**, v. 14, n. 8, p. 523–534, 27 jun. 2016.

DIMASI, J. A.; GRABOWSKI, H. G.; HANSEN, R. W. Innovation in the pharmaceutical industry: New estimates of R&D costs. **Journal of health economics**, v. 47, n. 47, p. 20–33, maio 2016.

ECDC. **Factsheet on COVID-19**. Disponível em: <https://www.ecdc.europa.eu/en/infectious-disease-topics/z-disease-list/covid-19/factsheet-covid-19>.

ENVALL, M. On the difference between mono-, holo-, and paraphyletic groups: a consistent distinction of process and pattern. **Biological Journal of the Linnean Society**, v. 94, n. 1, p. 217–220, 29 abr. 2008.

FAZOLO, T. et al. Pediatric COVID-19 patients in South Brazil show abundant viral mRNA and strong specific anti-viral responses. **Nature Communications**, v. 12, n. 1, p. 6844, 25 nov. 2021.

FORCHETTE, L.; SEBASTIAN, W.; LIU, T. A Comprehensive Review of COVID-19 Virology, Vaccines, Variants, and Therapeutics. **Current Medical Science**, v. 41, n. 6, 9 jul. 2021.

GNS, H. S. *et al.* An update on Drug Repurposing: Re-written saga of the drug's fate. **Biomedicine & Pharmacotherapy**, v. 110, p. 700–716, fev. 2019.

GOODMAN, B.; FLAXMAN, S. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". **AI Magazine**, v. 38, n. 3, p. 50–57, 2 out. 2017.

GORBALENYA, A. E. *et al.* The Species Severe Acute Respiratory syndrome-related Coronavirus: Classifying 2019-nCoV and Naming It SARS-CoV-2. **Nature Microbiology**, v. 5, n. 4, p. 1–9, 2 mar. 2020.

HALLAL, P. C. *et al.* SARS-CoV-2 antibody prevalence in Brazil: results from two successive nationwide serological household surveys. **The Lancet Global Health**, v. 8, n. 11, p. e1390–e1398, nov. 2020.

Handbook of Statistical Analysis and Data Mining Applications. Disponível em: <https://books.google.com.br/books?hl=en&lr=&id=U5np34a5fmQC&oi=fnd&pg=PP1&dq=Nisbet>. Acesso em: 3 set. 2024.

HELMY, Y. A. *et al.* The COVID-19 Pandemic: A Comprehensive Review of Taxonomy, Genetics, Epidemiology, Diagnosis, Treatment, and Control. **Journal of Clinical Medicine**, v. 9, n. 4, p. 1225, 24 abr. 2020.

HEMINGWAY, H. *et al.* Big data from electronic health records for early and late translational cardiovascular research: challenges and potential. **European Heart Journal**, v. 39, n. 16, p. 1481–1495, 29 ago. 2017.

Introduction to Artificial Intelligence. Disponível em: <https://books.google.com.br/books?hl=en&lr=&id=geFHDwAAQBAJ&oi=fnd&pg=PR5&dq=ERTEL>. Acesso em: 3 set. 2024.

JAHANAFROOZ, Z. *et al.* An overview of human proteins and genes involved in SARS-CoV-2 infection. **Gene**, v. 808, p. 145963, jan. 2022.

JOSHI, A. V. Machine Learning and Artificial Intelligence. [s.n.].

KADAM, S. B. *et al.* SARS-CoV-2, the pandemic coronavirus: Molecular and structural insights. **Journal of Basic Microbiology**, v. 61, n. 3, p. 180–202, 18 jan. 2021.

KARUNAKARAN, V. *et al.* A non-invasive ultrasensitive diagnostic approach for COVID-19 infection using salivary label-free SERS fingerprinting and artificial intelligence. **Journal of Photochemistry and Photobiology B-biology**, v. 234, p. 112545–112545, 1 set. 2022.

KIRTIPAL, N.; BHARADWAJ, S.; KANG, S. G. From SARS to SARS-CoV-2, insights on structure, pathogenicity and immunity aspects of pandemic human coronaviruses. **Infection, Genetics and Evolution**, v. 85, p. 104502, ago. 2020.

