UNIVERSIDADE FEDERAL DE MINAS GERAIS

Faculdade de Medicina

Programa de Pós-Graduação em Ciências da Saúde da Criança e do Adolescente

Adriano Lages dos Santos

CARACTERÍSTICAS CLÍNICAS E FATORES DE RISCO QUE AFETAM A MORTALIDADE DE CRIANÇAS E ADOLESCENTES COM COVID-19: UM ESTUDO DE COORTE RETROSPECTIVO DE ÂMBITO NACIONAL UTILIZANDO APRENDIZADO DE MÁQUINA

BELO HORIZONTE

2024

Adriano Lages dos Santos

CARACTERÍSTICAS CLÍNICAS E FATORES DE RISCO QUE AFETAM A MORTALIDADE DE CRIANÇAS E ADOLESCENTES COM COVID-19: UM ESTUDO DE COORTE RETROSPECTIVO DE ÂMBITO NACIONAL UTILIZANDO APRENDIZADO DE MÁQUINA

Tese apresentada ao Programa de Pós-Graduação em Ciências da Saúde da Criança e do Adolescente da Faculdade de Medicina da Universidade Federal de Minas Gerais como requisito parcial para obtenção do título de Doutor em Ciências da Saúde.

Orientador: Prof. Dr. Eduardo Araújo Oliveira

Coorientadora: Profª. Dra. Ana Cristina Simões e Silva

BELO HORIZONTE

2024

Santos, Adriano Lages dos.
SA237c Características clínicas e fatores de risco que afetam a mortalidade de crianças e adolescentes com COVID-19 [recursos eletrônicos]: um estudo de coorte retrospectivo de âmbito nacional utilizando aprendizado de máquina. / Adriano Lages dos Santos. - - Belo Horizonte: 2024.
123f.: il.
Formato: PDF.
Requisitos do Sistema: Adobe Digital Editions.

Orientador (a): Eduardo Araújo de Oliveira.
Coorientador (a): Ana Cristina Simões e Silva.
Área de concentração: Saúde da Criança e do Adolescente.
Tese (doutorado): Universidade Federal de Minas Gerais, Faculdade de Medicina.

1. COVID-19. 2. Inteligência Artificial. 3. Aprendizado de Máquina. 4. Criança. 5. Morte. 6. Dissertação Acadêmica. I. Oliveira, Eduardo Araújo de. II. Silva, Ana Cristina Simões e. III. Universidade Federal de Minas Gerais, Faculdade de Medicina. IV. Título.

NLM: WC 506

Bibliotecário responsável: Fabian Rodrigo dos Santos CRB-6/2697

UNIVERSIDADE FEDERAL DE MINAS GERAIS
FACULDADE DE MEDICINA - CENTRO DE PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS DA SAÚDE
SAÚDE DA CRIANÇA E DO ADOLESCENTE

ATA DE DEFESA DE TESE

Às oitos horas do dia dez de julho de dois mil e vinte e quatro na sala 618 (Auditório do CETES), no sexto (6°) andar da Faculdade de Medicina da Universidade Federal de Minas Gerais, realizou-se a sessão pública para a defesa de tese de Doutorado do aluno **ADRIANO LAGES DOS SANTOS**, número de registro 2022709777, graduado no curso de SISTEMAS DE INFORMAÇÃO, como requisito parcial para a obtenção do grau de Doutor em CIÊNCIAS DA SAÚDE - SAÚDE DA CRIANÇA E DO ADOLESCENTE, pelo Programa de Pós-Graduação em Ciências da Saúde-Saúde da Criança e do Adolescente. A Presidência da sessão coube ao Prof. Eduardo Araújo de Oliveira – Orientador (UFMG). Inicialmente o Presidente após dar conhecimento aos presentes sobre o teor das Normas Regulamentares do trabalho final de Pós-Graduação, fez a apresentação da Comissão Examinadora, assim, constituída pelos Professores Doutores: Eduardo Araújo de Oliveira – Orientador, Presidente (UFMG), Paulo Augusto Moreira Camargos (UFMG), Cristiane dos Santos Dias (FHEMIG/UFMG), Lilian Martins de Oliveira Diniz (FHEMIG/UFMG) e Ana Paula Couto da Silva (UFMG). Em seguida o Presidente autorizou o aluno a iniciar a apresentação de seu trabalho final intitulado **"CARACTERÍSTICAS CLÍNICAS E FATORES DE RISCO QUE AFETAM A MORTALIDADE DE CRIANÇAS E ADOLESCENTES COM COVID-19: UM ESTUDO DE COORTE RETROSPECTIVO DE ÂMBITO NACIONAL UTILIZANDO APRENDIZADO DE MÁQUINA"**. Seguiu-se à arguição pela comissão Examinadora, com a respectiva defesa do aluno. Logo após a Comissão reuniu-se sem a presença do candidato e do público para julgamento e expedição do resultado da avaliação do trabalho final do aluno e considerou a tese Aprovada. O resultado final foi comunicado publicamente ao aluno pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a sessão e lavrou a presente ata que, após lida, será assinada eletronicamente por todos os membros da Comissão Examinadora presente na sessão, através do SEI (Sistema Eletrônico de Informações) do Governo Federal.

Belo Horizonte, 10 de julho de 2024.

