PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS

Programa de Pós-Graduação em Biologia de Vertebrados

Paula Vitória Silva

O USO DE APRENDIZADO DE MÁQUINA NA PREVISÃO DE MAMÍFEROS TRANSMISSORES DE SARS-COV-2

Belo Horizonte

2024

Paula Vitória Silva

O USO DE APRENDIZADO DE MÁQUINA NA PREVISÃO DE MAMÍFEROS TRANSMISSORES DE SARS-COV-2

Dissertação apresentada ao Programa de Pós-Graduação em Biologia de Vertebrados da Pontifícia Universidade Católica de Minas Gerais, como requisito parcial para obtenção do título de Mestre em Biologia de Vertebrados.

Orientador: Prof.ª Dr.ª Cristiane Neri Nobre

Belo Horizonte

2024

FICHA CATALOGRÁFICA
Elaborada pela Biblioteca da Pontifícia Universidade Católica de Minas Gerais

Silva, Paula Vitória
S586u O uso de aprendizado de máquina na previsão de mamíferos transmissores de SARS-CoV-2 / Paula Vitória Silva. Belo Horizonte, 2024.
45 f. : il.

Orientadora: Cristiane Neri Nobre
Dissertação (Mestrado) - Pontifícia Universidade Católica de Minas Gerais. Programa de Pós-Graduação em Biologia de Vertebrados

1. Bioinformática. 2. COVID-19 (Doença). 3. SARS-CoV-2. 4. Aprendizado do computador. 5. Mamífero. 6. Transmissão de doença. 7. Doenças respiratórias. I. Nobre, Cristiane Neri. II. Pontifícia Universidade Católica de Minas Gerais. Programa de Pós-Graduação em Biologia de Vertebrados. III. Título.

CDU: 577.4
Ficha catalográfica elaborada por Fabiana Marques de Souza e Silva - CRB 6/2086

Paula Vitória Silva

O USO DE APRENDIZADO DE MÁQUINA NA PREVISÃO DE MAMÍFEROS TRANSMISSORES DE SARS-COV-2

Dissertação apresentada ao Programa de Pós-Graduação em Biologia de Vertebrados como requisito parcial para qualificação ao Grau de Mestre em Biologia de Vertebrados pela Pontifícia Universidade Católica de Minas Gerais.

Prof.ª Cristiane Neri Nobre – PUC Minas

Prof.ª Milene Barbosa Carvalho - UFSJ

Prof. Lucas Bleicher - UFMG

Belo Horizonte, 30 de Agosto de 2024

AGRADECIMENTOS

À Pontifícia Universidade Católica de Minas Gerais (PUC Minas) pela concessão da bolsa de estudos;

À minha orientadora, Cristiane, pelos ensinamentos e toda dedicação a este trabalho;

Aos meus pais, Angela e Jefferson, pelo incentivo em todos os momentos.

RESUMO

O SARS-CoV-2 é um vírus de RNA envelopado que causa graves doenças respiratórias em humanos e animais. A infecção ocorre quando a proteína Spike se liga à enzima conversora de angiotensina 2 (ACE2) do hospedeiro. Os morcegos são considerados os hospedeiros naturais do vírus, e a transmissão zoonótica é um risco significativo quando humanos entram em contato próximo com animais infectados. Portanto, compreender a interconexão entre a saúde humana, animal e ambiental é crucial para prevenir e controlar futuros surtos de coronavírus. Este trabalho visou revisar sistematicamente a literatura para identificar características que tornam os mamíferos transmissores adequados do vírus, levantar os principais métodos computacionais utilizados para avaliar o SARS-CoV-2 em mamíferos e identificar quais características tornam um mamífero bom transmissor deste vírus. Conseguinte, foram utilizados dados ecológicos, de história de vida e características biológicas de várias espécies de mamíferos, além de oito diferentes algoritmos de Aprendizado de Máquina (Naive Bayes, Decision Tree, Random Forest, XGBoost, AdaBoost, SVM, Regressão Logística e MLP) para prever a suscetibilidade das espécies ao SARS-CoV-2. A rede neural foi identificada como o melhor modelo devido à sua performance. Além disso, foi empregado o método CSSE (Agnostic Method of Counterfactual, Selected, and Social Explanations) para interpretar os resultados e identificar as características biológicas mais relevantes na suscetibilidade ao vírus. A análise contrafactual indicou que primatas apresentam alta suscetibilidade ao SARS-CoV-2. Fatores como densidade populacional, longevidade, tamanho do grupo social, frequência de ninhadas e atividade crepuscular foram identificados como determinantes na capacidade de uma espécie atuar como hospedeiro zoonótico. Os achados contribuem para a prevenção e controle de futuros surtos, fornecendo informações sobre fatores de transmissão e destacando a importância de métodos computacionais avançados no estudo de doenças infecciosas. Isso permite uma compreensão mais profunda dos padrões de transmissão e pode ajudar no desenvolvimento de estratégias de controle e intervenção mais eficazes. A pesquisa oferece uma ferramenta prática para identificar e monitorar potenciais hospedeiros de SARS-CoV-2 e outros patógenos emergentes.

Palavras-chave: Bioinformática. COVID-19. SARS-CoV-2. Aprendizado de máquina. Mamíferos.

ABSTRACT

SARS-CoV-2 is an enveloped RNA virus that causes severe respiratory illness in humans and animals. Infection occurs when the Spike protein is bound to the host's angiotensin-converting enzyme 2 (ACE2). Bats are considered the natural hosts of the virus, and zoonotic transmission is a significant risk when humans come into close contact with infected animals. Therefore, understanding the interconnection between human, animal, and environmental health is crucial to preventing and controlling future coronavirus outbreaks. This work aims to systematically review the literature to identify characteristics that make mammals suitable virus transmitters and identify the main computational methods used to evaluate SARS-CoV-2 in mammals and identify which characteristics make mammals good transmitters of this virus. Therefore, ecological data, life history, and biological characteristics of several species of mammals were used, in addition to eight different Machine Learning algorithms (Naive Bayes, Decision Tree, Random Forest, XGBoost, AdaBoost, SVM, Logistic Regression, and MLP) to Predict species susceptibility to SARS-CoV-2. The neural network was identified as the best model due to its performance. Furthermore, the CSSE method (Agnostic Method of Counterfactual, Selected, and Social Explanations) was implemented to interpret the results and identify the most relevant biological characteristics in susceptibility to the virus. A counterfactual analysis indicated that primates are highly susceptible to SARS-CoV-2. Population density, longevity, social group size, litter frequency, and twilight activity determine a species' ability to act as a zoonotic host. The findings of this research are not just theoretical, but also have practical implications for the prevention and control of future outbreaks. They provide crucial information on transmission factors and underscore the importance of advanced computational methods in the study of infectious diseases. This deeper understanding of transmission patterns can help develop more effective control and intervention strategies. The research offers a practical tool for identifying and monitoring potential hosts of SARS-CoV-2 and other emerging pathogens, thereby empowering the scientific community with the necessary knowledge and tools.

Keywords: Bioinformatics. COVID-19. SARS-CoV-2. Machine Learning. Mammals.

LISTA DE ABREVIATURAS E SIGLAS

ACE2 Enzima Conversira de Angiotensina 2

AM Aprendizado de Máquina

COVID-19 *Coronavirus Disease 2019*

CoV Coronavírus

MERS-CoV Coronavírus da Síndrome Respiratória do Oriente Médio

ML *Machine Learning*

OMS Organização Mundial da Saúde

RBD Domínio de Ligação ao receptor

S *Spike*

RNA Ácido Ribonucleico

SARS-CoV Coronavírus da Síndrome Respiratória Aguda Grave

SARS-CoV-2 Coronavírus da Síndrome Respiratória Aguda Grave 2

SUMÁRIO

1 INTRODUÇÃO

Os coronavírus (CoVs) são vírus envelopados, de RNA de sentido positivo (EVANS; LIU, 2021), classificados em quatro gêneros: *alfa* , *beta*, *gama* e *delta*. Os *alfacoronavírus* e *betacoronavírus* infectam principalmente mamíferos, enquanto os *gamacoronavírus* e *deltacoronavírus* infectam predominantemente aves (TIWARI et al., 2020). Esses vírus podem causar uma variedade de doenças em humanos e animais, que variam de infecções respiratórias leves a doenças graves. Nos últimos 20 anos, estes vírus causaram três grandes surtos de doenças respiratórias graves: Coronavírus da Síndrome Respiratória Aguda Grave (SARS-CoV), Coronavírus da Síndrome Respiratória do Oriente Médio (MERS-CoV) (ARORA et al., 2020), e o mais recente, Síndrome Respiratória Aguda Coronavírus Grave 2 (SARS-CoV-2), foi declarado pandêmico em 11 de março de 2020 (MUNIR et al., 2020).

O coronavírus da Síndrome Respiratória Aguda Grave (SARS-CoV) surgiu em novembro de 2002 em Guangdong, China, causando doenças respiratórias graves (LEE et al., 2003). Acredita-se que o vírus tenha se originado em morcegos-ferradura e sido transmitido aos humanos através de hospedeiros intermediários, como os civetas e cães-guaxinim. O surto de SARS resultou em mais de 8.000 casos e aproximadamente 800 mortes até meados de 2003 (KUIKEN et al., 2003).

O coronavírus da Síndrome Respiratória do Oriente Médio (MERS-CoV) foi identificado pela primeira vez na Arábia Saudita em 2012. Este vírus tem uma taxa de mortalidade elevada, em torno de 34,5% (ZHOU et al., 2023). Acredita-se que o MERS-CoV tenha se originado em morcegos e se espalhado para humanos através de camelos dromedários, que atuam como hospedeiros intermediários (MEYER et al., 2014). O vírus causou doenças respiratórias graves e falência renal nos indivíduos infectados (ZAKI et al., 2012).

Em dezembro de 2019, um novo coronavírus, o SARS-CoV-2, foi identificado em Wuhan, China, levando à pandemia de COVID-19 (HUANG et al., 2020). Este vírus causou uma crise de saúde global, com ampla transmissão e doenças respiratórias graves. Em 11 de março de 2020, a Organização Mundial da Saúde (OMS) declarou a COVID-19 uma pandemia. Sua rápida propagação se deve, em parte, à sua notável transmissibilidade e ao período de incubação assintomática, que dificulta a identificação precoce e controle de novos casos (MURATA et al., 2021).

O SARS-CoV-2 infecta as células hospedeiras ao ligar sua proteína Spike (S) aos

receptores da enzima conversora de angiotensina 2 (ACE2) na superfície celular (YOO; YOO, 2020). Essa ligação facilita a entrada do vírus na célula hospedeira, iniciando o processo de infecção.

A ACE2 é uma enzima de membrana expressa em diversos tecidos humanos, incluindo o pulmonar, cardíaco, renal e intestinal (TROUGAKOS et al., 2021). A presença ampla de receptores ACE2 em diferentes tecidos os torna alvos chave para o SARS-CoV-2, levando a manifestações sistêmicas da COVID-19.

A transmissão interespécies do SARS-CoV-2 foi observada, com o vírus infectando diversos animais, como gatos, cães, visons, tigres e leões. Esse transbordamento zoonótico destaca o potencial de humanos transmitirem o vírus para animais, representando riscos tanto para a vida selvagem quanto para animais domésticos (BAO et al., 2020; EGEREN et al., 2021; VOLZ et al., 2021).

A abordagem "One Health" enfatiza a interconexão entre a saúde humana, animal e ambiental (FUENTE; MERA; GORTÁZAR, 2021). Essa estratégia multidisciplinar é vital para compreender e controlar surtos de doenças infecciosas. Estudar a transmissão do vírus entre espécies é essencial para desenvolver intervenções eficazes de saúde pública e prevenir futuras pandemias (KORATH et al., 2022).

Recentes avanços na análise computacional aprimoraram significativamente nossa compreensão do SARS-CoV-2. Modelagem molecular, bioinformática e sequenciamento genômico têm sido instrumentais na identificação de alvos terapêuticos, monitoramento de mutações do vírus e previsão de possíveis novos surtos.

Métodos de Aprendizado de Máquina (AM) oferecem soluções inovadoras para a análise de dados biológicos complexos. No contexto do SARS-CoV-2, algoritmos de AM podem acelerar a descoberta de interações vírus-hospedeiro, identificar alvos terapêuticos potenciais e prever mutações virais e seus impactos (MOLLENTZE et al., 2022).

Modelos de AM interpretáveis são cruciais para obter *insights* sobre processos preditivos. Explicações contrafactuais, que identificam mudanças mínimas necessárias para alterar uma previsão, proporcionam transparência e aumentam a confiança nos modelos de AM. Esses métodos podem ajudar a identificar características chave que tornam certas espécies suscetíveis ao SARS-CoV-2 (ELSHAWI et al., 2021).

Este estudo tem como objetivo identificar as características que tornam os mamíferos capazes de transmitir o SARS-CoV-2, revisar e aplicar os principais métodos computacionais utilizados para analisar dados sobre a doença em mamíferos. Ao integrar a análise computacional com a abordagem One Health, buscamos aprimorar a compreensão das interações vírus-hospedeiro e desenvolver medidas preventivas e terapêuticas eficazes contra a COVID-19.

1.1 Problema

Ainda não há uma lista definitiva de mamíferos transmissores do SARS-CoV-2, o que dificulta a implementação de medidas eficazes para prevenir possíveis novos surtos de coronavírus. Assim, buscamos responder ao seguinte questionamento, objeto de discussão deste trabalho: quais características fazem com que um mamífero seja bom transmissor de SARS-CoV-2?

1.2 Objetivos

1.2.1 Objetivo geral

Identificar quais características tornam mamíferos bons transmissores de SARS-CoV-2 através do uso de Aprendizado de Máquina.

1.2.2 Objetivo específico

Como objetivo específico deste trabalho, estão:

- Realizar uma revisão sistemática da literatura para identificar o que torna um mamífero um bom trasmissor de SARS-CoV-2, além de levantar que métodos de aprendizado de máquina têm sido utilizados neste contexto;
- Coletar dados de sequenciamento de mamíferos relacionados com o SARS-CoV-2 disponíveis em bancos de dados genéticos;
- Avaliar diferentes algoritmos de Aprendizado de Máquina para previsão de mamíferos transmissores.

1.3 Justificativa

A pandemia de COVID-19 evidenciou a necessidade de se compreender melhor a relação entre doenças infecciosas e seus hospedeiros animais. A identificação de mamíferos transmissores de SARS-CoV-2 é importante para o desenvolvimento de medidas preventivas e terapêuticas eficazes. No entanto, métodos convencionais de identificação desses animais são limitados e muitas vezes não conseguem detectar espécies relevantes.

Estudos recentes têm demonstrado que o aprendizado de máquina pode ser uma ferramenta poderosa para a previsão de doenças infecciosas em mamíferos (YANG et al., 2020; WARDEH et al., 2021; WARDEH; SHARKEY; BAYLIS, 2020). Além disso, a

genômica comparativa e a análise filogenética têm sido usadas para entender a evolução e a dispersão do SARS-CoV-2 em diferentes hospedeiros animais (BONI et al., 2020; SARDAR et al., 2020; SHI et al., 2020). No entanto, o uso de aprendizado de máquina para prever mamíferos transmissores de SARS-CoV-2 ainda é pouco explorado.

Há uma lacuna significativa no conhecimento sobre os mamíferos transmissores de SARS-CoV-2, especialmente em relação àqueles que podem atuar como reservatórios para a transmissão do vírus aos seres humanos (GRYSEELS et al., 2021). Além disso, métodos convencionais de identificação desses animais são limitados e muitas vezes dependem de dados incompletos ou de baixa qualidade. O uso de aprendizado de máquina pode ajudar a preencher essas lacunas e permitir a identificação de espécies relevantes de uma forma mais precisa e eficaz.

Espera-se que o trabalho proposto forneça uma ferramenta para a previsão de mamíferos transmissores de SARS-CoV-2, permitindo que medidas preventivas e terapêuticas eficazes sejam desenvolvidas mais rapidamente. Além disso, o estudo pode contribuir para o avanço da compreensão teórica sobre a relação entre doenças infecciosas e seus hospedeiros animais.

1.4 Organização do trabalho

A dissertação "O uso de Aprendizado de Máquina na Previsão de Mamíferos Transmissores de SARS-CoV-2" foi elaborada na forma de artigo científico como previsto nas normas do Programa de Pós-graduação em Biologia de Vertebrados da Pontifícia Universidade Católica de Minas Gerais.

O Capítulo 2 desta dissertação contém o primeiro artigo, intitulado "Computational methods in the analysis of SARS-CoV-2 in mammals: a systematic review of the literature" publicado na revista *Computers in Biology and Medicine* em maio de 2024.

Este primeiro artigo trata-se de uma revisão sistemática da literatura com o objetivo de identificar características que tornam os mamíferos transmissores de SARS-CoV-2 além de levantar os principais métodos computacionais utilizados para avaliação deste vírus em mamíferos.

O Capítulo 3 apresenta o segundo artigo oriundo do presente trabalho, intitulado "Machine learning to explain the zoonotic ability of mammals to transmit SARS-CoV-2", a ser submetido à revista a ser definida.

O segundo artigo explora a aplicação de técnicas de Aprendizado de Máquina para a investigação da interação entre o vírus SARS-CoV-2 e a proteína ACE2, com foco na ligação essencial para a infecção viral. Este artigo tem como objetivo revisar

os avanços recentes e explorar o uso de algoritmos de Aprendizado de Máquina neste contexto, destacando seu impacto no desenvolvimento de estratégias eficazes contra a COVID-19.

O Capítulo 4 encerra com as considerações finais e trabalhos futuros.

2 CAPÍTULO 2

Computers in Biology and Medicine 173 (2024) 108264

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Computational methods in the analysis of SARS-CoV-2 in mammals: A systematic review of the literature

Paula Vitória Silva, Cristiane N. Nobre *

Pontifical Catholic University of Minas Gerais - PUC Minas, 500 Dom José Gaspar Street, Building 41, Coração Eucarístico, Belo Horizonte, MG 30535-901, Brazil

ARTICLE INFO

Keywords:
Bioinformatics
COVID-19
SARS-CoV-2
Machine learning
Mammals

ABSTRACT

SARS-CoV-2 is an enveloped RNA virus that causes severe respiratory illness in humans and animals. It infects cells by binding the Spike protein to the host's angiotensin-converting enzyme 2 (ACE2). The bat is considered the natural host of the virus, and zoonotic transmission is a significant risk and can happen when humans come into close contact with infected animals. Therefore, understanding the interconnection between human, animal, and environmental health is important to prevent and control future coronavirus outbreaks. This work aimed to systematically review the literature to identify characteristics that make mammals suitable virus transmitters and raise the main computational methods used to evaluate SARS-CoV-2 in mammals. Based on this review, it was possible to identify the main factors related to transmissions mentioned in the literature, such as the expression of ACE2 and proximity to humans, in addition to identifying the computational methods used for its study, such as Machine Learning, Molecular Modeling, Computational Simulation, between others. The findings of the work contribute to the prevention and control of future outbreaks, provide information on transmission factors, and highlight the importance of advanced computational methods in the study of infectious diseases that allow a deeper understanding of transmission patterns and can help in the development of more effective control and intervention strategies.

1. Introduction

Coronaviruses (CoVs) are enveloped, positive-sense Ribonucleic Acid (RNA) viruses divided into four genera [1]: *alpha* and *beta-coronaviruses*, capable of infecting mammals, and *gamma* and *delta-coronaviruses* that circulate mainly among birds [2]. Over the past 20 years, these viruses have caused three major outbreaks of severe respiratory illness: Severe Acute Respiratory Syndrome coronavirus (SARS-CoV), Middle East Respiratory Syndrome coronavirus (MERS-CoV) [3], and the latest, Acute Respiratory Syndrome Coronavirus 2 Severe (SARS-CoV-2), was declared a pandemic on March 11, 2020 [4].