KOHAVI, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. [s.n.]. Disponível em: <https://core.ac.uk/download/pdf/186743801.pdf>.

KUMAR, M. *et al.* Opportunities and challenges in application of artificial intelligence in pharmacology. **Pharmacological Reports**, v. 75, n. 1, p. 3–18, 9 jan. 2023.

LEVI, M. *et al.* Coagulation abnormalities and thrombosis in patients with COVID-19. **The Lancet Haematology**, v. 7, n. 6, p. e438–e440, jun. 2020.

LIANG, W. et al. Early triage of critically ill COVID-19 patients using deep learning. **Nature Communications**, v. 11, n. 1, p. 3543, 15 jul. 2020.

LIPSITCH, M.; SWERDLOW, D. L.; FINELLI, L. Defining the Epidemiology of Covid-19 — Studies Needed. **New England Journal of Medicine**, v. 382, p. 1194–1196, 19 fev. 2020.

LU, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. **The Lancet**, v. 395, n. 10224, p. 565–574, jan. 2020.

MARTINEZ, G. *et al.* An artificial neural network classification method employing longitudinally monitored immune biomarkers to predict the clinical outcome of critically ill COVID-19 patients. **PeerJ**, v. 10, p. e14487, 2022.

METZ, C. E. Basic principles of ROC analysis. **Seminars in Nuclear Medicine**, v. 8, n. 4, p. 283–298, out. 1978.

MOHANTY, S. *et al.* Application of Artificial Intelligence in COVID-19 drug repurposing. **Diabetes & Metabolic Syndrome: Clinical Research & Reviews**, v. 14, n. 5, p. 1027–1031, set. 2020.

MORAES, P. **Compressão de modelos em transferência de aprendizado de máquina**. 11 fev. 2022.

MUELLER, J. P.; MASSARON, L. **Aprendizado de Máquina por John Paul Mueller e Luca Massaron**. [s.l: s.n.]. Disponível em: <https://altabooks.com.br/wp-content/uploads/2019/11/AMOSTRA_AprendizadoMaquinaPL.pdf>.

MURALIDAR, S. *et al.* The emergence of COVID-19 as a global pandemic: Understanding the epidemiology, immune response and potential therapeutic targets of SARS-CoV-2. **Biochimie**, v. 179, n. 179, p. 85–100, dez. 2020.

NAVECA, F. G. *et al.* Spread of Gamma (P.1) Sub-Lineages Carrying Spike Mutations Close to the Furin Cleavage Site and Deletions in the N-Terminal Domain Drives Ongoing Transmission of SARS-CoV-2 in Amazonas, Brazil. **Microbiology Spectrum**, v. 10, n. 1, 23 fev. 2022.

NUNES, D.; MARQUES, A. **Inteligência artificial e Direito Processual: vieses algorítmicos e os riscos de atribuição de função decisória às máquinas = Artificial intelligence and Procedural Law: algorithmic bias and the risks of assignment of decision-making function to machines**. Disponível em: <https://oasisbr.ibict.br/vufind/Record/STJ-1_27ad28a4f877422391dcf59d93072996>. Acesso em: 3 set. 2024.

PENETRA, S. L. S. *et al.* Post-acute COVID-19 syndrome after reinfection and vaccine breakthrough by the SARS-CoV-2 Gamma variant in Brazil. **International Journal of Infectious Diseases**, v. 114, p. 58–61, jan. 2022.

PU, J. *et al.* Any unique image biomarkers associated with COVID-19? **European Radiology**, v. 30, n. 11, p. 6221–6227, 20 jul. 2020.

RANGARAJAN, K. *et al.* Artificial Intelligence–assisted chest X-ray assessment scheme for COVID-19. **European Radiology**, 20 jan. 2021.

RAZA, S. **Artificial intelligence for genomic medicine - PHG Foundation**. Disponível em: <https://www.phgfoundation.org/publications/reports/artificial-intelligence-for-genomic-medicine/>. Acesso em: 3 set. 2024.