Severe Acute Respiratory Syndrome coronavirus (SARS-CoV) was detected in November 2002 in patients in Guangdong, China [5]. Its reservoir host is the horseshoe bat (*Rhinolophus ferrumequinum*). In addition, the virus has been detected in its amplifier hosts, civet cat (*Paguma larvata*) and raccoon dog (*Nyctereutes procyonoides*), and also in non-human primates (PNH), causing fever and respiratory distress in elderly animals [6][7]. Until July 2003, the virus had infected 8439 human patients, and 812 (9.6%) died [8].

The Middle East Respiratory Syndrome coronavirus (MERS-CoV) has the highest fatality rate among CoVs: 34.5% [9]. The virus was first identified in Saudi Arabia in 2012 in a patient with acute pneumonia and subsequent kidney failure [10]. Subsequently, MERS-CoV was isolated in camels (*Camelus dromedarius*), considered intermediate hosts [11], and in bats, which are the probable origin of this virus [12].

In mid-December 2019, Coronavirus Disease (COVID-19), caused by the SARS-CoV-2 virus, was first identified in patients with pneumonia in Wuhan, China [13]. According to Lorusso et al. [14], the patients were exposed to the wild animal market, which may be a probable origin of the infection. According to Wong et al. [15], the bat is considered the virus's natural host, but it is still impossible to say. Natural SARS-Cov-2 infection has been reported in cats, dogs, minks, tigers, and lions [16]. The frequency of human interaction with animals contributes to zoonotic spillover when the virus that does not naturally infect humans starts to do so [17].

SARS-Cov-2 infects cells by binding the Spike (S) protein to the host's angiotensin-converting enzyme 2 (ACE2) [18]. According to Tiwari et al. [2], the virus has an excellent capacity for mutations due to the instability of the replicase enzyme and the lack of a nucleotide revision mechanism.

The interconnection between human, animal, and environmental health is recognized as essential to prevent and control infectious

* Corresponding author.
E-mail addresses: paula.silva.1138274@sga.pucminas.br (P.V. Silva), nobre@pucminas.br (C.N. Nobre).

https://doi.org/10.1016/j.compbiomed.2024.108264
Received 28 June 2023; Received in revised form 15 February 2024; Accepted 6 March 2024
Available online 16 March 2024

disease outbreaks; this approach is known as One Health [19]. Studying the virus transmission between humans and animals has been fundamental for preventing new episodes and developing effective public health strategies [20]. The One Health approach is proving increasingly important in understanding the SARS-CoV-2 pandemic.

Advances in computational analysis have been instrumental in understanding the biology of SARS-CoV-2. Molecular modeling and bioinformatics have allowed the identification of potential therapeutic targets and the development of new drugs to treat the infection [21]. In addition, computational analysis of virus genomic sequences has been used to monitor the spread of the pandemic in real-time and try to predict possible new outbreaks of zoonotic diseases [22].

Several works present a variety of approaches related to the use of computational analysis in the prediction of mammals that transmit SARS-CoV-2. Damas et al. [23] used a comparative and structural analysis of ACE2 to predict the wide range of hosts that may be susceptible to the virus. Kumar et al. [24] work used sequence homology to predict the susceptibility of different animals to the virus. The studies by Lam et al. [25] and Liu et al. [26] used molecular modeling and comparative sequence analysis to predict the likelihood of infection in a wide range of animals. Melin et al. [27] investigated ACE2 variation in primates and the risk of COVID-19 in these animals. Most of these researches employ comparative and structural analyses involving ACE2, the cellular receptor of the virus, to identify animal species with high binding affinity to SARS-CoV-2.

There is still no definitive list of mammals that transmit SARS-CoV-2, making it challenging to implement effective measures to prevent possible new coronavirus outbreaks. Therefore, this work aims to identify the characteristics that make mammals capable of transmitting SARS-CoV-2 and list the main computational methods used to analyze data on the disease in mammals. Identifying the mammals that transmit the virus is crucial to developing effective preventive and therapeutic measures. Currently, conventional identification methods are limited and often rely on incomplete or low-quality data, making using computational methods a promising tool. With this, it is expected to obtain significant results that can apply in preventing and controlling infectious diseases, mainly COVID-19.

2. Theoretical framework

There is information about the coronavirus and the One Health concept to understand this work better.

2.1. Virus evolution

Coronaviruses belong to the order Nidovirales, family Coronaviridae and subfamily Coronavirinae, being RNA viruses divided into four genera: alpha, beta, gamma and deltacoronavirus [28]. They can infect several species of animals and cause or not cause disease symptoms in their hosts [29]. Since the 1930s, they have been known to cause serious animal diseases, such as transmissible swine gastroenteritis virus, bovine CoV, feline infectious peritonitis virus, hepatitis virus, and infectious bronchitis virus [30].

The zoonotic potential of the coronavirus has previously been noted during the SARS and MERS outbreaks. A pandemic caused by SARS-CoV-2 is the third time coronaviruses have crossed the species barrier [2]. SARS-CoV-2, SARS-CoV and MERS-CoV are Betacoronaviruses with bats as their main reservoir [31]. Despite being highly pathogenic, SARS-CoV and MERS-CoV have not adapted well to humans, unlike their successor, SARS-CoV-2, which is more adaptable to the human host [32].

SARS-CoV, for example, caused an epidemic in 2002, which developed rapidly in various regions of the world and affected more than 8,000 people, causing around 800 deaths [32]. SARS-CoV is believe to have been transmitted to humans from civets or wildcats sold in live animal markets in China. The MERS-CoV virus, in turn, was discovered in 2012 and is believed to be transmitted to humans through camels (*Camelus dromedarius*) [33], causing symptoms such as fever, cough and shortness of breath, which can progress to pneumonia and respiratory pneumonia in cases graves [34].

SARS-CoV-2 shares 79% genome sequence identity with SARS-CoV and 50% with MERS-CoV [35]. Genomic analyses revealed several subgroups of the virus that harbor distinct environments, but with a relatively slow evolution compared to other RNA viruses [29], this is largely due to high nucleotide mutation rates that allow viruses to adapt to different environments quickly and changes in their hosts [28].

2.2. Mechanism of viral entry and transmission

Entry of SARS-CoV-2 into the host cell occurs in two main steps: host cell receptor recognition and virus-cell membrane fusion [36]. SARS-CoV-2 has four structural proteins: envelope (E), nucleocapsid (N), membrane (M), and spike (S) [37]. Protein S, responsible for viral entry, comprises three functional domains: *ectodomain*, *transmembrane anchor*, and *short cytoplasmic tail*. The ectodomain is formed by the S1 and S2 subunits, with the S1 subunit containing the receptor binding domain (RBD) that interacts with cell surface receptors (Fig. 1). Several ACE2 receptors are known to be expressed in many human organs, and the interaction between the SARS-CoV-2 S protein RBD and ACE2 is critical for viral entry into the host cell [38]. After receptor recognition, viral binding, and membrane fusion, SARS-CoV-2 releases genetic material into the cell, and viral proteins are synthesized in the host's Endoplasmic Reticulum for viral replication and release of virus particles occurs through the DNA complex Golgi [39].

Furthermore, the interaction between the Spike proteins of SARS-CoV and SARS-CoV-2 and ACE2 involves more than 15 touchpoints. Variations in many of these residues can significantly affect viral entry [40]. Coronaviruses have a high mutation rate, which can lead to changes in their antigenic profile, tissue tropism, and host range. This occurs through two main mechanisms: *antigenic drift*, which leads to the incorporation of wrong nucleotides during replication cycles, and *recombination*, which allows different coronavirus strains to synthesize a hybrid RNA. This process can lead to an adaptation to other species and an increase in viruses' pathogenicity [41].

2.3. One health

The One Health approach has gained prominence as a way to understand and address global health challenges more comprehensively and collaboratively [42]. The system recognizes that human, animal, and environmental health are interconnected and that diseases that affect one species can affect other species, including humans [43]. This means that the health of all species is interconnected and that a concerted effort is needed to prevent and control the disease.

This approach involves interdisciplinary collaboration among human, animal, and environmental health professionals, including human and veterinary clinicians, ecologists, epidemiologists, microbiologists, environmental scientists, and other health professionals [44,45]. The One Health approach allows for a broader view of the causes and effects of disease, enabling healthcare professionals to address illness more effectively and coordinatedly.

3. Systematic literature review

Systematic reviews objectively summarize large amounts of information, identifying gaps in scientific research [46]. The main objective of this Systematic Literature Review (SLR) is to identify which computational methods can be used to analyze SARS-CoV-2 data in mammals. Furthermore, we also seek to determine the characteristics of a suitable SARS-CoV-2 transmitter are. Thus, we developed the following research questions (RQs) to achieve the objective of the work:

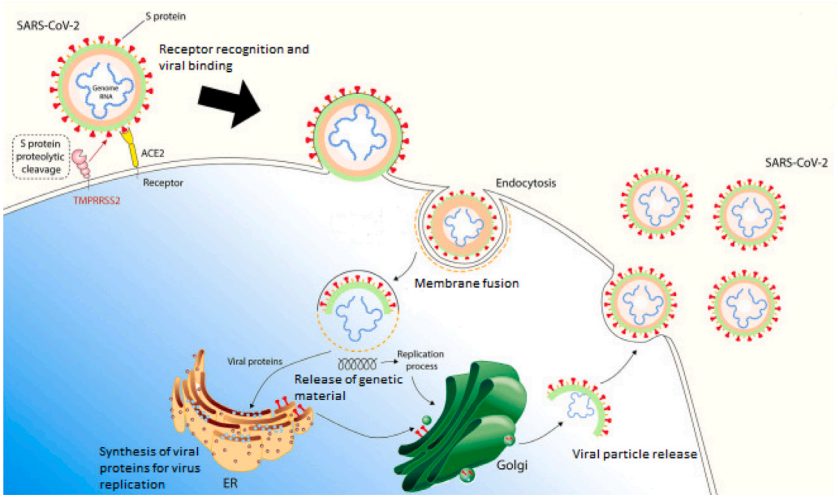


Fig. 1. Mechanism of entry and transmission of the SARS-CoV-2 virus.
Source: Adapted from Pasquarelli-do Nascimento et al. [39]

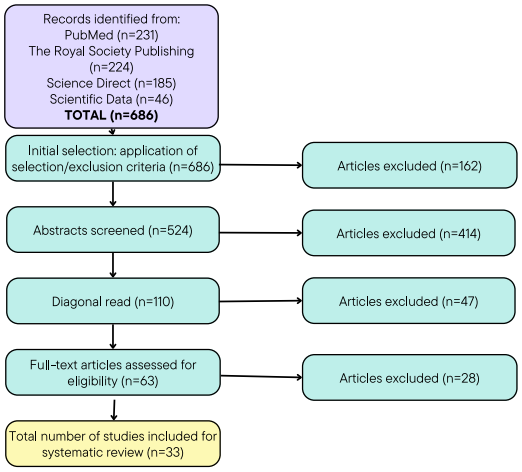


Fig. 2. PRISMA-chart illustrating the selection process used in this study.

1. RQ1: What computational methods are being used to analyze COVID-19 in mammals? What is being studied in these methods?
2. RQ2: What makes a mammal a suitable transmitter of SARS-CoV-2?

3.1. Source selection and search strings

To conduct the bibliographic research, we used the following databases: PubMed, The Royal Society Publishing, ScienceDirect, and Scientific Data. We specify the search period for articles from 2020 onward, date SARS-CoV-2 was considered a pandemic until November 2023. We then searched the four selected databases using three search strings described below:

1. Throughout the text: *COVID-19 AND Sars-Cov-2 AND Mammals AND Veterinary AND Zoonotic AND Transmission*
2. Throughout the text: *COVID-19 AND Sars-Cov-2 AND Mammals and Veterinary AND Zoonotic AND Computational*
3. Title only: *COVID-19 OR Sars-Cov-2 AND Mammals*.

Fig. 2 presents an overview of the selection process. To filter out irrelevant search results, screening was based on the next topic exclusion criteria.

3.2. Selection/exclusion criteria

To retain as many relevant works as possible, the deletion was carried out by applying the following criteria:

1. Works that were not freely available for reading
2. Manuscripts that were not full articles (books, tutorials, editorials, posters, panels, transcripts of lectures and round tables, materials from workshops and demonstrations, and workshop papers were disregarded)
3. SARS-CoV-2 vaccine research, as it is not the focus of this work
4. Articles that dealt exclusively with humans were disregarded, as this is not the purpose of this RSL
5. Publications that were not peer-reviewed
6. No computational method was applied in the research methodology (for the RQ1).

3.3. Research process

Table 1 presents the searched digital repositories and the number of articles maintained in each selection phase. Each report returned by the searches was reviewed individually by the authors of this work, and the final decision regarding the permanence of the article was taken when there was consensus among the authors. Selected only papers that answered at least one of the research questions. Table 1 presents the number of articles found initially and the quantity kept in each selection step.

1. Initial selection: application of selection/exclusion criteria
2. Elimination by summary
3. Diagonal read elimination
4. A full reading of articles not eliminated in the previous steps.

After the complete reading of the 63 articles, another 28 were eliminated, leaving 33 papers used in this systematic literature review.

3.4. Analysis of research questions

This section discusses the research questions presented in Section 4, according to the data and information in the 34 articles selected by the review.

RQ1: What computational methods are being used to analyze COVID-19 in mammals? What is being studied in these methods?

The main computational methods used to analyze COVID-19 reported in the researched literature were machine learning [47–50],

Table 1
The number of articles retained after each of the four filtering phases of the review.

Data base	Selection	Reading by summary	Diagonal reading	Full reading
PUBMED	335	90	53	28
The Royal Society Publishing	3	1	1	1
ScienceDirect	171	11	5	1
Scientific Data	15	7	3	3
Total	**524**	**110**	**63**	**33**

analysis of genomic and phylogenetic data [36,51–53], computational simulation of molecular dynamics [21,40,54–59], creation of tools [60–62], sequencing analysis [63], and comparison *scripts* [64,65]. Each article will be presented below.

1. **Machine Learning**

- Mollentze et al. [47] evaluated the effectiveness of computational models based on ACE2 receptor variation to predict the susceptibility of animals to SARS-CoV-2. The results showed that these models, although they have high precision in the prediction, are not based on processes mechanically linked to the biology of the infection but correlated with the phylogeny of the host. The limited availability of ACE2 sequences misleads projections of the number and geographic distribution of species at risk, and model predictions must be combined with local knowledge of exposure risk to guide surveillance.
- Wardeh et al. [48] used genomic sequencing data from different coronavirus species and information about the biology of potential hosts to develop the prediction model of mammalian hosts that may be sources of new zoonotic coronaviruses. The study relied on a molecular phylogeny analysis and machine learning techniques to identify critical factors contributing to the virus's transmission between species. They found far more associations of coronavirus hosts, potential recombination hosts, and host species with four or more different coronavirus subgenera than have been observed. This demonstrates the significant underestimation of the possible scale of generation of new coronaviruses in wild and domesticated animals. The authors identified high-risk species for coronavirus surveillance, such as the Asian palm civet (*Paradoxurus hermaphroditus*), the great horseshoe bat (*Rhinolophus ferrumequinum*), the intermediate horseshoe bat (*Rhinolophus affinis*), and the pangolin (*Manis javanica*).
- Becker et al. [49] used statistical models to predict possible reservoirs of zoonotic viruses in bats. They developed eight models and found that models based on ecological traits better predict new hosts. They also found that including host ecology in predictive models is essential. By reviewing their models, they were able to expect more than 400 bat species globally that may be undetected hosts of betacoronavirus. The study highlights the importance of systematic validation of models and the need for a dynamic forecasting process, data collection, verification, and updating.
- Kou et al. [50] proposed a *Deep Learning* model to predict the risk of pandemics based on the viral genome sequences of coronaviruses of animal origin. The model combines a network of recurrent units with a one-dimensional convolution and uses a pre-trained DNA vector and attention mechanism to obtain the best performances. The results show that the model can accurately predict the risk of cross-species infection and provides an early warning for the next pandemic.

Table 2 highlights the Machine Learning algorithms used and the address of the data, if available.

2. **Analysis of genomic and phylogenetic data**

- Gupta et al. [36] combined phylogenetic, bioinformatics, and molecular modeling analyses to evaluate the hypothesized host ACE2-interacting residues to the spike protein receptor in SARS-CoV-2 isolates from bats and pangolins. Based on the comparative analysis, the authors support the view that Guangdong pangolins are the intermediate hosts that adapted to SARS-CoV-2 and represented a significant evolutionary link in the transmission path of the SARS-CoV-2 virus. Furthermore, the article discusses the role of intermediate hosts in the origin of the omicron variant.
- Lytras et al. [51] examined the genetic structure of SARS-CoV-2 and compared it to other known coronaviruses, including those found in bats and pangolins. The study also looked at the molecular evolution of the virus and how it might have mutated and recombined over time. The evolutionary history of these coronaviruses shows relatively recent geographic movement and co-circulation among bat hosts. The analysis highlights the need for more wildlife sampling to identify the exact origins of the virus and possible intermediate species that facilitated transmission to humans.
- Magateshvaren Saras et al. [52] performed genomic analysis of SARS-CoV-2 in different animal hosts and geographic regions, using various genomic and statistical analysis techniques to assess the genetic diversity and evolution of the virus. Some recurrent mutations were identified, and SARS-CoV-2 variants with these mutations were seen in human and non-human sequences from the same country, indicating distinct epidemiological dynamics. The results highlight the importance of surveilling viral evolution in non-human hosts during pandemics.
- Rojas-Cruz et al. [53] used an integrative genetic, structural, and functional analysis approach to identify genetic mutations that may be related to the ability of SARS-CoV-2 and other betacoronavirus coronaviruses to cross host species barriers. The results showed that mutations in viral proteins, mainly in the S1 subunit of the S protein, boosted viral diversification. Furthermore, highly conserved RNA structures were found in Betacoronavirus genomes, suggesting essential functions in viral biology that have yet to be investigated. According to the authors, more research is needed to examine the potential of encoding small RNAs derived from viruses and to develop new antiviral therapeutic strategies.

3. **Computer simulation of molecular dynamics**

- Lupala et al. [54] analyzed the sequences of some mammalian ACE2 proteins and predicted the structures of the ACE2-RBD complexes by homology modeling, and the complexes were refined using molecular dynamics simulation. Sequence, structure, and dynamics analyses provide valuable insight into interactions between ACE2 and RBD. The analysis results suggest that ACE2 from cattle, cats, and pandas form solid binding interactions with RBD, whereas in the cases of rats, bats, horses, pigs, mice, and civets, ACE2 proteins interact weakly with RBD.

Table 2
Works, methods and data availability.

Article	Method	Data availability
Mollentze et al. [47]	Gradient-boosted classification tree	https://zenodo.org/record/7185111#.ZB9lXtDMLrc
Wardeh et al. [48]	Gradient Boosting Machine (GBM)	https://doi.org/10.6084/m9.figshare.13110896
Becker et al. [49]	Random Forest	github.com/viralemergence/Fresnel_Jun
Kou et al. [50]	Deep Learning	https://ngdc.cncb.ac.cn/ncov/

- Rajendran and Babbitt [55] used comparative molecular dynamics simulations to predict the infectivity of SARS-CoV-2 variants in different animal species. High-resolution simulation models of the ACE2 receptor from other animals, including humans, primates, dogs, cats, ferrets, and pangolins, were used. They predicted that there is still a significant risk of mammalian cross-infectivity of the human variants during the successive waves of infection as COVID-19 transitions from a pandemic to an endemic state.
- Celik and Tallei [56] performed computer simulations to analyze the binding mechanism between the active metabolite of molnupiravir (MPT) and the RNA-dependent RNA polymerase of SARS-CoV-2. According to the results of this study, MTP has a high probability of becoming widely used as an anti-SARS-CoV-2 agent. The fact that MTP is not only cytotoxic but also mutagenic for mammalian cells, as well as the possibility of causing damage to the host's DNA, were raised as possible concerns.
- Khaledian et al. [40] presented a combined laboratory and computer network science approach to identify ACE2-independent determinants of human cell entry in bat arboviruses. The results show a similar receptor on ACE2-independent viruses that can infect human and bat cells in culture. These sequence determinants of human cell entry map to an exposed protrusion on the surface of the predicted bat arbovirus spike receptor binding domain structure. The findings provide additional evidence for a group of bat-derived arboviruses with zoonotic potential.
- Hemmati and Tabein [57] used a computational approach to screen a library of insect protease inhibitors for their potential to bind and inhibit the SARS-CoV-2 major protease. The screening method used molecular docking and molecular dynamics simulations to predict the binding affinity of each inhibitor to the protease active site and to assess the stability of the inhibitor-protease complex.
- Chen et al. [58] used computational tools to simulate the interaction between the SARS-CoV-2 spike protein and the ACE2 receptor in different animal species, including dogs, cats, tigers, lions, pangolins, bats, and humans. They analyzed structural differences in ACE2 proteins from other animals and evaluated the binding affinity of the spike protein with these ACE2 proteins. It was observed that most of the incorrect mutations in the RBD region of the interaction interface did not significantly affect ACE2-S binding. However, some mutations within the RBD region have increased the virus's binding affinity for human ACE2, which may make it more contagious. On the other hand, modeling the interactions between animal ACE2 molecules and the SARS-CoV-2 spike protein revealed that many pets and wild animals showed different levels of virus-binding ability.
- Sekar et al. [21] investigated the ability of alpha-helical antimicrobial peptides derived from frog skin to inhibit the interaction between the spike protein of SARS-CoV-2 and the ACE2 receptor using computer simulations and in vitro assays. After analyses, they concluded that the antimicrobials studied may be strong candidates for a therapeutic scaffold to prevent SARS-CoV-2 infection.
- Khan et al. [59] used computational methods to investigate possible SARS-CoV-2 S protein receptor mutations that may lead to increased binding affinity with the human ACE2 receptor. They predicted the possible structural variants of residue 501, which impose a more robust interaction response and infectivity. The results show that some variants that have shown aptitude in animals, such as ferrets, may aggravate the situation further.

4. Sequencing analyzes

- In Peng et al. [63], the authors used sequencing methods to analyze pangolin samples and identify different coronaviruses related to SARS-CoV-2. The analysis was based on collecting samples of pangolins illegally trafficked in China and other Asian countries. The work revealed that the genetic diversity of pangolin-CoVs is substantially greater than previously estimated. Given the potential infectivity of pangolin-CoVs, the high genetic diversity of pangolin-CoVs warns of the ecological risk of zoonotic evolution and transmission of pathogenic SC2r-CoVs.

5. Use of bioinformatics tools

- Burkholz et al. [64] analyzed the viral genomes of human and mink (*Neovison vison*) samples and identified paired mutations in the spike protein unique to mink samples. Viral genomic mutations observed in mink in the Netherlands and Denmark show the potential for new mutations in the RBD spike protein of SARS-CoV-2 to be introduced into humans by zoonotic transfer.
- King and Singh [65] presented a comparative analysis of the genomes of several mammalian species, including humans, to investigate their susceptibility to coronavirus infections. The authors found evidence of adaptive amino acid substitutions in the ACE2-spike interaction, while variation within ACE2 proteins in primates and some mammals is inconsistent with evolutionary adaptations.
- Dong et al. [60] developed a bioinformatics tool called LETRS (Local Exhaustive Transcriptomic Reconstruction from Short-reads) to analyze the subgenomic RNA of SARS-CoV-2 in different types of samples and investigate the presence of subgenomic mRNAs (sg mRNAs), molecules that SARS -CoV-2 produces when infecting cells. Their results can be used to evaluate the biology of SARS-CoV-2 in clinical and non-clinical samples, mainly to evaluate different variants and medical countermeasures that can influence viral RNA synthesis.
- Kaushik et al. [61] created and implemented a framework-based method to identify vertebrate susceptibility to SARS-CoV-2. They used amino acid sequences from ACE2 proteins to predict 299 binding affinities (via dissociation constants) with the spike protein from SARS-CoV-2. The results show that the SARS-CoV-2 spike protein can bind to several ACE2-carrying vertebrate species, implying a broad host range at the viral entry level, which may contribute to cross-species transmission and subsequent virus evolution.
- Robinson et al. [62] used a previously developed computational tool called *EpitopeBuilder*, which predicts the structural interactions between an antibody and its target

antigen. They then used this tool to analyze a panel of previously reported antibodies that bind to the spike protein of coronaviruses, including some that have been shown to neutralize SARS-CoV-2.

Table 3 presents the main strengths, weaknesses, and applications of the computational methods found during this work.

3.4.1. Comparison between works

Mollentze et al. [47] highlight the limitations of computational models based on ACE2 receptor variation, emphasizing the need for combination with local knowledge to guide surveillance. Multidisciplinary research gains prominence, highlighting the diversity of approaches, from genomic analyses to molecular simulations, fundamental to understanding the SARS-CoV-2-mammal interaction.

The importance of international collaborations is highlighted in the implications for public policy and future research, underlining the need for integrated approaches at local, national, and international levels. Emphasis on continued surveillance is crucial for a global approach to monitoring and controlling zoonotic diseases.

Understanding transmission pathways and natural reservoirs guides prevention policies, including wildlife trade regulations and biosecurity measures. Furthermore, detailed investigation of the molecular interactions between the virus and host policy information on developing and distributing antiviral therapies and vaccines.

The need for a coordinated response to the monitoring and controlling of zoonotic diseases is highlighted, including international collaborations, sharing of data and resources, and rapid response protocols. General conclusions emphasize a comprehensive approach to studies, contributing to the global understanding of the biology and epidemiology of SARS-CoV-2 in different contexts and host species.

Various approaches, from phylogenetic analyses to molecular dynamics simulations, highlight the importance of diverse methods in understanding and preventing the spread of zoonotic viruses. The warning of the need for continuous and collaborative research reflects the virus's ongoing evolution and future perspectives, emphasizing the importance of a proactive and integrated approach.

RQ2: What makes a mammal a suitable transmitter of SARS-CoV-2?

Several factors can contribute to making a mammal a suitable transmitter of SARS-CoV-2. Here are some features found during the review that may influence the mammalian transmission of the virus:

1. **Presence of ACE2 receptor**

   - Liu et al. [76] analyzed the ACE2 characteristics of several species to determine their ability to withstand the entry of SARS-CoV-2. Through analysis of five specific residues in ACE2, 80 mammalian ACE2 proteins that could mediate virus entry were identified. Among these proteins, 44 ACE2 orthologs, including domestic animals, pets, farm animals, and animals found in zoos and aquariums, were shown to bind to the SARS-CoV-2 Spike protein and facilitate its entry into cells.
   - Zhao et al. [77] investigated ACE2 receptor activity in 14 mammalian species to determine their ability to support the entry of SARS-CoV-2. They found that ACE2 from several species can access virus pseudotyped with the SARS-CoV-2 S protein. ACE2 receptor activity varied between species, with human/rhesus monkey ACE2 having the highest training and rat/mouse ACE2 having the lowest. Furthermore, rabbit and pangolin ACE2s demonstrated strong binding to the S1 subunit of the SARS-CoV-2 S protein and efficiently supported pseudotyped virus infection.
   - Kim et al. [78] studied the impact of SARS-CoV-2 S protein mutations on host cell receptor interactions and antibody-neutralizing ability. Using pseudoviruses and cell lines from nine animal species, it was observed that all species tested, except minks, allowed viral entry of pseudoviruses containing the ancestral S protein at levels comparable to the human ACE2 receptor.
   - Khaledian et al. [40] identified viruses derived from bats capable of infecting human cells, even without using the ACE2 protein. These viruses share a similar receptor binding motif, which makes it possible for other bat viruses to enter human cells.
   - Tan et al. [79] highlighted that SARS-CoV-2 could infect cells of several mammalian host species, mainly due to the conservation of angiotensin-converting enzyme 2 (ACE2). The authors found that circulation of SARS-CoV-2 in mink and deer resulted in some degree of viral adaptation to its animal host but not in high mutation rates or significant changes in the evolutionary landscape of the virus, highlighting a generalistic nature. SARS-CoV-2 as a mammalian pathogen as the mutational prerequisite for efficient transmission of SARS-CoV-2 in new hosts is low.

2. **Interaction with humans**

   - Stout et al. [80] highlighted evidence that the virus could be transmitted from humans to cats and other pets. Still, whether these animals can transmit it back to humans remains unclear. In this sense, the authors emphasize the importance of continuous surveillance for the presence of SARS-CoV-2 in pets and the need for adequate preventive measures to protect both animals and humans.
   - Tiwari et al. [2] discussed the relationship between the Covid-19 pandemic and animals, highlighting the importance of surveillance and monitoring of animals, especially those that live near humans. Repeated human–animal interactions in the market or the animal industry without using environmental biosecurity have been considered significant risk factors for the emergence of zoonotic diseases such as COVID-19.
   - Prince et al. [81] address the possibility that animals act as reservoirs for SARS-CoV-2 and play a role in transmitting the virus to humans. One example is livestock, an area of particular concern given close human–animal contact and the potential for a high population density of some livestock species and a threat to food supply chains.

3. **Ability to carry out recombination and mutations**

   - Dhama et al. [82] point out that in natural hosts or reservoirs, coronaviruses adapt well; however, as they are stable RNA viruses, they continue to multiply continuously without producing diseases, thus allowing persistence or survivability and accumulation of mutations over time., resulting in new virus strains that occasionally spread to other species, adapting to their bodily systems and expanding the range of biological hosts for evolutionary sustainability.

4. **Other features**

   - Ye et al. [32] discuss how bats play a significant role in cross-species virus spread due to their broad array of virus species and favorable traits such as longevity, densely packed colonies, close social interaction, and strong ability to fly.
   - Ruiz-Aravena et al. [83] point out that co-infections with multiple pathogens can influence the transmission of SARS-CoV-2 to conspecifics and spillover hosts. Cross-protective

Table 3
Strengths, limitations and applicability of computational methods.

Computational method	Strengths	Weaknesses	Applications
Machine Learning	When exposed to new data, machine learning models can adapt independently; they learn from previous computations to produce reliable, repeatable decisions and results. Using statistical methods, algorithms are trained to make classifications or predictions and discover essential insights in data mining projects [66]. It can identify complex patterns in large data sets, can be applied in several areas, including biological data analysis, and has the potential to predict future relationships and behaviors.	The development of machine learning models is complex and heavily depends on the quality and quantity of available data [66]. Depending on the quality and quantity of training data, it can be challenging to interpret model decisions, especially in more complex models. Choosing the appropriate algorithm and adjusting the parameters requires specialized knowledge.	Machine learning has the potential to contribute to clinical research by increasing the power and efficiency of pre-trial basic/translational research and enhancing the planning, conduct, and analysis of clinical trials. [67]. Other fields of application of machine learning are intelligent decision-making through data-based predictive analysis [68], prediction of protein structures, classification of genes associated with diseases, and identification of patterns in genomic data.
Analysis of genomic and phylogenetic data	Efficiency and low cost of sequence data acquisition and the development of analytical methods to deal with many characters with a small number of states [69]. Allows understanding of genetic and evolutionary diversity; Facilitates the identification of phylogenetic relationships between species; Important for studies of evolution and taxonomy.	The lack of linkage information limits the ability to use metagenomic data for phylogenetic and population genetic analysis since most current methods assume complete linkage information is available [70]. It may require significant computing resources; Interpretation of results can be challenging in cases of rapid evolution or horizontal gene transfer events.	Community composition from metagenomes is the taxonomic classification of metagenome sequences between others [70]. Reconstruction of phylogenetic trees; Identification of genetic markers associated with specific characteristics.
Computer simulation of molecular dynamics	Notable improvements in the speed, accuracy, and accessibility of the simulation, combined with the proliferation of experimental structural data, have increased the appeal of biomolecular simulation. These simulations have proven to be valuable in understanding the functional mechanisms of proteins and other biomolecules in discovery of the structural basis of diseases and the design and optimization of small molecules, peptides, and proteins [71]. Provides detailed insights into molecular dynamics and interactions; Allows the study of biochemical processes on an atomic scale; Valuable for understanding the structure and function of macromolecules.	To carry out reliable and high-quality work in molecular dynamics, it is necessary to identify some critical issues: (1) design simulations appropriately; (2) configure these simulations carefully; (3) meticulously analyze the simulations; (4) Consider several sources of error that can affect the results and the expected statistical fluctuation from one simulation to another.; and (5) compare the results with available experimental data and, when possible, (6) design follow-up experiments to validate the results further [71]. High computational cost; Simplifications in simulations may affect the accuracy of the results.	Determination of Structures and Movements of Biomolecules, Assessment of Accuracy and Refinement of Modeled Structures, Flexibility of Molecules, Determination of mechanism in which a biomolecular system will respond to perturbation [72]. Investigating protein dynamics and ligand–receptor interactions; Developing medicines through understanding molecular interactions.
Sequencing analyses	It can improve assembly, mapping certainty, transcript isoform identification, and detection of structural variants. Long-read sequencing of native molecules eliminates amplification bias while preserving base modifications [73]. It allows for identifying specific genetic sequences, is fundamental for genomic and transcriptomic studies, and facilitates the discovery of genetic variants associated with diseases.	Requires tailored analysis tools due to the qualitative difference from second-generation sequencing. The fast-paced development of such devices can be overwhelming [73]. Requires careful handling of sequencing errors; Analyzing large volumes of data can be challenging.	Used for a broad range of applications in genomics for model and non-model organisms [73]. Identification of genetic mutations associated with diseases; Gene expression studies.
Use of bioinformatics tools	It allows for more efficient and accurate analysis of large volumes of information, can be used to sift through enormous amounts of data from multiple studies, they exponentially increase the usefulness of past data as researchers mine information to make new connections. Facilitates the analysis of biological data; Offers a variety of tools for different purposes; and Automates complex data processing tasks.	It is not as accessible to most biologists, and the heterogeneity of how data are analyzed, annotated, and displayed and the lack of connectivity among the available data [74]. Dependence on the quality of the tools used; There can be integration challenges when using multiple tools.	Genomics, proteomics, metabolomics, transcriptomics, molecular phylogenomics, development of biomarkers to create safer and more personalized drugs, among others [75]. DNA sequencing analysis; Prediction of protein structures; Gene expression studies.

immunity from infection by related pathogens can reduce susceptibility or transmission. In contrast, offsets in the immune response to one pathogen may increase exposure and facilitate the transmission of another.

- Petrovan et al. [84] discussed non-taxon-specific features that also play a role in zoonotic pathogen risk, such as species migration that can significantly impact pathogen dispersal.

4. Final considerations

This systematic review of the literature on using computational methods to assess SARS-CoV-2 infection in mammals revealed valuable insights relevant to understanding and controlling the spread of the disease. The application of Machine Learning techniques and other computational methods allowed the efficient identification and classification of mammalian species with the potential to be carriers or intermediate hosts of the virus, contributing to the early detection of potential threats to public health.

The studies analyzed in this review highlighted the utility of computational methods in identifying specific patterns and characteristics in mammalian epidemiological and genetic data. They highlighted the utility of computational methods in identifying particular patterns and characteristics in mammalian epidemiological and genetic data. This information makes it possible to identify groups of animals most likely to harbor the virus, which can help with the adoption of specific preventive measures. Data from work can help identify areas where more research is needed, factors that increase the risk and make it challenging to control zoonoses, and strategic actions in surveillance, research, communication, and training that can support the formation of a cooperation network. In addition, this data can be used to educate the public about zoonoses, their risks, and how to prevent them.

Furthermore, the review also demonstrated that combining data from different sources, such as environmental and climate data, with mammalian data can further improve the accuracy of Machine Learning models in predicting the spread of SARS-CoV-2. This integrated approach can provide valuable information for implementing more effective surveillance and control strategies.

Although the studies reviewed showed promising results, it is vital to highlight some challenges and limitations in using computational methods for analyzing zoonotic diseases:

- *Data availability*: The effectiveness of computational models depends on the quality and quantity of available data. In many cases, there may be a need for comprehensive, reliable, and up-to-date data on various aspects of zoonotic diseases. This can limit the accuracy and reliability of the models.
- *Generalization*: Generalization is one of the main goals in training machine learning models. It refers to the ability of a model to make accurate predictions on unseen data, that is, in situations beyond those in which it was trained. Several factors can affect the generalizability of a model: (1) *Size and Representativeness of the Training Set*: If the training set is too small or does not adequately represent the diversity of data that the model will encounter in practice, the model may have difficulty generalizing; (2) *Overfitting*: Overfitting occurs when a model overfits to the specific details of the training set, capturing noise or random variations. This can harm the model's ability to generalize to new data; (3) *Underfitting*: Contrary to overfitting, underfitting occurs when a model is too simple to capture the complexity of the data. This can result in poor performance on both the training set and new data; (4) *Irrelevant or Redundant Features*: The presence of irrelevant or redundant features in the data can hinder the generalization of the model. Feature selection techniques can be applied to mitigate this problem; (5) *Bias in Data*: If the training set contains bias, the model can learn to reproduce it. This can lead to biased predictions and a lack of generalization to more diverse data, among other factors. Furthermore, models are often developed based on specific sets of data, which may not represent all situations or regions. This may limit the generalization of the models.
- *Need for continuous updates*: As new data becomes available, machine learning algorithms may need to be updated to reflect emerging patterns or changes in the distribution of the data. Keeping models up to date allows them to continue to make accurate predictions. On the other hand, Zoonotic diseases are influenced by various factors, including genetic, ecological, socioeconomic, and climatic factors. These factors can change over time, requiring continual updates to models to ensure they remain accurate and relevant.
- *Complexity of interactions*: Zoonotic diseases involve complex interactions between humans, animals, and the environment. Incorporating these interactions into models can be challenging but is crucial for a holistic understanding of disease transmission.

Furthermore, we know that a systematic literature review is a rigorous and methodological approach to analyzing and synthesizing existing research on a specific topic. However, like any method, it has some limitations, namely:

1. *Publication bias*: Systematic review may be subject to publication bias, as studies that do not produce significant results may be less likely to be published. This can lead to a distorted view of the body of available evidence.
2. *Language bias:* The limitation of only including studies published in certain languages may result in an incomplete view of the body of literature. Relevant studies in other languages may be excluded, introducing a linguistic bias.
3. *Selection of sources and materials*: The selection of sources and materials can influence the results. The review may not be fully comprehensive if certain databases, journals, or specific sources are excluded. This work conducted a bibliographical search in four electronic databases: Pubmed, ScienceDirect, Scientific Data, and The Royal Society Publishing. Therefore, additional relevant studies available on other databases may need to be included. Additionally, we excluded articles that were not freely available, which may have excluded work pertinent to this study.
4. *Variation in the quality of included studies*: The studies included in a systematic review may vary in methodological quality. Some studies may be more robust than others, which may affect the reliability of the review's conclusions.
5. *Challenges in identifying relevant studies*: Searching for relevant studies can be challenging, even with well-defined search strategies. Selecting search terms and determining inclusion criteria can influence results.
6. *Quick search changes*: In dynamic research areas, the results of a systematic review can quickly become outdated due to the constant emergence of new studies and discoveries.
7. *Difficulties in synthesizing diverse data*: Heterogeneity in methods and study populations can make synthesis and direct comparison of results difficult. This may limit the Ability to reach definitive conclusions.
8. *Time and resource constraints:* Time and resource constraints may restrict the scope and extent of the review. A complete and comprehensive review may be impractical in some cases.
9. *Generalization uncertainty:* The generalizability of the results of a systematic review of different contexts or populations can be uncertain, especially if the diversity of included studies is limited.

Despite these limitations, systematic reviews remain valuable tools for synthesizing evidence and informing decision-making. Researchers must recognize these limitations when interpreting systematic review results and consider the need to seek additional sources of evidence when appropriate.