ROTHE, C. *et al.* Transmission of 2019-nCoV Infection from an Asymptomatic Contact in Germany. **New England Journal of Medicine**, v. 382, n. 10, 30 jan. 2020.

RUSSELL, S.; NORVIG, P. **Artificial Intelligence A Modern Approach Third Edition**. [s.l: s.n.]. Disponível em: <https://people.engr.tamu.edu/guni/csce421/files/AI_Russell_Norvig.pdf>.

SANTOS, A. J. DOS *et al.* **Inteligência artificial e COVID-19. Construção de conhecimento no curso da pandemia de COVID-19**: aspectos biomédicos, clínico-assistenciais, epidemiológicos e sociais, 2020.

SARTOR, S. *et al.* Association between obesity and hospitalization in mild COVID-19 adult outpatients in Brazil: a prospective cohort study. **Archives of Endocrinology and Metabolism**, v. 66, n. 4, p. 512–521, 1 jan. 2022.

SHARMA, A.; AHMAD FAROUK, I.; LAL, S. K. COVID-19: A Review on the Novel Coronavirus Disease Evolution, Transmission, Detection, Control and Prevention. **Viruses**, v. 13, n. 2, p. 202, 29 jan. 2021.

SHIRI, I. et al. Ultra-low-dose chest CT imaging of COVID-19 patients using a deep residual neural network. **European Radiology**, v. 31, n. 3, p. 1420–1431, 3 set. 2020.

SIMOLINI, A. V. *et al.* FATORES DE RISCO E ADOECIMENTO PSÍQUICO EM ESTUDANTES UNIVERSITÁRIOS DURANTE A PANDEMIA POR CORONAVÍRUS. **Cadernos Cajuína**, v. 9, n. 2, p. e249223–e249223, 2 maio 2024. Disponível em: https://doi.org/10.52641/cadcajv9i2.222

STEHMAN, S. V. Selecting and interpreting measures of thematic classification accuracy. **Remote Sensing of Environment**, v. 62, n. 1, p. 77–89, out. 1997.

SYED, H. H. et al. A Rapid Artificial Intelligence-Based Computer-Aided Diagnosis System for COVID-19 Classification from CT Images. **Behavioural Neurology**, v. 2021, p. 1–13, 27 dez. 2021.

TAN, C. *et al.* A Survey on Deep Transfer Learning. **Artificial Neural Networks and Machine Learning – ICANN 2018**, p. 270–279, 2018.

VARGA, Z. *et al.* Endothelial cell infection and endotheliitis in COVID-19. **The Lancet**, v. 395, n. 10234, 21 abr. 2020.

WAN, Y. *et al.* Receptor recognition by novel coronavirus from Wuhan: An analysis based on decade-long structural studies of SARS. **Journal of Virology**, v. 94, n. 7, 29 jan. 2020.

WANG, H. *et al.* Decoding COVID-19 pneumonia: comparison of deep learning and radiomics CT image signatures. **European Journal of Nuclear Medicine and Molecular Imaging**, v. 48, n. 5, p. 1478–1486, 1 maio 2021.

WEISS, S.; NAVAS-MARTIN, S. **Coronavirus Pathogenesis and the Emerging Pathogen Severe Acute Respiratory Syndrome Coronavirus**. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/16339739/>.

WIERSINGA, W. J. *et al.* Pathophysiology, Transmission, Diagnosis, and Treatment of Coronavirus Disease 2019 (COVID-19): A Review. **JAMA**, v. 324, n. 8, p. 782–793, 10 jul. 2020.

WORLD HEALTH ORGANIZATION. **Coronavirus Disease (COVID-19) Pandemic**. Disponível em: <https://www.who.int/emergencies/diseases/novel-coronavirus-2019>.

WU, F. et al. A new coronavirus associated with human respiratory disease in China. **Nature**, v. 579, n. 7798, p. 265–269, 3 fev. 2020.

ZHU, N. *et al.* A Novel Coronavirus from Patients with Pneumonia in China, 2019. **New England Journal of Medicine**, v. 382, n. 8, 24 jan. 2020.2, 1987.