In conclusion, using computational methods to evaluate SARS-CoV-2-transmitting mammals can provide valuable information for preventing and controlling epidemic outbreaks. These innovative approaches

can contribute to the early identification of possible virus reservoirs, allowing the implementation of mitigation measures and targeted actions to minimize public health risks, as mentioned in Section 4.1. However, it is essential to continue investing in research and development to improve the capabilities of these models and ensure their applicability in different epidemiological contexts.

4.1. Public policies and future research

Implications for public policy and future research center on the need for integrated and collaborative approaches, considering the complexity of interactions between the virus, mammals, and the environment. This approach requires ongoing monitoring, multidisciplinary research, and coordinated action at local, national, and international levels.

Understanding transmission pathways and natural reservoirs emerges as a crucial direction for prevention policies, especially in scenarios where zoonotic transmission is a concern. This may involve implementing wildlife trade regulations and biosecurity measures.

Detailed analysis of the molecular interactions between the virus and the host provides valuable insights to guide policies for developing and distributing antiviral therapies and vaccines.

The findings highlight the importance of a global approach to monitoring and controlling zoonotic diseases. This global approach includes international collaborations, sharing data and resources, and establishing rapid response protocols to address emerging challenges effectively.

Thus, based on the studies above on the interaction of SARS-CoV-2 with mammals and the implications for public policy, future research can explore several areas to expand knowledge and improve prevention strategies. Some proposals for future work include:

1. *Deepening into ACE2 Receptor Variation:* Further investigation into ACE2 receptor variation in different species may improve predictive models by considering specific genetic nuances that affect virus–host interaction.
2. *Improvement of Computational Models:* Development and improvement of computational models for more accurate predictions of the susceptibility of different mammals to SARS-CoV-2. This may include integrating genomic, phylogenetic, and molecular dynamics data.
3. *Study of Epidemiology in Wild Environments:* Expanding epidemiological research in wild environments to identify natural reservoirs, transmission routes, and environmental factors that can influence the spread of the virus.
4. *Experimental Validation of Models:* Carrying out experimental studies to validate the developed predictive models, providing a more robust approach and ensuring the practical applicability of the findings.
5. *Diverse Species Research:* Expanding research to include a wider variety of mammal species, especially those that interact closely with humans, such as pets and wild animals that frequent urban areas.
6. *Investigation of Viral Evolution in Hosts:* Studies on viral evolution in different hosts focus on how the virus adapts and evolves in mammals over time, contributing to control strategies.
7. *Development of Biosafety Strategies:* Development of specific biosafety strategies for different contexts, considering the identified transmission routes and host characteristics.
8. *Impact Assessment on Human and Animal Populations:* Assessing the impact of zoonotic transmission on human and animal populations, including studying possible outbreaks and transmission dynamics between different groups.
9. *Implementation of Findings-Based Preventative Measures:* Discovery-based preventive measures, regulations, and public policies to mitigate risks of zoonotic transmission, such as regulations on wildlife trade and enhancement of biosecurity measures at human–animal interfaces.
10. *Study of the Role of Intermediate Reservoirs:* Investigation of the role of possible intermediate reservoirs in the transmission of the virus between species, with a focus on understanding how certain animals can facilitate the spread of SARS-CoV-2.
11. *Assessment of Therapies and Vaccines in Different Species:* Assessment of antiviral therapies and vaccines in different species to develop effective prevention and treatment strategies, considering the specific characteristics of the hosts.
12. *Socioeconomic Impact Analysis:* Analysis of the socioeconomic impacts of implemented policies, considering the interaction between human, animal, and environmental health, to ensure sustainable and equitable approaches.

CRediT authorship contribution statement

Paula Vitória Silva: Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Cristiane N. Nobre:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] John P. Evans, Shan-Lu Liu, Role of host factors in SARS-CoV-2 entry, J. Biol. Chem. 297 (1) (2021).
[2] Ruchi Tiwari, Kuldeep Dhama, Khan Sharun, Mohd Iqbal Yatoo, Yashpal Singh Malik, Rajendra Singh, Izabela Michalak, Ranjit Sah, D Katterine Bonilla-Aldana, Alfonso J Rodriguez-Morales, COVID-19: Animals, veterinary and zoonotic links, Vet. Q. 40 (1) (2020) 169–182.
[3] Pooja Arora, Mohammad Jafferany, Torello Lotti, Roxanna Sadoughifar, Mohamad Goldust, Learning from history: Coronavirus outbreaks in the past, Dermatol. Ther. 33 (4) (2020) e13343.
[4] Khalid Munir, Shoaib Ashraf, Isra Munir, Hamna Khalid, Mohammad Akram Muneer, Noreen Mukhtar, Shahid Amin, Sohaib Ashraf, Muhammad Ahmad Imran, Umer Chaudhry, et al., Zoonotic and reverse zoonotic events of SARS-CoV-2 and their impact on global health, Emerg. Microbes Infect. 9 (1) (2020) 2222–2235.
[5] Nelson Lee, David Hui, Alan Wu, Paul Chan, Peter Cameron, Gavin M Joynt, Anil Ahuja, Man Yee Yung, CB Leung, KF To, et al., A major outbreak of severe acute respiratory syndrome in Hong Kong, N. Engl. J. Med. 348 (20) (2003) 1986–1994.
[6] Aasish Gautam, Krishna Kaphle, Birendra Shrestha, Samiksha Phuyal, Susceptibility to SARS, MERS, and COVID-19 from animal health perspective, Open vet. J. 10 (2) (2020) 164–177.
[7] Saskia L Smits, Anna De Lang, Judith MA Van Den Brand, Lonneke M Leijten, Wilfred F Van Ijcken, Marinus JC Eijkemans, Geert Van Amerongen, Thijs Kuiken, Arno C Andeweg, Albert DME Osterhaus, et al., Exacerbated innate host response to SARS-CoV in aged non-human primates, PLoS Pathog. 6 (2) (2010) e1000756.
[8] Thijs Kuiken, Ron AM Fouchier, Martin Schutten, Guus F Rimmelzwaan, Geert Van Amerongen, Debby Van Riel, Jon D Laman, Ton De Jong, Gerard Van Doornum, Wilina Lim, et al., Newly discovered coronavirus as the primary cause of severe acute respiratory syndrome, Lancet 362 (9380) (2003) 263–270.
[9] Ziqi Zhou, Abraham Ali, Elias Walelign, Getnet F Demissié, Ihab El Masry, Takele Abayneh, Belayneh Getachew, Pavithra Krishnan, Daisy YM Ng, Emma Gardner, et al., Genetic diversity and molecular epidemiology of middle east respiratory syndrome Coronavirus in dromedaries in Ethiopia, 2017 to 2020, Emerg. Microbes Infect. (just-accepted) (2023) 2164218.
[10] Ali M Zaki, Sander Van Boheemen, Theo M Bestebroer, Albert DME Osterhaus, Ron AM Fouchier, Isolation of a novel Coronavirus from a man with pneumonia in Saudi Arabia, N. Engl. J. Med. 367 (19) (2012) 1814–1820.
[11] Benjamin Meyer, Marcel A Müller, Victor M Corman, Chantal BEM Reusken, Daniel Ritz, Gert-Jan Godeke, Erik Lattwein, Stephan Kallies, Artem Siemens, Janko van Beek, et al., Antibodies against MERS Coronavirus in dromedary camels, united Arab Emirates, 2003 and 2013, Emerg. Infect. Dis. 20 (4) (2014) 552.
[12] Sidra Rahman, Sana Ullah, Zabta Khan Shinwari, Muhammad Ali, Bats-associated beta-Coronavirus detection and characterization: First report from Pakistan, Infect. Genet. Evol. 108 (2023) 105399.

[13] Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, et al., Clinical features of patients infected with 2019 novel Coronavirus in Wuhan, China, Lancet 395 (10223) (2020) 497–506.
[14] Alessio Lorusso, Paolo Calistri, Antonio Petrini, Giovanni Savini, Nicola Decaro, Novel Coronavirus (SARS-CoV-2) epidemic: A veterinary perspective, Vet. Italiana 56 (1) (2020) 5–10.
[15] Gary Wong, Yu-Hai Bi, Qi-Hui Wang, Xin-Wen Chen, Zhi-Gang Zhang, Yong-Gang Yao, Zoonotic origins of human Coronavirus 2019 (HCoV-19/SARS-CoV-2): Why is this work important? Zool. Res. 41 (3) (2020) 213.
[16] Mohamed A.A. Mahdy, Waleed Younis, Zamzam Ewaida, An overview of SARS-CoV-2 and animal infection, Front. Vet. Sci. 7 (2020) 596391.
[17] Zoë L Grange, Tracey Goldstein, Christine K Johnson, Simon Anthony, Kirsten Gilardi, Peter Daszak, Kevin J Olival, Tammie O'Rourke, Suzan Murray, Sarah H Olson, et al., Ranking the risk of animal-to-human spillover for newly discovered viruses, Proc. Natl. Acad. Sci. 118 (15) (2021) e2002324118.
[18] Han Sang Yoo, Dongwan Yoo, COVID-19 and veterinarians for one health, zoonotic-and reverse-zoonotic transmissions, J. Vet. Sci. 21 (3) (2020).
[19] José de la Fuente, Isabel G Fernández de Mera, Christian Gortázar, Challenges at the host-arthropod-Coronavirus interface and COVID-19: A one health approach, Front. Biosci.-Landmark 26 (8) (2021) 379–386.
[20] Anna DJ Korath, Jozef Janda, Eva Untersmayr, Milena Sokolowska, Wojciech Feleszko, Ioana Agache, Ahmed Adel Seida, Katrin Hartmann, Erika Jensen-Jarolim, Isabella Pali-Schöll, One health: EAACI position paper on Coronaviruses at the human-animal interface, with a specific focus on comparative and zoonotic aspects of SARS-Cov-2, Allergy 77 (1) (2022) 55–71.
[21] P Chandra Sekar, E Srinivasan, G Chandrasekhar, D Paul, G Sanjay, S Surya, NS Kumar, R Rajasekaran, Probing the competitive inhibitor efficacy of frog-skin alpha helical AMPs identified against ACE2 binding to SARS-CoV-2 S1 spike protein as therapeutic scaffold to prevent COVID-19, J. Mol. Model. 28 (5) (2022) 1–13.
[22] Gregory F Albery, Daniel J Becker, Liam Brierley, Cara E Brook, Rebecca C Christofferson, Lily E Cohen, Tad A Dallas, Evan A Eskew, Anna Fagre, Maxwell J Farrell, et al., The science of the host–virus network, Nat. Microbiol. 6 (12) (2021) 1483–1492.
[23] Joana Damas, Graham M Hughes, Kathleen C Keough, Corrie A Painter, Nicole S Persky, Marco Corbo, Michael Hiller, Klaus-Peter Koepfli, Andreas R Pfenning, Huabin Zhao, et al., Broad host range of SARS-CoV-2 predicted by comparative and structural analysis of ACE2 in vertebrates, Proc. Natl. Acad. Sci. 117 (36) (2020) 22311–22322.
[24] Ashutosh Kumar, Sada N Pandey, Vikas Pareek, Ravi K Narayan, Muneeb A Faiq, Chiman Kumari, Predicting susceptibility for SARS-CoV-2 infection in domestic and wildlife animals using ACE2 protein sequence homology, Zoo Biol. 40 (1) (2021) 79–85.
[25] SD Lam, N Bordin, VP Waman, HM Scholes, P Ashford, N Sen, L Van Dorp, C Rauer, NL Dawson, CSM Pang, et al., SARS-CoV-2 spike protein predicted to form complexes with host receptor protein orthologues from a broad range of mammals, Sci. Rep. 10 (1) (2020) 1–14.
[26] Zhixin Liu, Xiao Xiao, Xiuli Wei, Jian Li, Jing Yang, Huabing Tan, Jianyong Zhu, Qiwei Zhang, Jianguo Wu, Long Liu, Composition and divergence of Coronavirus spike proteins and host ACE2 receptors predict potential intermediate hosts of SARS-CoV-2, J. Med. Virol. 92 (6) (2020) 595–601.
[27] Amanda D. Melin, Mareike C. Janiak, F. Marrone III, Paramjit S. Arora, James P. Higham, Comparative ACE2 variation and primate COVID-19 risk, Commun. Biol. 3 (1) (2020) 641.
[28] Diego Forni, Rachele Cagliani, Mario Clerici, Manuela Sironi, Molecular evolution of human Coronavirus genomes, Trends Microbiol. 25 (1) (2017) 35–48.
[29] Devika Singh, Soojin V. Yi, On the origin and evolution of SARS-CoV-2, Exper. Mol. Med. 53 (4) (2021) 537–547.
[30] L.J. Saif, Animal coronaviruses: What can they teach us about the severe acute respiratory syndrome? Revue scientifique et technique (Int. Off. Epizootics) 23 (2) (2004) 643–660.
[31] Roger Frutos, Jordi Serra-Cobo, Lucile Pinault, Marc Lopez Roig, Christian A Devaux, Emergence of bat-related betacoronaviruses: Hazard and risks, Front. Microbiol. 12 (2021) 591535.
[32] Zi-Wei Ye, Shuofeng Yuan, Kit-San Yuen, Sin-Yee Fung, Chi-Ping Chan, Dong-Yan Jin, Zoonotic origins of human Coronaviruses, Int. J. Biol. Sci. 16 (10) (2020) 1686.
[33] Gytis Dudas, Luiz Max Carvalho, Andrew Rambaut, Trevor Bedford, MERS-CoV spillover at the camel-human interface, elife 7 (2018) e31257.
[34] W Widagdo, Syriam Sooksawasdi Na Ayudhya, Gadissa B Hundie, Bart L Haagmans, Host determinants of MERS-CoV transmission and pathogenesis, Viruses 11 (3) (2019) 280.
[35] Roujian Lu, Xiang Zhao, Juan Li, Peihua Niu, Bo Yang, Honglong Wu, Wenling Wang, Hao Song, Baoying Huang, Na Zhu, et al., Genomic characterisation and epidemiology of 2019 novel Coronavirus: implications for virus origins and receptor binding, Lancet 395 (10224) (2020) 565–574.
[36] Shishir K Gupta, Rashmi Minocha, Prithivi Jung Thapa, Mugdha Srivastava, Thomas Dandekar, Role of the pangolin in origin of SARS-CoV-2: An evolutionary perspective, Int. J. Mol. Sci. 23 (16) (2022) 9115.
[37] Cody B. Jackson, Michael Farzan, Bing Chen, Hyeryun Choe, Mechanisms of SARS-CoV-2 entry into cells, Nat. Rev. Mol. Cell Biol. 23 (1) (2022) 3–20.
[38] Guangzhi Zhang, Bin Li, Dongwan Yoo, Tong Qin, Xiaodong Zhang, Yaxiong Jia, Shangjin Cui, Animal coronaviruses and SARS-CoV-2, Transbound. Emerg. Dis. 68 (3) (2021) 1097–1110.
[39] Gabriel Pasquarelli-do Nascimento, Heloisa Antoniella Braz-de Melo, Sara Socorro Faria, Igor de Oliveira Santos, Gary P Kobinger, Kelly Grace Magalhães, Hypercoagulopathy and adipose tissue exacerbated inflammation may explain higher mortality in COVID-19 patients with obesity, Front. Endocrinol. 11 (2020) 530.
[40] Ehdieh Khaledian, Sinem Ulusan, Jeffery Erickson, Stephen Fawcett, Michael C Letko, Shira L Broschat, Sequence determinants of human-cell entry identified in ACE2-independent bat Sarbecoviruses: A combined laboratory and computational network science approach, EBioMedicine 79 (2022) 103990.
[41] Nicola Decaro, Alessio Lorusso, Novel human Coronavirus (SARS-CoV-2): A lesson from animal Coronaviruses, Vet. Microbiol. 244 (2020) 108693.
[42] E. Paul J. Gibbs, Tara C. Anderson, et al., One world-one health and the global challenge of epidemic diseases of viral aetiology, Vet. Italiana 45 (1) (2009) 35–44.
[43] Henrik Lerner, Charlotte Berg, The concept of health in one health and some practical implications for research and education: What is one health? Infect. Ecol. Epidemiol. 5 (1) (2015) 25300.
[44] Delphine Destoumieux-Garzón, Patrick Mavingui, Gilles Boetsch, Jérôme Boissier, Frédéric Darriet, Priscilla Duboz, Clémentine Fritsch, Patrick Giraudoux, Frédérique Le Roux, Serge Morand, et al., The one health concept: 10 years old and a long road ahead, Front. Vet. Sci. (2018) 14.
[45] Michael J. Day, One health: The importance of companion animal vector-borne diseases, Parasites Vect. 4 (2011) 1–6.
[46] Jacqueline K. Owens, Systematic reviews: Brief overview of methods, limitations, and resources, Nurse Author Ed. 31 (3–4) (2021) 69–72.
[47] Nardus Mollentze, Deborah Keen, Uuriintuya Munkhbayar, Roman Biek, Daniel G Streicker, Variation in the ACE2 receptor has limited utility for SARS-CoV-2 host prediction, bioRxiv (2022).
[48] Maya Wardeh, Matthew Baylis, Marcus S.C. Blagrove, Predicting mammalian hosts in which novel Coronaviruses can be generated, Nat. Commun. 12 (1) (2021) 1–12.
[49] Daniel J Becker, Gregory F Albery, Anna R Sjodin, Timothée Poisot, Laura M Bergner, Binqi Chen, Lily E Cohen, Tad A Dallas, Evan A Eskew, Anna C Fagre, et al., Optimising predictive models to prioritise viral discovery in zoonotic reservoirs, Lancet Microbe (2022).
[50] Zheng Kou, Yi-Fan Huang, Ao Shen, Saeed Kosari, Xiang-Rong Liu, Xiao-Li Qiang, Prediction of pandemic risk for animal-origin coronavirus using a deep learning method, Infect. Dis. Poverty 10 (05) (2021) 62–70.
[51] Spyros Lytras, Joseph Hughes, Darren Martin, Phillip Swanepoel, Arné de Klerk, Rentia Lourens, Sergei L Kosakovsky Pond, Wei Xia, Xiaowei Jiang, David L Robertson, Exploring the natural origins of SARS-CoV-2 in the light of recombination, Genome Biol. Evol. 14 (2) (2022) evac018.
[52] Murali Aadhitya Magateshvaren Saras, L Ponoop Prasad Patro, Patil Pranita Uttamrao, Thenmalarchelvi Rathinavelan, Geographical distribution of SARS-CoV-2 amino acids mutations and the concomitant evolution of seven distinct clades in non-human hosts, Zoonoses Public Health (2022).
[53] Alexis Felipe Rojas-Cruz, Juan Carlos Gallego-Gómez, Clara Isabel Bermúdez-Santana, RNA structure-altering mutations underlying positive selection on spike protein reveal novel putative signatures to trace crossing host-species barriers in Betacoronavirus, RNA Biol. 19 (1) (2022) 1019–1044.
[54] Cecylia Severin Lupala, Vikash Kumar, Xiao-dong Su, Chun Wu, Haiguang Liu, Computational insights into differential interaction of mammalian angiotensin-converting enzyme 2 with the SARS-CoV-2 spike receptor binding domain, Comput. Biol. Med. 141 (2022) 105017.
[55] Madhusudan Rajendran, Gregory A. Babbitt, Persistent cross-species SARS-CoV-2 variant infectivity predicted via comparative molecular dynamics simulation, bioRxiv (2022).
[56] Ismail Celik, Trina E. Tallei, A computational comparative analysis of the binding mechanism of molnupiravir's active metabolite to RNA-dependent RNA polymerase of wild-type and delta subvariant AY. 4 of SARS-CoV-2, J. Cell. Biochem. 123 (4) (2022) 807–818.
[57] Seyed Ali Hemmati, Saeid Tabein, Insect protease inhibitors; Promising inhibitory compounds against SARS-CoV-2 main protease, Comput. Biol. Med. 142 (2022) 105228.
[58] Ping Chen, Jingfang Wang, Xintian Xu, Yuping Li, Yan Zhu, Xuan Li, Ming Li, Pei Hao, Molecular dynamic simulation analysis of SARS-CoV-2 spike mutations and evaluation of ACE2 from pets and wild animals for infection risk, Comput. Biol. Chem. 96 (2022) 107613.
[59] Abbas Khan, Sarfaraz Hussain, Sajjad Ahmad, Muhammad Suleman, Imrana Bukhari, Taimoor Khan, Farooq Rashid, Abul Kalam Azad, Muhammad Waseem, Wajid Khan, et al., Computational modelling of potentially emerging SARS-CoV-2 spike protein RBDs mutations with higher binding affinity towards ACE2: A structural modelling study, Comput. Biol. Med. 141 (2022) 105163.

[60] Xiaofeng Dong, Rebekah Penrice-Randal, Hannah Goldswain, Tessa Prince, Nadine Randle, Francisco J Salguero, Julia Tree, Ecaterina Vamos, Charlotte Nelson, Jordan Clark, et al., Analysis of SARS-CoV-2 known and novel subgenomic mRNAs in cell culture, animal model, and clinical samples using LeTRS, a bioinformatic tool to identify unique sequence identifiers, GigaScience 11 (2022).
[61] Rahul Kaushik, Naveen Kumar, Kam YJ Zhang, Pratiksha Srivastava, Sandeep Bhatia, Yashpal Singh Malik, A novel structure-based approach for identification of vertebrate susceptibility to SARS-CoV-2: Implications for future surveillance programmes, Environ. Res. 212 (2022) 113303.
[62] Sarah A Robinson, Matthew IJ Raybould, Constantin Schneider, Wing Ki Wong, Claire Marks, Charlotte M Deane, Epitope profiling using computational structural modelling demonstrated on Coronavirus-binding antibodies, PLoS Comput. Biol. 17 (12) (2021) e1009675.
[63] Min-Sheng Peng, Jian-Bo Li, Zheng-Fei Cai, Hang Liu, Xiaolu Tang, Ruochen Ying, Jia-Nan Zhang, Jia-Jun Tao, Ting-Ting Yin, Tao Zhang, et al., The high diversity of SARS-CoV-2-related Coronaviruses in Pangolins alerts potential ecological risks, Zool. Res. 42 (6) (2021) 834.
[64] Scott Burkholz, Suman Pokhrel, Benjamin R Kraemer, Daria Mochly-Rosen, Richard T Carback III, Tom Hodge, Paul Harris, Serban Ciotlos, Lu Wang, CV Herst, et al., Paired SARS-CoV-2 spike protein mutations observed during ongoing SARS-CoV-2 viral transfer from humans to minks and back to humans, Infect. Genet. Evol. 93 (2021) 104897.
[65] Sean B. King, Mona Singh, Comparative genomic analysis reveals varying levels of mammalian adaptation to Coronavirus infections, PLoS Comput. Biol. 17 (11) (2021) e1009560.
[66] Alan Brnabic, Lisa M. Hess, Systematic literature review of machine learning methods used in the analysis of real-world data for patient-provider decision making, BMC Med. Inform. Decis. Making 21 (1) (2021) 1–19.
[67] E Hope Weissler, Tristan Naumann, Tomas Andersson, Rajesh Ranganath, Olivier Elemento, Yuan Luo, Daniel F Freitag, James Benoit, Michael C Hughes, Faisal Khan, et al., The role of machine learning in clinical research: Transforming the future of evidence generation, Trials 22 (1) (2021) 1–15.
[68] Iqbal H. Sarker, Machine learning: Algorithms, real-world applications and research directions, SN Comput. Sci. 2 (3) (2021) 160.
[69] Gonzalo Giribet, Morphology should not be forgotten in the era of genomics–a phylogenetic perspective, Zoologischer Anzeiger-A J. Comparat. Zool. 256 (2015) 96–103.
[70] Aaron E Darling, Guillaume Jospin, Eric Lowe, Frederick A Matsen IV, Holly M Bik, Jonathan A Eisen, PhyloSift: Phylogenetic analysis of genomes and metagenomes, PeerJ 2 (2014) e243.
[71] Scott A. Hollingsworth, Ron O. Dror, Molecular dynamics simulation for all, Neuron 99 (6) (2018) 1129–1143.
[72] Mohammad Sufian Badar, Shazmeen Shamsi, Jawed Ahmed, Md Afshar Alam, Molecular dynamics simulations: Concept, methods, and applications, in: Transdisciplinarity, Springer, 2022, pp. 131–151.
[73] Shanika L Amarasinghe, Shian Su, Xueyi Dong, Luke Zappia, Matthew E Ritchie, Quentin Gouil, Opportunities and challenges in long-read sequencing data analysis, Genome Biol. 21 (1) (2020) 1–16.
[74] Seung Yon Rhee, Bioinformatics. Current limitations and insights for the future, Plant Physiol. 138 (2) (2005) 569–570.
[75] Iuliia Branco, Altino Choupina, Bioinformatics: New tools and applications in life science and personalized medicine, Appl. Microbiol. Biotechnol. 105 (2021) 937–951.
[76] Yinghui Liu, Gaowei Hu, Yuyan Wang, Wenlin Ren, Xiaomin Zhao, Fansen Ji, Yunkai Zhu, Fei Feng, Mingli Gong, Xiaohui Ju, et al., Functional and genetic analysis of viral receptor ACE2 orthologs reveals a broad potential host range of SARS-CoV-2, Proc. Natl. Acad. Sci. 118 (12) (2021) e2025373118.
[77] Xuesen Zhao, Danying Chen, Robert Szabla, Mei Zheng, Guoli Li, Pengcheng Du, Shuangli Zheng, Xinglin Li, Chuan Song, Rui Li, et al., Broad and differential animal angiotensin-converting enzyme 2 receptor usage by SARS-CoV-2, J. Virol. 94 (18) (2020) e00940–20.
[78] Yunjeong Kim, Natasha N Gaudreault, David A Meekins, Krishani D Perera, Dashzeveg Bold, Jessie D Trujillo, Igor Morozov, Chester D McDowell, Kyeong-Ok Chang, Juergen A Richt, Effects of spike mutations in SARS-CoV-2 variants of concern on human or animal ACE2-mediated virus entry and neutralization, Microbiol. Spectrum 10 (3) (2022) e01789–21.
[79] Cedric Tan, Su Datt Lam, Damien Richard, Christopher J Owen, Dorothea Berchtold, Christine Orengo, Meera Surendran Nair, Suresh V Kuchipudi, Vivek Kapur, Lucy van Dorp, et al., Transmission of SARS-CoV-2 from humans to animals and potential host adaptation, Nature Commun. 13 (1) (2022) 1–13.
[80] Alison E. Stout, Nicole M. André, Javier A. Jaimes, Jean K Millet, Gary R Whittaker, Coronaviruses in cats and other companion animals: Where does SARS-CoV-2/COVID-19 fit? Vet. Microbiol. 247 (2020) 108777.
[81] Tessa Prince, Shirley L. Smith, Alan D. Radford, Tom Solomon, Grant L. Hughes, Edward I. Patterson, SARS-CoV-2 infections in animals: Reservoirs for reverse zoonosis and models for study, Viruses 13 (3) (2021) http://dx.doi.org/10.3390/v13030494, URL https://www.mdpi.com/1999-4915/13/3/494.
[82] Kuldeep Dhama, Shailesh Kumar Patel, Khan Sharun, Mamta Pathak, Ruchi Tiwari, Mohd Iqbal Yatoo, Yashpal Singh Malik, Ranjit Sah, Ali A Rabaan, Parmod Kumar Panwar, et al., SARS-CoV-2 jumping the species barrier: Zoonotic lessons from SARS, MERS and recent advances to combat this pandemic virus, Travel Med. Infect. Dis. 37 (2020) 101830.
[83] Manuel Ruiz-Aravena, Clifton McKee, Amandine Gamble, Tamika Lunn, Aaron Morris, Celine E Snedden, Claude Kwe Yinda, Julia R Port, David W Buchholz, Yao Yu Yeo, et al., Ecology, evolution and spillover of Coronaviruses from bats, Nat. Rev. Microbiol. 20 (5) (2022) 299–314.
[84] Silviu O Petrovan, David C Aldridge, Harriet Bartlett, Andrew J Bladon, Hollie Booth, Steven Broad, Donald M Broom, Neil D Burgess, Sarah Cleaveland, Andrew A Cunningham, et al., Post COVID-19: A solution scan of options for preventing future zoonotic epidemics, Biol. Rev. 96 (6) (2021) 2694–2715.

3 CAPÍTULO 3

Machine learning to explain the zoonotic ability of mammals to transmit SARS-CoV-2.

Paula V. Silva[a,*], André de O. Brandão[a], Samuel Q. Lopes[a], Ariane C. B. da Silva [a], Marcelo de S. Balbino [a], Gisele Dantas[a], Matheus Libório[a] and Cristiane N. Nobre[a,*]

[a]*Pontifical Catholic University of Minas Gerais - PUC Minas - 500 Dom José Gaspar Street, Building 41, Coração Eucarístico, Belo Horizonte, MG 30535-901, Brazil*

[b]*Federal Center for Technological Education of Minas Gerais - 121, 19 de Novembro Street, Centro Norte, Timóteo, MG 35180-008, Brazil*

ARTICLE INFO

Keywords:
Bioinformatician
COVID-19
SARS-CoV-2
Machine Learning
Mammals

ABSTRACT

This article provides a comprehensive overview of the SARS-CoV-2 virus, delving into the infection process and focusing on studies that explore its connection with the angiotensin-converting enzyme 2 (ACE2). A key aspect of our research is the innovative use of Machine Learning techniques to investigate these complex interactions. Our database, which includes ecological information, life history, and biological characteristics of several mammal species, was used to run eight different ML algorithms (Naive Bayes, Decision Tree, Random Forest, XGBoost, AdaBoost, SVM, Logistic Regression, and MLP) to predict species susceptibility to SARS-CoV-2. The neural network emerged as the top-performing model, showcasing its potential in this field. Furthermore, the CSSE method (Agnostic Method of Counterfactual, Selected, and Social Explanations) was used to interpret the results and identify the most relevant biological characteristics in susceptibility to the virus. The results indicated that population density, longevity, social group size, litter frequency, and twilight activity are crucial determinants of a species' ability to act as a zoonotic host. The counterfactual analysis revealed that primates, in particular, showed high susceptibility to SARS-CoV-2. This research expands understanding of the dynamics of zoonotic transmission and offers a practical tool to identify and monitor potential hosts of SARS-CoV-2 and other emerging pathogens, thereby engaging the scientific community in the potential real-world applications of the findings. Counterfactual explanations provide a detailed view of how different biological characteristics influence susceptibility to the virus, further piquing the interest of the readers.

1. Introduction

The SARS-CoV-2 virus, which caused the COVID-19 pandemic that emerged in late 2019, has triggered an unprecedented global health crisis, prompting tireless efforts by researchers and healthcare professionals to unravel its biology and the mechanisms underlying its dissemination and pathogenicity (Gralinski & Menachery, 2020; Munir et al., 2020). Belonging to the Coronaviridae family, SARS-CoV-2 is a single-stranded RNA virus responsible for COVID-19, a disease that ranges from mild to severe manifestations, including pneumonia, acute respiratory failure, and death (Evans & Liu, 2021; Nasserie, Hittle, & Goodman, 2021). Its rapid spread is partly due to its notable transmissibility and asymptomatic incubation period, which makes early identification and control of new cases difficult (Murata et al., 2021).

SARS-CoV-2 infects cells by binding to angiotensin-converting enzyme 2 (ACE2) through the receptor-binding domain (RBD) of its spike (S) protein (Cevik, Bamford, & Ho, 2020; Praharaj et al., 2022). Upon virus entry, an uncontrolled inflammatory immune response occurs, which drives cytokine storm, aggressive inflammation, and collateral tissue damage due to the broad organotropism of SARS-CoV-2, leading to systemic failure (Trougakos et al., 2021). ACE2 is expressed in several cells found in lung, heart, kidney, and intestinal tissues. It plays a crucial role in viral infection, mediating the connection between the viral protein and the ACE2 domains (Li, Li, Zhang, & Wang, 2020).

The widespread presence of ACE2, combined with the high prevalence of SARS-CoV-2 in the human population, has explained several infections since the virus's emergence in 2019. In the context of spillover infections, humans

*Corresponding author
paula.silva.1138274@sga.pucminas.br (P.V. Silva); nobre@pucminas.br (C.N. Nobre)
ORCID(s):

transmit the SARS-CoV virus -2, causing infections in non-human animals. This situation poses a threat to both wildlife and domestic animals (Shi et al., 2020).

Furthermore, repeated infections may result in the establishment of new animal hosts, from which SARS-CoV-2 may pose a risk of secondary infection in humans, possibly through intermediate hosts (Guth, Visher, Boots, & Brook, 2019). This occurred in the wild in Denmark, where SARS-CoV-2 spread from humans to farm-raised minks (*Neovison vison*), resulting in a subsequent spillover of the SARS-CoV-2 variant from minks back to humans (Organization, 2020). It has also been observed in the laboratory in the case of two new human variants that managed to infect laboratory rats, overcoming the species barrier (Bao et al., 2020). This raises great concern due to the possibility of mutant strains emerging that could affect the host range, increase transmissibility between humans, reduce the effectiveness of neutralizing antibodies, and reduce the effectiveness of vaccines (Van Egeren et al., 2021; Volz et al., 2021).

Although the presence of ACE2 is the main point to consider whether or not an animal can transmit the SARS-CoV-2 virus, other characteristics can also be considered, such as the level of interaction with humans and the virus' ability to mutate (Silva & Nobre, 2024). Thus, in addition to the level of binding strength of ACE2 with SARS-CoV-2, the database used in this work, made available by Fischhoff, Castellanos, Rodrigues, Varsani, and Han (2021), brings together ecological information, life history, phylogenetic and biological characteristics of some databases.

Given the complexity of molecular interactions in this context, using machine learning (ML) techniques emerges as a promising approach. The application of ML algorithms can catalyze the discovery of new insights into the virus-host relationship, accelerate the search for targeted therapies, and predict viral mutations and their spread (Mollentze, Keen, Munkhbayar, Biek, & Streicker, 2022). Furthermore, the analysis of the virus's genomic sequences using AM makes it possible to identify relevant mutations and correlate them with characteristics such as transmissibility, virulence, and resistance to treatments, contributing to public health strategies (COVIDSurg Collaborative, 2021; Serna García, Al Khalaf, Invernici, Ceri, & Bernasconi, 2023).

However, although the use of Machine Learning models in healthcare is increasing, professionals need more intuition and explanation of their predictions to understand and trust the models. For this reason, the demand for interpretability methods that can provide insights into the prediction process of Machine Learning models (ElShawi, Sherif, Al-Mallah, & Sakr, 2021) is growing. Among the interpretability methods, we highlight counterfactual explanations, whose principle consists of identifying the minimum changes necessary in the input so that a contrastive output is obtained (Stepin, Alonso, Catala, & Pereira-Fariña, 2021).

Thus, the biology of the virus, the interaction with the ACE2 protein, and the application of Machine Learning techniques represent a multidisciplinary and promising approach to understanding and addressing the challenges of SARS-CoV-2 viral infection. This article aims to review recent advances in understanding the interaction between the SARS-CoV-2 virus and the ACE2 protein, focusing on the essential link for viral infection, and explore the use of Machine Learning algorithms in this context, highlighting its impact on the development of effective strategies against COVID-19.

To this end, we evaluated the performance of eight different machine learning algorithms (Naive Bayes, Decision Tree, Logistic Regression and AdaBoost Classifier, Random Forest, XGBClassifier, SVC, and MLP Classifier) to consider whether or not a species is susceptible to the virus. Using the best learning models obtained, Logistic Regression, and Neural Network, the most significant characteristics of the species that make them susceptible to the SARS-CoV-2 virus were evaluated. To interpret the Neural Network, a non-interpretable method, we use the *Agnostic Method of Counterfactual, Selected and Social Explanations (CSSE)* described in (de Sousa Balbino, Gálvez, & Nobre, 2023).

Deepening our understanding of virus biology and its interactions is critical to meeting the challenges posed by this global pandemic and preparing for future public health emergencies. The knowledge generated by these investigations can provide crucial insights into disease prevention and effective treatment, better preparing us to deal with present and future public health crises.

2. Theoretical Framework

2.1. Mechanism of viral entry and transmission

The invasion of SARS-CoV-2 into the host cell occurs in two main phases: identification of the cell's receptor and fusion of the viral membrane with the cell (Gupta, Minocha, PJ, Srivastava, & Dandekar, 2022). SARS-CoV-2 is composed of four structural proteins: envelope (E), nucleocapsid (N), membrane (M) and spike (S) (Jackson, Farzan, Chen, & Choe, 2022). The S protein, which facilitates virus entry, is formed by three functional domains: *ectodomain*,

transmembrane anchor and *short cytoplasmic tail*. The ectodomain is composed of the S1 and S2 subunits, with the S1 subunit containing the receptor-binding domain (RBD) that binds to receptors on the cell surface. It is known that several ACE2 receptors are expressed in many cells of different human tissues, and the interaction between the RBD of the SARS-CoV-2 S protein and ACE2 is crucial for the entry of the virus into the host cell (Zhang et al., 2021) . After receptor recognition, viral binding and membrane fusion, SARS-CoV-2 releases its genetic material into the cell, and viral proteins are synthesized in the host's Endoplasmic Reticulum for viral replication and the release of viral particles occurs through the Golgi complex DNA (Pasquarelli-do Nascimento et al., 2020).

Furthermore, the interaction between SARS-CoV and SARS-CoV-2 S proteins and host ACE2 involves more than 15 contact points. Changes in many of these residues can significantly influence viral entry (Khaledian et al., 2022). Coronaviruses have a high mutation rate, which can result in changes in their antigenic profile, tissue tropism and host range. This occurs through two main mechanisms: *antigenic drift*, which leads to the incorporation of wrong nucleotides during replication cycles, and *recombination*, which allows different strains of coronavirus to synthesize a hybrid RNA. This process can lead to adaptation to other species and an increase in the pathogenicity of viruses (Decaro & Lorusso, 2020).

2.2. Method of Interpretability of Machine Learning Models

It has become increasingly common to use machine learning systems in high-risk problems that have a profound impact on human life and society. The models have provided support for issues where a correct decision is essential (Rudin, 2019). From this scenario, there was a movement in the scientific community in search of increasingly better models from the point of view of predictive performance. Much research has been directed in this direction and relevant results have been obtained, especially through so-called "black box" (Abdul, Vermeulen, Wang, Lim, & Kankanhalli, 2018) methods.

However, improved predictive performance has been achieved at the expense of greater complexity and less transparency in (Du, Liu, & Hu, 2019) decisions. Rudin (2019) and El Shawi, Sherif, Al-Mallah, and Sakr (2019) also highlight this *trade-off* between performance and transparency and highlight how limitations in the interpretability of decisions undermine confidence in the model and its consequent use in practice.

Furthermore, in many domains, having a model with high predictive performance is only a partial solution to the problem. In high-risk scenarios, such as the doctor, the decision maker feels insecure without explaining the results of the (Carvalho, Pereira, & Cardoso, 2019; Tjoa & Guan, 2020) model. Therefore, it is necessary to present elements that allow developers and users to understand the system's decisions better, increasing confidence in the results and, when required, allowing incorrect decisions to be noticed.

In this context, questions arise regarding quality measures of machine learning models based only on predictive performance. For example, a model used in a banking system that suggests whether or not to lend a loan to a particular customer should aim to reduce the default rate and not discriminate against anyone based on race or where they live. In general, machine learning techniques are only concerned with optimizing the loan default metric and do not care about other factors, such as discrimination. Therefore, it is necessary to include other essential factors in evaluating machine learning techniques, interpretability being one of them (Karim, Mishra, Newton, & Sattar, 2018).

Therefore, XAI (Explainable Artificial Intelligence), a field of study focusing on the interpretability of ML systems, has gained the scientific community's attention. The objective is to contribute to creating methods that allow the models to be interpreted, preserving high levels of predictive performance (Adadi & Berrada, 2018; Miller, 2019).

Explanation is a way in which an observer can gain understanding. Therefore, one way to increase the level of interpretability of systems is to provide appropriate explanations to users (Miller, 2019). However, research points to inadequacies in current interpretability methods regarding the ability to communicate with end users (Du et al., 2019; Karim et al., 2018; Miller, 2019; Molnar, 2020).

In this sense, Miller (2019) states that if we want to design and implement intelligent agents that can explain people, it is essential to understand how humans define, generate, select, evaluate, and present explanations. However, the author highlights that most research and practices use researchers' intuitions about what is considered a "good" explanation. The solution is to expand studies beyond computational issues. The author adds that in the fields of philosophy, psychology/cognitive sciences, and social psychology, a vast and mature number of works study precisely these topics.

Given this, based on studies that consider computational and social science aspects, Miller (2019) lists three points that should be considered to build a genuinely explainable AI: contrastive, selected, and social explanations.

According to Biran and Cotton (2017), the level of interpretability of a model is related to the degree to which an observer can understand its decisions. Explanation is precisely one way in which an observer can gain such understanding. In healthcare, interpretability can help professionals understand the logic behind predictions, allowing them to accept or reject the prediction (ElShawi et al., 2021).

In classification problems, interpretable machine learning aims to uncover the reasons behind predictions made by uninterpretable models. For this purpose, one of the most valuable methods is using counterfactual explanations. A counterfactual explanation reveals what would have to be different about an instance to change its classification (Guidotti, 2022). For example, for an individual who received a positive diagnosis for a disease, the counterfactual explanation presents what should have been done differently that would have resulted in a negative diagnosis.

The *Agnostic Method of Counterfactual, Selected, and Social Explanations* (CSSE), a counterfactual explanation method, can generate local explanations for classification models using a genetic algorithm. It offers counterfactual explanations for learning models, presents explanations with diversity, without verbosity, and allows the user to restrict the attributes in the explanation (de Sousa Balbino et al., 2023).

2.3. Discriminant Analysis

Discriminant Analysis is a multivariate statistical method that seeks to minimize the variance of elements within a class and, at the same time, maximize the distance between the means of elements from two or more courses. These characteristics make Discriminant Analysis a valuable method for classifying elements, reducing dimensionality, and identifying discriminating factors in (Xanthopoulos et al., 2013) classification. Discriminant Analysis has also been very useful for validating classifications made by other machine learning methods (Riveiro-Valiño, Álvarez-López, & Marey-Pérez, 2009; Zebardast, Mazaherian, Rahmani, & Nouri, 2024), especially in the areas of public and clinical health (Dhamnetiya, Goel, Jha, Shalini, & Bhattacharyya, 2022).

The model of Discriminant Analysis is constructed on the solid foundation of Wilk's Lambda criterion and F statistics. These statistical tools play a crucial role in identifying the most discriminating variables and determining which variables should be included in the model. Understanding and applying these principles is key to mastering Discriminant Analysis.

3. Related Work

Research into mammals' zoonotic capacity to transmit SARS-CoV-2 has become an issue of extreme relevance amid the pandemic caused by this virus. This work seeks to address this issue by applying Machine Learning techniques to predict the potential for zoonotic virus transmission in different species of mammals. To contextualize our approach, reviewing previous work exploring related themes is essential. This section presents a critical analysis of prior studies that focused on predictions using computational methods.

Rodrigues et al. (2020) investigated the structural properties of several ACE2 orthologs linked to the Spike (S) protein of SARS-CoV-2, observing that species known not to be susceptible to SARS-CoV-2 infection present non-conservative mutations in several residues of amino acids of ACE2, which impairs essential polar and charged contacts with the viral S protein. Their models, created using the MODELLER 9.24 program and custom Python scripts (Available at: https://github.com/joaorodrigues//ace2-animal-models/), also made it possible to predict mutations that increase affinity and that can be used to design ACE2 variants for therapeutic purposes.

Ahmed, Hasan, Siddiki, and Islam (2021) used ACE2 and TMPRSS2 protein sequences from species common in the South Asian region to predict hosts of SARS-CoV-2. The authors performed homology modeling to simulate the interaction between ACE2 and S using the MODELLER software built into CHIMERA. Their results point to the Pangolin as an intermediate host and cows, buffaloes, goats, and sheep as having a high potential for infection.

In the study by Damas et al. (2020), the authors studied the conservation of ACE2 and its potential to be used as a receptor by SARS-CoV-2. To do this, they assigned a 5-category binding score based on the conservation properties of 25 amino acids essential for the binding between ACE2 and S. In their dataset, which has 410 species, 18 are in the very high category; among these are Old World primates and Great Apes.

Liu et al. (2020) sought to predict the interaction between ACE2 from various species and the S protein of SARS-CoV-2 through systematic comparison and analysis. Their results indicated three possible intermediate hosts: pangolins, snakes, and turtles. Luan, Jin, Lu, and Zhang (2020) When comparing the primary amino acids in ACE2 from different species to determine their ability to bind to the S protein, they found that snakes and turtles are not intermediate hosts of SARS-CoV-2.

X. Huang, Zhang, Pearce, Omenn, and Zhang (2020) developed a computational pipeline to identify potential intermediate hosts by modeling the binding affinity between the ACE2 protein of potential intermediate hosts and the S protein of SARS-CoV-2. Using this pipeline, they found 96 mammals permissive to SARS-CoV-2, including primates, rodents, and carnivores.

In Lam et al. (2020)'s work, researchers wanted to predict the risks of animals becoming infected with the SARS-CoV-2 virus. To do this, they modeled ACE2-S complexes and calculated changes in the energy of the complex caused by mutations in each species. Thus, they concluded that SARS-CoV-2 can infect many mammals but few fish, birds, and reptiles.

Han, Schmidt, Bowden, and Drake (2015) used a Machine Learning algorithm, Boosted Regression Trees (BRT), on a dataset with biological, ecological, and life history characteristics of rodents that carry numerous zoonotic pathogens to identify species with high probabilities of harboring undiscovered pathogens. Their model predicted rodent status with 90% accuracy, identifying rodent species that may be new zoonotic reservoirs and regions where new emerging pathogens are more likely. It also described trait profiles that distinguish reservoirs from non-reservoirs.

Yang and Han (2018) sought to identify intrinsic characteristics capable of predicting which ticks of the genus *Ixodes* are confirmed or suspected to be vectors of zoonotic pathogens. To do this, they used a Machine Learning algorithm, Generalized Boosted Regression, which predicted with 91% accuracy and identified 14 species with a high probability of transmitting infections.

Han et al. (2016) used the Machine Learning algorithm, Generalized Boosted Regression, to characterize filovirus-positive bat species with 87% accuracy. They reported a profile of intrinsic characteristics that discriminate host from non-host animals and identified several species most likely to be positive for filoviruses based on similarity to bats known to be positive for these viruses.

Han et al. (2015), Han et al. (2016) and Yang and Han (2018) used Machine Learning techniques to predict animals capable of transmitting some pathogens, close to what we did in this work. Ahmed et al. (2021); X. Huang et al. (2020); Lam et al. (2020); Rodrigues et al. (2020) focused on modeling the binding between host ACE2 and the viral S protein to simulate the binding of SARS-CoV-2.

Thus, although all the works cited in this section deal with predictions, they all take different approaches. All explore the species' biological, structural, and epidemiological characteristics. Our work uses a counterfactual explanation method to offer a more explanatory perspective for a greater understanding of viral transmission.

4. Materials and Methods

4.1. Database description

The database used in this work was made available by Fischhoff et al. (2021), who also analyzed the zoonotic capacity of mammals. This database was chosen for its scope and detail, which is essential for applying machine learning algorithms to predict susceptibility to SARS-CoV-2. The database has ACE2 sequences from NCBI GenBank and MEROPS and the results of binding simulations with the SARS-CoV-2 Spike protein using the HADDOCK software.

HADDOCK (High Ambiguity Driven biomolecular DOCKing) is a software for modeling biomolecular complexes. It is widely used in the scientific community to study molecular interactions, such as the binding between an enzyme and its substrate or the formation of protein complexes (van Zundert & Bonvin, 2014).

The database also has ecological, life history, phylogenetic, and biological characteristics data from some repositories described below: AnAge (De Magalhaes, Costa, & J, 2009), Amniote Life History Database (ALDH) (Myhrvold et al., 2015), EltonTraits (Wilman et al., 2014) and PanTHERIA (Jones et al., 2009). The database has 73 attributes and 113 instances. Table 1 describes the attributes and their origins. This database had already been classified a priori and was used to train the Machine Learning algorithms. The authors also made available a database containing 5400 that were not classified, which we classified using the best learning model obtained.

Below is a brief description of each repository:

- AnAge (De Magalhaes et al., 2009): This database compiles life history parameters, focusing on senescence, for a wide range of taxa, mainly animals but also including some plants and fungi. Parameters collected include maximum longevity, age at sexual maturity for males and females, metabolic rates, litter sizes, gestation/incubation periods, adult mass, and litters per year.

- Amniote Life History Database (ALDH) (Myhrvold et al., 2015): This is a database that allows you to carry out comparative analyses of birds, mammals, and reptiles. It also contains information about animals' life histories.
- EltonTraits (Wilman et al., 2014): This global species-level database contains critical attributes for all 9993 bird species and 5400 extant and recently extinct mammal species. Attributes include diet, foraging stratum, activity time, and body size.
- PanTHERIA (Jones et al., 2009): PanTHERIA is a species-level database compiled to analyze all known extant and recently extinct mammals' life history, ecology, and geography. It contains over 100,000 lines of biological data for extant and recently extinct mammal species.

Table 1
Sources of attributes present in the Fischhoff et al. (2021) dataset

Source	Attribute
EltonTraits	ForStrat.ground, ForStrat.understory, ForStrat.arboreal, ForStrat.arboreal, ForStrat.aerial, ForStrat.marine, ForStrat_terrestrial, ForStrat_aquatic, Activity.Nocturnal, Activity.Crepuscular, Activity.Diurnal, diet_breadth
ALHD	femal_maturity_d, male_maturity_d, weaning_d, development_d, log_litterclutch_size_n, litters_or_clutches_per_y, log _inter_litterbirth_interval_y, log_birthhatching_weight_g, log_weaning_weight_g, log_adult_body_mass_g, longevity_y, log_female_body_mass_g, log_male_body_mass_g, adult_svl_cm
AnAge	infantMortalityRate_per_year, mortalityRateDoublingTime_y, metabolicRate_W, temperature_K
PanTHERIA	AgeatEyeOpening, GestationLen, SocialGrpSize, TeatNumber, TrophicLevel, WeaningHeadBodyLen, MaxLat, MinLat, MidRangeLat, MaxLong, MinLong, MidRangeLong, HuPopDen_Change, Precip_Mean, Temp_Mean_01degC, AET_Mean, PET_Mean, DispersalAge, HomeRange_km2, HomeRange_Indiv_km2, PopulationDensity, PopulationGrpSize, HuPopDen_Min, HuPopDen_Mean, HuPopDen_5p NeonateHeadBodyLen
Fischhoff et al. (2021)	tnc_ecoregion_breadth, mass_specific_production, log_range_size, AA_83_y, AA_30_negative, log_WOS_hits_synonyms

4.2. Data Preprocessing

To guarantee data quality, during the database pre-processing stage, techniques were explored that include:

1. **Class categorization:** The class of instances in the database is represented by continuous values, indicating the strength of connection of each species. A limit was established to transform these values into discrete numbers, categorizing the instances into two classes: 0 (weak connection) and 1 (strong connection). The decision was based on the "binary_haddock_score" attribute, classifying values lower than -129 as 1 and the rest as 0. This value is between two HADDOCK scores: the domestic cat *(Felis catus)*, which is currently the species with the weakest predicted link between animals with confirmed conspecific transmission (Bosco-Lauth et al., 2020), and the pig/boar *(Sus scrofa)*, which shows the most vital predicted link between the species for which experimental inoculation failed to cause detectable infection (Shi et al., 2020). The -129 threshold was also adopted by Fischhoff et al. (2021).
2. **Removal of outlier values:** Operations were performed to remove values considered outliers, aiming to improve statistical accuracy and machine learning models. The identification criteria were values that exceeded three times the standard deviation of the attribute. Most of the data lies within the range of 3 standard deviations from

the mean. Therefore, 432 outliers were identified and had their values changed to missing values and subsequently imputed during the processing of missing data by the `KNNImputer` algorithm.

3. **Normalization:** Since the database contained attributes with very different variation scales, the `MinMaxScaler` class from the Scikit-learn library was used to normalize the data, scaling each value of each attribute to a range specific between 0 and 1.

4. **Missing Data Treatment:** To deal with a large number of missing values, data imputation was performed using the algorithm `k-NearestNeighbor` `KNNImputer` from the Scikit-learn library, using the average of the five neighboring values closest for each missing value. However, attributes that were more than 90% missing were removed (haddock_score_mean, infantilityMortalityRate_per_year, mortalityRateDoublingTime_y, log_DispersalAge_d, log_PopulationGrpSize, X13.3_WeaningHeadBodyLen_mm, e X24.1_TeatNumber)

5. **Dimensionality reduction:** Seeking to eliminate unrepresentative instances and attributes (irrelevant, noisy and/or redundant), in terms of learning, the ant colony method (Dorigo, Birattari, & Stutzle, 2006) was used. Dimensionality reduction, performed in databases through Instance Selection (SI) and Attribute Selection (SA), aims to obtain a subset of the input data that is capable of increasing the performance metrics of trained ML models , by eliminating irrelevant instances and attributes that could be harmful or misleading to the algorithm that learns a model. In this way, the quality of the selected data would lead to high quality results and reduced computational training costs, achieving a truly representative sample with a minimum size. Thus, 39 attributes were selected from the initial 73 and 64 instances from the 113 that made up the total base[1]. The selected attributes were: ForStrat_terrestrial, ForStrat_aquatic, Activity.Nocturnal, Activity.Crepuscular, Activity.Diurnal, female_maturity_d, weaning_d, litters_or_clutches_per_y, longevity_y, adult_svl_cm, development_d, log_litterclutch_size_n, log_birthhatching_weight_g, log_weaning_weight_g, log_male_body_mass_g, log_range_size, X2.1_AgeatEyeOpening_d, X9.1_GestationLen_d, Afrosoricida, Didelphimorphia, Erinaceomorpha, Macroscelidea, X10.2_SocialGrpSize, X6.2_TrophicLevel, X26.2_GR_MaxLat_dd, X26.4_GR_MidRangeLat_dd, X26.5_GR_MaxLong_dd, Primates, Proboscidea, X27.4_HuPopDen_Change, X30.1_AET_Mean_mm, X30.2_PET_Mean_mm, Rodentia, Soricomorpha, nchar, log_HomeRange_km2, log_PopulationDensity_n.km2, log_NeonateHeadBodyLen_mm, bin_haddock_score.

6. **Separation of the set for training, validation and testing:** Of the data set, 10% was reserved for testing, resulting in 7 instances, plus the 49 instances that the ant colony did not select for creating the models of learning. That is, the test set contains 56 instances. The 10-fold cross-validation method was used for the remaining 90%. The training database was then divided as follows: 58 instances of class 1 (susceptible to the virus) and 55 instances of class 0 (susceptible to the virus).

4.3. Description of Methods

Eight machine-learning algorithms were run to predict whether or not the species were susceptible to the virus, examining the database attributes using different strategies. They were: *Naive Bayes, Decision Tree, Random Forest, XGBoost, AdaBoost, SVM, Logistic Regression* and *MLP*. Table 2 presents the hyperparameters used in each algorithm, adjusted with the Random Search (Bergstra & Bengio, 2012) method.

To evaluate the quality of the tested models, we use the metrics *Precision*[2], *Recall*[3], and *F-measure*[4].

4.4. Interpretability of Machine Learning models

To interpret the results obtained from the best model obtained by Machine Learning algorithms, we use the *Agnostic Method of Counterfactual, Selected, and Social Explanations* (CSSE) (de Sousa Balbino et al., 2023). CSSE is a counterfactual explanation method, which reveals which attribute should be different so that the animal could no longer be susceptible to SARS-CoV-2.

[1] Instructions for using the python library for SI and SA with ant colonies can be obtained at `https://test.pypi.org/project/antcolony-is/`

[2] $\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$

[3] $\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$

[4] $\text{F} - \text{Measure} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision+Recall}}$

Table 2
Control Parameters applied to learning algorithms.

Algorithm	Control Parameters	Values
Naive Bayes	priors	'None'
	var_smoothing	'1e-09'
Decision Tree	splitter	'best'
	max_features	'sqrt'
	max_depth	10
	criterion	'entropy'
Random Forest	n_estimators	30
	min_samples_leaf	5
	max_features	0.2
	max_depth	5
	criterion	'entropy'
XGBClassifier	n_estimators	500
	max_depth	5
	learning_rate	0.05
AdaBoostClassifier	n_estimators	90
	learning_rate	0.01
SVC	kernel	'linear'
	gamma	1
	C	10
LogisticRegression	solver	'liblinear'
	penalty	'l2'
	multi_class	'ovr'
	max_iter	4000
	class_weight	'balanced'
	C	10
MLPClassifier	solver	'adam'
	max_iter	6000
	learning_rate	'constant'
	hidden_layer_sizes	(15, 15)
	batch_size	32
	alpha	0.0001
	activation	'relu'

5. Results and Discussion

For all machine learning algorithms investigated, evaluation metrics for both classes (susceptible or non-susceptible) were calculated. Despite the small amount of data in the sample space, the neural network and logistic regression obtained the best results and were chosen to be investigated for interpretability.

This section presents the results of the learning algorithms, explanations of the knowledge acquired by them, and the result of the discriminant analysis carried out based on the classification of the 5400 instances based on the best models obtained.

5.1. Performance of algorithms regarding SARS-CoV-2 transmission susceptibility

Figure 1 presents the results obtained from each of the learning algorithms investigated. Evaluations were conducted for each 'susceptible' and 'non-susceptible' class, observing the precision, recall, and F-measure metrics. It is observed that the recall for the 'susceptible' class reached a rate of 100% with the SVM, logistic regression, and Neural Network algorithms. In other words, of all the test instances with this classification, these models got 100% of the cases correct. As for accuracy, the Naive Bayes algorithm managed to achieve a 100% accuracy rate; however, it only got a 67% recall. All other classifiers had a performance ranging from 67% to 75%, except for the Decision Tree and Random Forest algorithms, which had the worst performances, 33% and 50%, respectively. Thus, considering the harmonic mean of recall and precision, given by the F-measure metric, for the 'susceptible' class, we have that the two best algorithms were logistic regression and neural network, reaching a rate of 86%.

For the 'non-susceptible' class, we observed an opposite behavior regarding precision and recall; that is, logistic regression and the Neural Network obtained a precision of 100%, but a recall of 75%. Thus, Fmeasure followed the pattern of the 'susceptible' class, that is, a rate of 86% obtained by the logistic regression and neural network algorithms. The Decision Tree and Random Forest algorithms also received the worst performances.

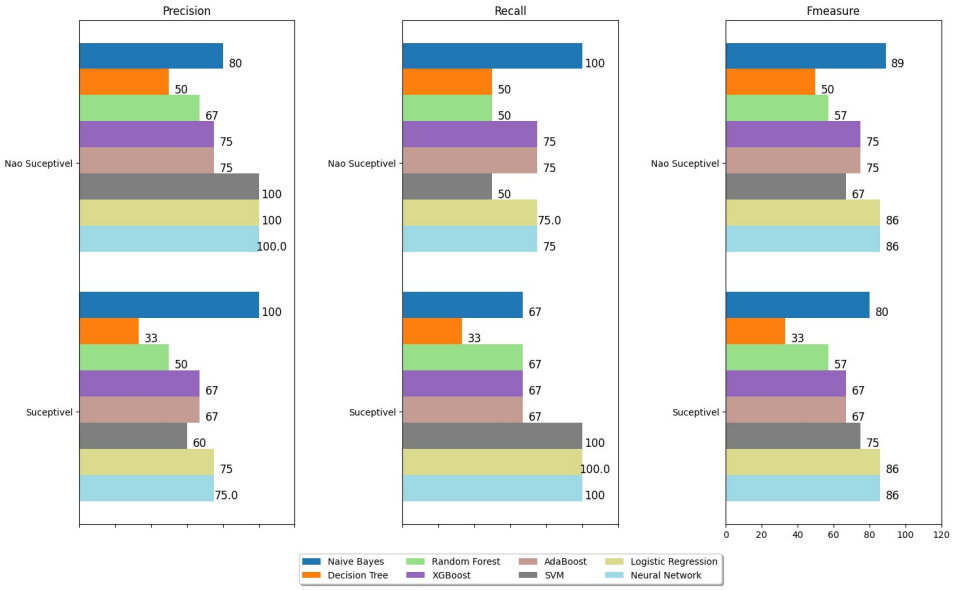


Figure 1: Results obtained with the machine learning algorithms investigated to evaluate susceptible and non-susceptible species to the SARS-CoV-2 virus

5.2. Analysis of the importance of attributes for the transmission of SARS-CoV-2

Based on the previous results, we sought to identify which biological characteristics make a species susceptible or not to the virus. In other words, in addition to pursuing a high predictive capacity, we also want to identify which characteristics make the mammal a suitable transmitter of SARS-CoV-2. In this sense, we interpreted the two best models in this classification task. In other words, we investigated what logistic regression and the Neural Network considered relevant.

Logistic regression is an interpretable method; therefore, it was possible to identify the relevant characteristics directly. Each attribute's importance for separating the two classes of interest is plotted. Figure 2(a) shows the attributes identified by Logistic Regression, ordered in order of importance.

On the other hand, the Neural Network is not a naturally interpretable method. For this, we use the CSSE (de Sousa Balbino et al., 2023), a counterfactual explanation method, which evaluates the relevance of the attributes that contributed most to the classification. The characteristics are indicated in Figure 2(b).

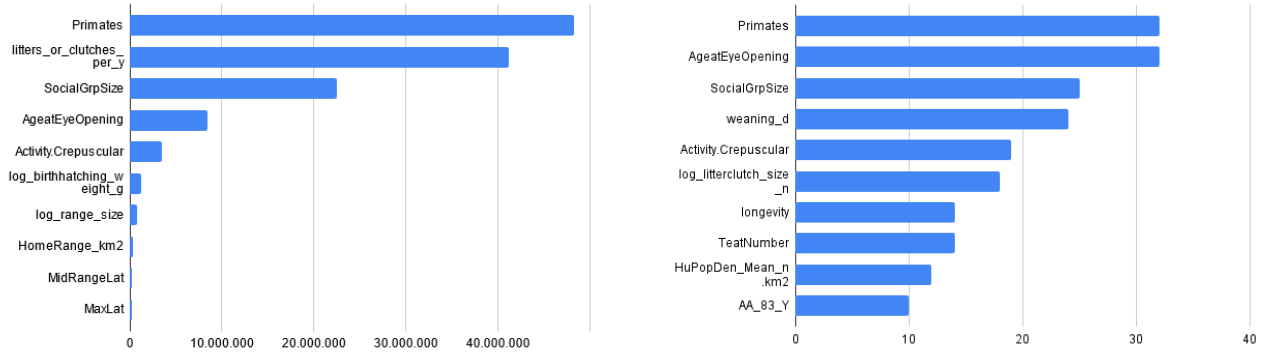
Thus, from these two results, we have the following explanations for some characteristics considered important by these two learning methods.

1. **Primates**:

   Primates, particularly great apes such as chimpanzees and gorillas, are highly susceptible to transmitting diseases to humans. This is due to its evolutionary proximity to us. Disease transmission among wild primates is more frequent between species that are geographically close and closely related (Pedersen & Davies, 2009).

   An important aspect contributing to this transmission is the primates' social structure. One study showed that social structure is intrinsically linked to parasite load. This suggests that contact between group members, rather than group size itself, is associated with parasite transmission.

   Transmission of viral diseases among non-human primates occurs mainly through direct or indirect contact with infected blood and other bodily fluids. A notable example of this is HIV, which is caused by a virus that jumped from wild primates to humans through infected bodily fluids. Therefore, the close interaction and evolutionary



(a) Features identified by Logistic Regression

(b) Features identified by the Neural Network, using CSSE

Figure 2: Most relevant characteristics in SARS-CoV-2 infection, according to Logistic Regression and Neural Network, using CSSE.

proximity with primates may have significant implications for the transmission of viral diseases to humans, highlighting the urgent need for further research in this area.

2. **Social group size**

   Social group size in mammalian species can significantly impact the transmission of viral diseases. Reservoir species with low population densities may support more minor, less genetically diverse viral populations that harbor fewer mutations capable of infecting humans. In contrast, reservoir species with short lifespans and rapid population turnover can support more significant, more genetically diverse viral populations, including variants capable of infecting humans (Lucatelli, Mariano-Neto, & Japyassú, 2021).

   Mortality from the disease can induce decreases in population density, leading to fewer associations. Furthermore, adaptive behavioral responses, by which animals identify infected individuals (and whether they are infected), can trigger quarantine or self-isolation behaviors that reduce encounters between infected and healthy individuals.

   A general pattern has been reported suggesting that smaller mammal species that utilize home ranges more intensively experience a more significant risk of invasion by environmentally transmitted macroparasites. On the other hand, larger hosts that exhibit a high degree of social group living could be more easily invaded by directly transmitted microparasites.

   Studies show that social structure, and not necessarily group size, is connected to parasite load. However, the size of the social group can significantly affect the spread of infections in animal societies. Larger social groups can facilitate the spread of infections, while smaller groups can limit it. Therefore, social structure and group size play a crucial role in the transmission of viral diseases between mammalian species.

3. **Weaning duration**

   Early weaning may shape long-term immune and metabolic responses in mammals such as pigs, which may have implications for disease susceptibility (Fardisi et al., 2023).

4. **Litter size**

   There is evidence to suggest that animals that mature early and have frequent litters tend to invest less in some immunological defenses, making them more suitable hosts for pathogens. Additionally, one study showed that litter size is intrinsically related to social group size. This is relevant because social group size is associated with parasite transmission. Therefore, it can be inferred that litter size can indirectly influence parasite transmission through its impact on social group size (Ostfeld et al., 2014).

   Incubation time and clutch size have been linked to the species' immune response. At the same time, body mass may be a proxy for size-scale life history traits such as fecundity, metabolic requirements, and age at first reproduction.(Z. Y. Huang et al., 2013).

5. **Longevity**

   Reservoir species with low population densities may support more minor, less genetically diverse viral populations that harbor fewer mutations capable of infecting humans. Alternatively, reservoir species with short lifespans and rapid population turnover may support more significant, more genetically diverse viral populations that – by chance – include variants capable of infecting humans (Nuismer, Basinski, Schreiner, Whitlock, & Remien, 2022).

   Life history theory suggests that short-lived (generally relatively small) species invest more in reproduction and less in immune defenses. Smaller species may, therefore be more susceptible to infection (Wang et al., 2021).

   A study showed that reservoir host longevity, viral tolerance, and constitutive immunity impact the evolution of viral traits that cause virulence after spillover to humans (Brook, Rozins, Guth, & Boots, 2023).

6. **Human population density within species range**

   Human population density and urbanization can increase the risk of transmission of zoonotic diseases. With more people living in dense conditions, contact between individuals becomes more frequent, facilitating the transmission of diseases.

   A recent study highlighted that meteorological factors, along with population density and living conditions—particularly in urban and semi-urban areas—play a crucial role in the intensity, evolution, and spread of SARS-CoV-2. This suggests that environmental and social conditions can significantly impact the spread of viral diseases. Mishra, Mishra, and Arora (2021).

   Furthermore, human population density can influence the transmission of viral diseases from mammals to humans. For example, virus transmission from range-shifting mammalian species is predicted to concentrate in areas with high human population density, such as parts of Asia and Africa. This suggests that interactions between humans and animals in densely populated regions may be a key factor in the spread of zoonotic diseases.

   It is important to highlight that providing detailed explanations of all the characteristics identified through the interpretability of the learning models was not possible. Despite efforts to offer complete and comprehensive documentation, we encountered limitations due to the lack of information available in specialized literature.

   However, in the work of Silva and Nobre (2024), which conducted a systematic review of the literature to identify relevant characteristics in animals that transmit SARS-CoV-2, some of the characteristics found in this study were cited, such as interaction with humans, longevity, and social group size.

5.3. Discriminant Analysis to evaluate the quality of the classification performed by the Neural Network

As mentioned, this work relied on two databases. The first base was already labeled with the species being susceptible or not to the virus. This base was used to create the machine learning models described in the previous sections. It contains 73 attributes and 113 instances. All attributes were presented in Table 1. The second database has 5400 instances, and has the same attributes described in Table 1, but without the class label. In other words, for this database, we do not know whether or not the species is susceptible to the virus.

Using the best classification model obtained, we classified the 5400 unlabeled instances. As we have two models that had equivalent behavior, we selected the neural network for this classification task. In this way, each of the 5400 instances were labeled, obtaining 3446 instances classified in class 1 = 'susceptible to the virus' and 1954 instances in class 0 = 'not susceptible to the virus'. From these already classified instances, we apply Discriminant Analysis to validate the quality of this classification.

As presented in Section 2.3, a discriminant analysis is statistically adequate when the proportion of observations classified into their original classes is close to 1.00, the Lambda value is greater than 0.015, the eigenvalues and canonical correlations are close to 1.00.

In the tests carried out, the value found for correctly classified observations was 0.89 (the sum of the hits in classes 0 and 1), as presented in Table 3, and Table 4 guarantees that the results found are appropriate, validating the classification carried out by the Neural Network.

Table 3
Discriminant analysis classification results for the 5400 instances.

	Class	0	1	Total
Original count	0	1809	145	1954
	1	408	3038	3446
%	0	92,6	7,4	100,0
	1	11,8	88,2	100,0

Table 4
Metrics for evaluating discriminant analysis.

Teste de funções	Lambda de Wilks	Qui-Quadrado	df	Sig	Autovalor	Correlação canônia
1	,440	4428,614	21	< ,001	1,275	,749

5.4. Machine Learning Model Generalization

Using machine learning to predict susceptibility based on ecological, life history, and biological traits holds great potential for application to zoonotic diseases beyond SARS-CoV-2. This approach is especially valuable in situations involving complex interactions between species and pathogens, particularly when susceptibility is linked to factors such as the presence of specific cellular receptors, like ACE2, as described in this work. For other zoonotic viruses that utilize similar mechanisms to enter host cells, this methodology can be adapted and applied. Machine learning-based predictive models can help identify species most likely to host and transmit emerging pathogens, thus supporting surveillance efforts and the development of more effective control strategies (Heesterbeek et al., 2015).

The ecological and biological traits selected in the model—such as population density, longevity, social group size, and litter frequency—are variables that can be useful for predicting susceptibility to other zoonotic viruses. These traits, along with factors like proximity to humans and host ecology, serve as robust predictors of susceptibility, making them adaptable for studies on other zoonoses, such as avian influenza or viruses within the coronavirus family. The influence of these traits on pathogen transmission and hosting potential makes them critical candidates for inclusion in new models aimed at predicting zoonotic disease risks (Karesh et al., 2012; Plowright et al., 2017).

6. Conclusion

This article adopted a multidisciplinary approach to understand and address the challenges of SARS-CoV-2 viral infection, combining the biology of the virus, the interaction with the ACE2 protein and the application of Machine Learning techniques. The objective was to review recent advances in understanding the interaction between the SARS-CoV-2 virus and the ACE2 protein, which is fundamental for viral infection. Furthermore, we explore the use of Machine Learning algorithms in this context, highlighting their impact on developing effective strategies against COVID-19.

With an already pre-classified data set, we adopted a set of pre-processing steps, which, among others, included two very important steps: the selection of instances and attributes. This process was fundamental to select the most representative instances and attributes and ensure that the model obtained a good result capable of helping to understand the susceptibility of mammals to the virus.

Using the concept of interpretability of machine learning models, our research has revealed some of the most significant characteristics that make species susceptible to the SARS-CoV-2 virus. This step provided a more in-depth understanding of the evaluated context and also suggests gaps for future research.

Our model indicated that certain attributes, such as population density, longevity and social group size, are crucial determinants of a species' ability to act as a zoonotic host. Counterfactual analysis allowed for a deeper understanding of how these characteristics interact and influence susceptibility to SARS-CoV-2 infection.

The results indicate that species with high population density and complex social structure, such as many mammals, are more likely to facilitate viral transmission. Furthermore, species with a fast life cycle and high reproduction rate also showed greater potential to harbor and transmit pathogens.

Furthermore, in this work we used Discriminant Analysis to validate the classification carried out in the database that did not indicate susceptibility to transmission of SARS-CoV-2. This base of 5400 instances was classified using the model trained by the Neural Network and the Discriminant Analysis confirmed the classification of 89% of these instances, indicating that the characteristics found are indeed efficient for classification.

This research not only expands our understanding of the dynamics of zoonotic transmission but also provides a practical tool for identifying and monitoring potential hosts of SARS-CoV-2 and other emerging pathogens. Applying interpretability methods of learning models, such as counterfactual explanations, could be a valuable strategy for future epidemiology and infectious disease control research. These results highlight the potential of using machine learning algorithms in epidemiological research. They improve our understanding of the spread of SARS-CoV-2 and provide a solid foundation for developing more effective prevention and treatment strategies.

For future studies, it is recommended that the database be expanded with more biological and ecological attributes and that other machine learning techniques be explored that can complement and improve the results obtained. Interdisciplinary collaboration will be essential to refine these approaches and ensure an effective response to future pandemics.

References

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 chi conference on human factors in computing systems* (pp. 1–18). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3287560.3287574 doi: 10.1145/3173574.3174156

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160.

Ahmed, R., Hasan, R., Siddiki, A. M. A. M. Z., & Islam, M. S. (2021, January). Host range projection of SARS-CoV-2: South asia perspective. *Infect. Genet. Evol.*, *87*(104670), 104670.

Bao, L., Deng, W., Huang, B., Gao, H., Liu, J., Ren, L., … others (2020). The pathogenicity of sars-cov-2 in hace2 transgenic mice. *Nature*, *583*(7818), 830–833.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, *13*(2).

Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A survey. In *Ijcai-17 workshop on explainable ai (xai)* (Vol. 8, pp. 8–13). Retrieved from http://www.cs.columbia.edu/~orb/papers/xai_survey_paper_2017.pdf

Bosco-Lauth, A. M., Hartwig, A. E., Porter, S. M., Gordy, P. W., Nehring, M., Byas, A. D., … Bowen, R. A. (2020). Experimental infection of domestic dogs and cats with sars-cov-2: Pathogenesis, transmission, and response to reexposure in cats. *Proceedings of the National Academy of Sciences*, *117*(42), 26382–26388.

Brook, C. E., Rozins, C., Guth, S., & Boots, M. (2023). Reservoir host immunology and life history shape virulence evolution in zoonotic viruses. *Plos Biology*, *21*(9), e3002268.

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8), 832.

Cevik, M., Bamford, C., & Ho, A. (2020). Covid-19 pandemic—a focused review for clinicians. *Clinical Microbiology and Infection*, *26*(7), 842–847.

COVIDSurg Collaborative, G. C. (2021). Sars-cov-2 vaccination modelling for safe surgery to save lives: data from an international prospective cohort study. *British Journal of Surgery*, *108*(9), 1056–1063.

Damas, J., Hughes, G. M., Keough, K. C., Painter, C. A., Persky, N. S., Corbo, M., … others (2020). Broad host range of sars-cov-2 predicted by comparative and structural analysis of ace2 in vertebrates. *Proceedings of the National Academy of Sciences*, *117*(36), 22311–22322.

de Sousa Balbino, M., Gálvez, L. E. Z., & Nobre, C. N. (2023). Csse - an agnostic method of counterfactual, selected, and social explanations for classification models. *Expert Systems with Applications*, *228*, 120373. Retrieved from https://www.sciencedirect.com/science/article/pii/S0957417423008758 doi: https://doi.org/10.1016/j.eswa.2023.120373

Decaro, N., & Lorusso, A. (2020). Novel human coronavirus (sars-cov-2): A lesson from animal coronaviruses. *Veterinary microbiology*, *244*, 108693.

De Magalhaes, J., Costa, & J. (2009). A database of vertebrate longevity records and their relation to other life-history traits. *Journal of evolutionary biology*, *22*(8), 1770–1774.

Dhamnetiya, D., Goel, M. K., Jha, R. P., Shalini, S., & Bhattacharyya, K. (2022). How to perform discriminant analysis in medical research? explained with illustrations. *Journal of Laboratory Physicians*, *14*(04), 511–520.

Dorigo, M., Birattari, M., & Stutzle, T. (2006). Ant colony optimization. *IEEE computational intelligence magazine*, *1*(4), 28–39.

Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM*, *63*(1), 68–77.

El Shawi, R., Sherif, Y., Al-Mallah, M., & Sakr, S. (2019). Interpretability in healthcare a comparative study of local machine learning interpretability techniques. In *2019 ieee 32nd international symposium on computer-based medical systems (cbms)* (p. 275-280). IEEE. Retrieved from https://ieeexplore.ieee.org/document/8787506

ElShawi, R., Sherif, Y., Al-Mallah, M., & Sakr, S. (2021). Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. *Computational Intelligence*, *37*(4), 1633–1650.

Evans, J. P., & Liu, S.-L. (2021). Role of host factors in sars-cov-2 entry. *Journal of Biological Chemistry*, *297*(1).

Fardisi, M., Thelen, K., Groenendal, A., Rajput, M., Sebastian, K., Contreras, G. A., & Moeser, A. J. (2023). Early weaning and biological sex shape long-term immune and metabolic responses in pigs. *Scientific reports*, *13*(1), 15907.

Fischhoff, I. R., Castellanos, A. A., Rodrigues, J. P., Varsani, A., & Han, B. A. (2021). Predicting the zoonotic capacity of mammals to transmit sars-cov-2. *Proceedings of the Royal Society B*, *288*(1963), 20211651.

Gralinski, L. E., & Menachery, V. D. (2020). Return of the coronavirus: 2019-ncov. *Viruses*, *12*(2), 135.

Guidotti, R. (2022). Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 1–55.

Gupta, S., Minocha, R., PJ, T., Srivastava, M., & Dandekar, T. (2022). Role of the pangolin in origin of sars-cov-2: An evolutionary perspective. *International Journal of Molecular Sciences*, *23*(16), 9115.

Guth, S., Visher, E., Boots, M., & Brook, C. E. (2019). Host phylogenetic distance drives trends in virus virulence and transmissibility across the animal–human interface. *Philosophical Transactions of the Royal Society B*, *374*(1782), 20190296.

Han, B. A., Schmidt, J. P., Alexander, L. W., Bowden, S. E., Hayman, D. T., & Drake, J. M. (2016). Undiscovered bat hosts of filoviruses. *PLoS neglected tropical diseases*, *10*(7), e0004815.

Han, B. A., Schmidt, J. P., Bowden, S. E., & Drake, J. M. (2015). Rodent reservoirs of future zoonotic diseases. *Proceedings of the National Academy of Sciences*, *112*(22), 7039–7044.

Heesterbeek, H., Anderson, R. M., Andreasen, V., Bansal, S., De Angelis, D., Dye, C., … others (2015). Modeling infectious disease dynamics in the complex landscape of global health. *Science*, *347*(6227), aaa4339.

Huang, X., Zhang, C., Pearce, R., Omenn, G. S., & Zhang, Y. (2020). Identifying the zoonotic origin of sars-cov-2 by modeling the binding affinity between the spike receptor-binding domain and host ace2. *Journal of proteome research*, *19*(12), 4844–4856.

Huang, Z. Y., de Boer, W. F., van Langevelde, F., Olson, V., Blackburn, T. M., & Prins, H. H. (2013). Species' life-history traits explain interspecific variation in reservoir competence: a possible mechanism underlying the dilution effect. *PLoS One*, *8*(1), e54341.

Jackson, C. B., Farzan, M., Chen, B., & Choe, H. (2022). Mechanisms of sars-cov-2 entry into cells. *Nature reviews Molecular cell biology*, *23*(1), 3–20.

Jones, K. E., Bielby, J., Cardillo, M., Fritz, S. A., O'Dell, J., Orme, C. D. L., … others (2009). Pantheria: a species-level database of life history, ecology, and geography of extant and recently extinct mammals: Ecological archives e090-184. *Ecology*, *90*(9), 2648–2648.

Karesh, W. B., Dobson, A., Lloyd-Smith, J. O., Lubroth, J., Dixon, M. A., Bennett, M., … others (2012). Ecology of zoonoses: natural and unnatural histories. *The Lancet*, *380*(9857), 1936–1945.

Karim, A., Mishra, A., Newton, M. H., & Sattar, A. (2018). Machine learning interpretability: A science rather than a tool. Retrieved from https://arxiv.org/abs/1807.06722 doi: 10.48550/ARXIV.1807.06722

Khaledian, E., Ulusan, S., Erickson, J., Fawcett, S., Letko, M. C., & Broschat, S. L. (2022). Sequence determinants of human-cell entry identified in ace2-independent bat sarbecoviruses: A combined laboratory and computational network science approach. *EBioMedicine*, *79*, 103990.

Lam, S., Bordin, N., Waman, V., Scholes, H., Ashford, P., Sen, N., … others (2020). Sars-cov-2 spike protein predicted to form complexes with host receptor protein orthologues from a broad range of mammals. *Scientific reports*, *10*(1), 1–14.

Li, M.-Y., Li, L., Zhang, Y., & Wang, X.-S. (2020). Expression of the sars-cov-2 cell receptor gene ace2 in a wide variety of human tissues. *Infectious diseases of poverty*, *9*(02), 23–29.

Liu, Z., Xiao, X., Wei, X., Li, J., Yang, J., Tan, H., … Liu, L. (2020). Composition and divergence of coronavirus spike proteins and host ace2 receptors predict potential intermediate hosts of sars-cov-2. *Journal of medical virology*, *92*(6), 595–601.

Luan, J., Jin, X., Lu, Y., & Zhang, L. (2020). Sars-cov-2 spike protein favors ace2 from bovidae and cricetidae. *Journal of medical virology*, *92*(9), 1649–1656.

Lucatelli, J., Mariano-Neto, E., & Japyassú, H. F. (2021). Social interaction, and not group size, predicts parasite burden in mammals. *Evolutionary Ecology*, *35*, 115–130.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38.

Mishra, J., Mishra, P., & Arora, N. K. (2021). Linkages between environmental issues and zoonotic diseases: with reference to covid-19 pandemic. *Environmental Sustainability*, *4*(3), 455–467.

Mollentze, N., Keen, D., Munkhbayar, U., Biek, R., & Streicker, D. G. (2022). Variation in the ace2 receptor has limited utility for sars-cov-2 host prediction. *bioRxiv*.

Molnar, C. (2020). *Interpretable machine learning*. Lulu.com. Retrieved from https://christophm.github.io/interpretable-ml-book/

Munir, K., Ashraf, S., Munir, I., Khalid, H., Muneer, M. A., Mukhtar, N., … others (2020). Zoonotic and reverse zoonotic events of sars-cov-2 and their impact on global health. *Emerging microbes & infections*, *9*(1), 2222–2235.

Murata, T., Sakurai, A., Suzuki, M., Komoto, S., Ide, T., Ishihara, T., & Doi, Y. (2021). Shedding of viable virus in asymptomatic sars-cov-2 carriers. *MSphere*, *6*(3), e00019–21.

Myhrvold, N. P., Baldridge, E., Chan, B., Sivam, D., Freeman, D. L., & Ernest, S. M. (2015). An amniote life-history database to perform comparative analyses with birds, mammals, and reptiles: Ecological archives e096-269. *Ecology*, *96*(11), 3109–3109.

Nasserie, T., Hittle, M., & Goodman, S. N. (2021). Assessment of the frequency and variety of persistent symptoms among patients with covid-19: a systematic review. *JAMA network open*, *4*(5), e2111417–e2111417.

Nuismer, S. L., Basinski, A. J., Schreiner, C., Whitlock, A., & Remien, C. H. (2022). Reservoir population ecology, viral evolution and the risk of emerging infectious disease. *Proceedings of the Royal Society B*, *289*(1982), 20221080.

Organization, W. H. (2020). *Disease outbreak news.* Retrieved from https://www.who.int/emergencies/disease-outbreak-news/item/2020-DON301 (Access at: 2024/04/08)

Ostfeld, R. S., Levi, T., Jolles, A. E., Martin, L. B., Hosseini, P. R., & Keesing, F. (2014, 09). Life history and demographic drivers of reservoir competence for three tick-borne zoonotic pathogens. *PLOS ONE*, *9*(9), 1-8. Retrieved from https://doi.org/10.1371/journal.pone.0107387 doi: 10.1371/journal.pone.0107387

Pasquarelli-do Nascimento, G., Braz-de Melo, H. A., Faria, S. S., Santos, I. d. O., Kobinger, G. P., & Magalhães, K. G. (2020). Hypercoagulopathy and adipose tissue exacerbated inflammation may explain higher mortality in covid-19 patients with obesity. *Frontiers in endocrinology*, *11*, 530.

Pedersen, A. B., & Davies, T. J. (2009). Cross-species pathogen transmission and disease emergence in primates. *EcoHealth*, *6*, 496–508.

Plowright, R. K., Parrish, C. R., McCallum, H., Hudson, P. J., Ko, A. I., Graham, A. L., & Lloyd-Smith, J. O. (2017). Pathways to zoonotic spillover. *Nature Reviews Microbiology*, *15*(8), 502–510.

Praharaj, M. R., Garg, P., Kesarwani, V., Topno, N. A., Khan, R. I. N., Sharma, S., … others (2022). Sars-cov-2 spike glycoprotein and ace2 interaction reveals modulation of viral entry in wild and domestic animals. *Frontiers in Medicine*, *8*, 775572.

Riveiro-Valiño, J., Álvarez-López, C., & Marey-Pérez, M. F. (2009). The use of discriminant analysis to validate a methodology for classifying farms based on a combinatorial algorithm. *Computers and Electronics in Agriculture*, *66*(2), 113–120.

Rodrigues, J. P., Barrera-Vilarmau, S., Mc Teixeira, J., Sorokina, M., Seckel, E., Kastritis, P. L., & Levitt, M. (2020). Insights on cross-species transmission of sars-cov-2 from structural modeling. *PLoS computational biology*, *16*(12), e1008449.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

Serna García, G., Al Khalaf, R., Invernici, F., Ceri, S., & Bernasconi, A. (2023). Coveffect: interactive system for mining the effects of sars-cov-2 mutations and variants based on deep learning. *GigaScience*, *12*, giad036.

Shi, J., Wen, Z., Zhong, G., Yang, H., Wang, C., Huang, B., … others (2020). Susceptibility of ferrets, cats, dogs, and other domesticated animals to sars–coronavirus 2. *Science*, *368*(6494), 1016–1020.

Silva, P. V., & Nobre, C. N. (2024). Computational methods in the analysis of sars-cov-2 in mammals: A systematic review of the literature. *Computers in Biology and Medicine*, 108264.

Stepin, I., Alonso, J. M., Catala, A., & Pereira-Fariña, M. (2021). A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, *9*, 11974–12001.

Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, 1-21. doi: 10.1109/TNNLS.2020.3027314

Trougakos, I. P., Stamatelopoulos, K., Terpos, E., Tsitsilonis, O. E., Aivalioti, E., Paraskevis, D., … Dimopoulos, M. A. (2021). Insights to sars-cov-2 life cycle, pathophysiology, and rationalized treatments that target covid-19 clinical complications. *Journal of Biomedical Science*, *28*, 1–18.

Van Egeren, D., Novokhodko, A., Stoddard, M., Tran, U., Zetter, B., Rogers, M., … others (2021). Risk of rapid evolutionary escape from biomedical interventions targeting sars-cov-2 spike protein. *PloS one*, *16*(4), e0250780.

van Zundert, G. C., & Bonvin, A. M. (2014). Modeling protein–protein complexes using the haddock webserver "modeling protein complexes with haddock". *Protein structure prediction*, 163–179.

Volz, E., Mishra, S., Chand, M., Barrett, J. C., Johnson, R., Geidelberg, L., … others (2021). Transmission of sars-cov-2 lineage b. 1.1. 7 in england: Insights from linking epidemiological and genetic data. *MedRxiv*, 2020–12.

Wang, Y. X., Matson, K. D., Santini, L., Visconti, P., Hilbers, J. P., Huijbregts, M. A., … others (2021). Mammal assemblage composition predicts global patterns in emerging infectious disease risk. *Global change biology*, *27*(20), 4995–5007.

Wilman, H., Belmaker, J., Simpson, J., de la Rosa, C., Rivadeneira, M. M., & Jetz, W. (2014). Eltontraits 1.0: Species-level foraging attributes of the world's birds and mammals: Ecological archives e095-178. *Ecology*, *95*(7), 2027–2027.

Xanthopoulos, P., Pardalos, P. M., Trafalis, T. B., Xanthopoulos, P., Pardalos, P. M., & Trafalis, T. B. (2013). Linear discriminant analysis. *Robust data mining*, 27–33.

Yang, L. H., & Han, B. A. (2018). Data-driven predictions and novel hypotheses about zoonotic tick vectors from the genus ixodes. *BMC ecology*, *18*(1), 1–6.

Zebardast, E., Mazaherian, H., Rahmani, M., & Nouri, M. (2024). Developing a methodology for identifying urban neighborhoods with severe housing deprivation in iran. *Social Indicators Research*, *172*(1), 29–58.

Zhang, G., Li, B., Yoo, D., Qin, T., Zhang, X., Jia, Y., & Cui, S. (2021). Animal coronaviruses and sars-cov-2. *Transboundary and Emerging Diseases*, *68*(3), 1097–1110.

4 CONSIDERAÇÕES FINAIS

A revisão sistemática da literatura sobre o uso de métodos computacionais para avaliar a infecção por SARS-CoV-2 em mamíferos revelou conceitos pertinentes para compreender a disseminação viral. A aplicação de técnicas de Aprendizado de Máquina e outros métodos computacionais permitiram a identificação e classificação eficientes das espécies de mamíferos com potencial para serem portadoras ou hospedeiras intermediárias do vírus, contribuindo para a detecção precoce de ameaças potenciais à saúde pública.

Os estudos analisados destacaram a utilidade dos métodos computacionais na identificação de padrões específicos e características nos dados epidemiológicos e genéticos de mamíferos. Estas informações permitem identificar grupos de animais mais propensos a abrigar o vírus, ajudando na adoção de medidas preventivas específicas. Os dados do trabalho podem ajudar a identificar áreas onde mais pesquisas são necessárias, fatores que aumentam o risco e dificultam o controle de zoonoses, além de ações estratégicas em vigilância, pesquisa, comunicação e treinamento que podem apoiar a formação de uma rede de cooperação. Além disso, esses dados podem ser usados para educar o público sobre zoonoses, seus riscos e como preveni-las.

A pesquisa adotou uma abordagem multidisciplinar para entender os desafios da infecção viral por SARS-CoV-2, combinando a biologia do vírus, a interação com a proteína ACE2, a aplicação de técnicas de Aprendizado de Máquina, a interpretabilidade destes modelos de aprendizado e um pré-processamento dos dados bastante rigoroso, que envolveu, por exemplo, seleção de instâncias, eliminação de *outlier*, normalização, e redução de dimensionalidade. A análise da interpretabilidade dos modelos de aprendizado revelou características significativas que tornam as espécies susceptíveis ao vírus, melhorando a interpretabilidade dos modelos de aprendizado de máquina e fornecendo material para futuras pesquisas.

A integração de dados de diferentes fontes, como dados ambientais e climáticos, com dados de mamíferos pode melhorar ainda mais a precisão dos modelos de Aprendizado de Máquina na previsão da disseminação do SARS-CoV-2. A pesquisa demonstrou que certos atributos, como densidade populacional, longevidade e tamanho do grupo social, são determinantes na capacidade de uma espécie atuar como hospedeiro zoonótico.

Espécies com alta densidade populacional e estrutura social complexa, bem como aquelas com ciclo de vida rápido e alta taxa de reprodução, também mostraram maior potencial para abrigar e transmitir patógenos.

Os resultados não apenas ampliam o entendimento sobre a dinâmica da transmissão zoonótica, mas também oferecem uma ferramenta prática para identificar e monitorar potenciais hospedeiros de SARS-CoV-2 e outros patógenos emergentes. A análise da interpretabilidade dos modelos de aprendizado, que envolveu explicações contrafactuais para explicar o aprendizado da Rede Neural, pode ser uma estratégia valiosa para futuras pesquisas em epidemiologia e controle de doenças infecciosas.

Uma das limitações deste trabalho foi o tamanho da base de dados inicialmente rotulada. Para estudos futuros, recomenda-se a expansão da base de dados com mais atributos biológicos e ecológicos, além de explorar outras técnicas de aprendizado de máquina que possam complementar e aprimorar os resultados obtidos. A colaboração interdisciplinar será essencial para refinar estas abordagens e garantir uma resposta eficaz a futuras pandemias. Esses resultados destacam o potencial do uso de algoritmos de aprendizado de máquina na pesquisa epidemiológica, fornecendo uma base sólida para o desenvolvimento de estratégias de prevenção e tratamento mais eficazes.

REFERÊNCIAS

ARORA, P. et al. Learning from history: coronavirus outbreaks in the past. Dermatologic Therapy, Wiley Online Library, v. 33, n. 4, p. e13343, 2020.

BAO, L. et al. The pathogenicity of sars-cov-2 in hace2 transgenic mice. Nature, Nature Publishing Group, v. 583, n. 7818, p. 830–833, 2020.

BONI, M. F. et al. Evolutionary origins of the sars-cov-2 sarbecovirus lineage responsible for the covid-19 pandemic. Nature microbiology, Nature Publishing Group UK London, v. 5, n. 11, p. 1408–1417, 2020.

EGEREN, D. V. et al. Risk of rapid evolutionary escape from biomedical interventions targeting sars-cov-2 spike protein. PloS one, Public Library of Science, v. 16, n. 4, p. e0250780, 2021.

ELSHAWI, R. et al. Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. Computational Intelligence, Wiley Online Library, v. 37, n. 4, p. 1633–1650, 2021.

EVANS, J. P.; LIU, S.-L. Role of host factors in sars-cov-2 entry. Journal of Biological Chemistry, ASBMB, v. 297, n. 1, 2021.

FUENTE, J. de la; MERA, I. G. F. de; GORTÁZAR, C. Challenges at the host-arthropod-coronavirus interface and covid-19: a one health approach. Frontiers in Bioscience-Landmark, IMR Press, v. 26, n. 8, p. 379–386, 2021.

GRYSEELS, S. et al. Risk of human-to-wildlife transmission of sars-cov-2. Mammal Review, Wiley Online Library, v. 51, n. 2, p. 272–292, 2021.

HUANG, C. et al. Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. The lancet, Elsevier, v. 395, n. 10223, p. 497–506, 2020.

KORATH, A. D. et al. One health: Eaaci position paper on coronaviruses at the human-animal interface, with a specific focus on comparative and zoonotic aspects of sars-cov-2. Allergy, Wiley Online Library, v. 77, n. 1, p. 55–71, 2022.

KUIKEN, T. et al. Newly discovered coronavirus as the primary cause of severe acute respiratory syndrome. The Lancet, Elsevier, v. 362, n. 9380, p. 263–270, 2003.

LEE, N. et al. A major outbreak of severe acute respiratory syndrome in hong kong. New England Journal of Medicine, Mass Medical Soc, v. 348, n. 20, p. 1986–1994, 2003.

MEYER, B. et al. Antibodies against mers coronavirus in dromedary camels, united arab emirates, 2003 and 2013. Emerging infectious diseases, Centers for Disease Control and Prevention, v. 20, n. 4, p. 552, 2014.

MOLLENTZE, N. et al. Variation in the ace2 receptor has limited utility for sars-cov-2 host prediction. BIORXIV, Cold Spring Harbor Laboratory, 2022.

MUNIR, K. et al. Zoonotic and reverse zoonotic events of sars-cov-2 and their impact on global health. EMERGING MICROBES & INFECTIONS, Taylor & Francis, v. 9, n. 1, p. 2222–2235, 2020.

MURATA, T. et al. Shedding of viable virus in asymptomatic sars-cov-2 carriers. MSPHERE, Am Soc Microbiol, v. 6, n. 3, p. e00019–21, 2021.

SARDAR, R. et al. Comparative analyses of sar-cov2 genomes from different geographical locations and other coronavirus family genomes reveals unique features potentially consequential to host-virus interaction and pathogenesis. BIORXIV, Cold Spring Harbor Laboratory, p. 2020–03, 2020.

SHI, J. et al. Susceptibility of ferrets, cats, dogs, and other domesticated animals to sars–coronavirus 2. SCIENCE, American Association for the Advancement of Science, v. 368, n. 6494, p. 1016–1020, 2020.

TIWARI, R. et al. Covid-19: animals, veterinary and zoonotic links. VETERINARY QUARTERLY, Taylor & Francis, v. 40, n. 1, p. 169–182, 2020.

TROUGAKOS, I. P. et al. Insights to sars-cov-2 life cycle, pathophysiology, and rationalized treatments that target covid-19 clinical complications. JOURNAL OF BIOMEDICAL SCIENCE, Springer, v. 28, p. 1–18, 2021.

VOLZ, E. et al. Transmission of sars-cov-2 lineage b. 1.1. 7 in england: Insights from linking epidemiological and genetic data. MEDRXIV, Cold Spring Harbor Laboratory Press, p. 2020–12, 2021.

WARDEH, M. et al. Divide-and-conquer: machine-learning integrates mammalian and viral traits with network features to predict virus-mammal associations. NATURE COMMUNICATIONS, Nature Publishing Group UK London, v. 12, n. 1, p. 3954, 2021.

WARDEH, M.; SHARKEY, K. J.; BAYLIS, M. Integration of shared-pathogen networks and machine learning reveals the key aspects of zoonoses and predicts mammalian reservoirs. PROCEEDINGS OF THE ROYAL SOCIETY B, The Royal Society, v. 287, n. 1920, p. 20192882, 2020.

YANG, X. et al. Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method. COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL, Elsevier, v. 18, p. 153–161, 2020.

YOO, H. S.; YOO, D. Covid-19 and veterinarians for one health, zoonotic-and reverse-zoonotic transmissions. JOURNAL OF VETERINARY SCIENCE, The Korean Society of Veterinary Science, v. 21, n. 3, 2020.

ZAKI, A. M. et al. Isolation of a novel coronavirus from a man with pneumonia in saudi arabia. NEW ENGLAND JOURNAL OF MEDICINE, Mass Medical Soc, v. 367, n. 19, p. 1814–1820, 2012.

ZHOU, Z. et al. Genetic diversity and molecular epidemiology of middle east respiratory syndrome coronavirus in dromedaries in ethiopia, 2017 to 2020. EMERGING MICROBES & INFECTIONS, Taylor & Francis, n. just-accepted, p. 2164218, 2023.