Documento assinado eletronicamente por **Paulo Augusto Moreira Camargos**, **Membro de comissão**, em 12/07/2024, às 11:40, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Ana Paula Couto da Silva**, **Professora do Magistério Superior**, em 12/07/2024, às 13:16, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Eduardo Araujo de Oliveira**, **Professor do Magistério Superior**, em 12/07/2024, às 15:23, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Lilian Martins Oliveira Diniz**, **Membro**, em 15/07/2024, às 07:38, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Cristiane dos Santos Dias**, **Professora do Magistério Superior**, em 15/07/2024, às 11:04, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador 3371733 e o código CRC 22C73137.

Para a minha filha Ana Clara e minha
esposa Andreia, com todo meu amor.

AGRADECIMENTOS

Aos professores Eduardo Araújo Oliveira e Ana Cristina Simões e Silva pela constante disponibilidade e por acreditarem no meu trabalho. Essa confiança é fundamental para minha motivação e prazer em realizar esta pesquisa. Gostaria também de agradecê-los por me permitirem fazer parte de uma equipe de pesquisa incrível. Sou uma simples engrenagem de um motor maior que é o grupo de pesquisa do qual eles fazem parte. Eles são essenciais para que tudo aconteça. Sou um simples ajudante e tenho orgulho de fazer parte desta equipe. O objetivo principal é sempre contribuir para a evolução da ciência.

Ao Programa de Pós-Graduação em Saúde da Criança e do Adolescente por acreditar na minha proposta de pesquisa e me dar a oportunidade de estar em um programa tão renomado.

A todos os professores com quem tive contato na Faculdade de Medicina da UFMG durante o curso de doutorado. Tive o prazer de participar de excelentes aulas, aprendi muito e me motivou cada vez mais em minha trajetória dentro da pós-graduação. Aos professores integrantes da banca de seleção do doutorado que também me motivaram e acreditaram no trabalho, dando sugestões importantes que não foram esquecidas e estão incluídas neste presente trabalho.

À Universidade Federal de Minas Gerais por me dar a oportunidade de estudar gratuitamente e com qualidade excepcional. Eu amo esta universidade e fazer parte desta comunidade.

Este estudo foi apoiado pelo CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) e FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais).

RESUMO

A pandemia de COVID-19 impulsionou a aplicação de tecnologias digitais avançadas, como a inteligência artificial (IA), para prever a mortalidade em pacientes adultos. No entanto, o desenvolvimento de modelos de aprendizado de máquina (ML) para prever desfechos em crianças e adolescentes com COVID-19 ainda é limitado. Este estudo teve como objetivo avaliar o desempenho de múltiplos modelos de aprendizado de máquina na previsão de mortalidade entre pacientes pediátricos hospitalizados com COVID-19 e analisar sua viabilidade quando aplicados a grandes bases de dados. Neste estudo de coorte, utilizamos o banco de dados SIVEP-Gripe, um recurso público mantido pelo Ministério da Saúde, para monitorar a síndrome respiratória aguda grave (SRAG) no Brasil. Para criar subconjuntos destinados ao treinamento e teste dos modelos de aprendizado de máquina (ML), dividimos o banco de dados primário em três partes. Com esses subconjuntos, desenvolvemos e treinamos 12 algoritmos de ML para prever os desfechos. Avaliamos o desempenho desses modelos utilizando diversas métricas, como acurácia, precisão, sensibilidade, revocação e a área sob a curva característica de operação do receptor (AUC). Entre as 37 variáveis examinadas, 24 foram identificadas como potenciais indicadoras de mortalidade, conforme determinado pelo teste de independência do qui-quadrado. O algoritmo de regressão logística (LR) obteve o maior desempenho, com uma acurácia de 92,5% e uma AUC de 80,1% no conjunto de dados otimizado. Os algoritmos de *Gradient Boosting Classifier* (GBC) e *Adaptive Boosting* (ADA) apresentaram resultados semelhantes aos do algoritmo LR. Nosso estudo também revelou que a saturação de oxigênio reduzida na linha de base, a presença de comorbidades e a idade avançada foram os fatores mais relevantes na previsão de mortalidade em crianças e adolescentes hospitalizados. O uso de modelos de ML pode ser uma ferramenta valiosa na tomada de decisões clínicas e na implementação de estratégias de gestão de pacientes baseadas em evidências, o que pode melhorar os desfechos dos pacientes e a qualidade geral dos cuidados médicos. Os modelos LR, GBC e ADA demonstraram eficiência na previsão precisa de mortalidade em pacientes pediátricos com COVID-19.

Palavras-chave: COVID-19; inteligência artificial; aprendizado de máquina; criança; morte.

ABSTRACT

The COVID-19 pandemic has catalyzed the application of advanced digital technologies such as artificial intelligence (AI) to predict mortality in adult patients. However, the development of machine learning (ML) models for predicting outcomes in children and adolescents with COVID-19 remains limited. This study aimed to evaluate the performance of multiple machine learning models in forecasting mortality among hospitalized pediatric COVID-19 patients and assess their feasibility when applied to large-scale datasets. In this cohort study, we used the SIVEP-Gripe dataset, a public resource maintained by the Ministry of Health, to track severe acute respiratory syndrome (SARS) in Brazil. To create subsets for training and testing the machine learning (ML) models, we divided the primary dataset into three parts. Using these subsets, we developed and trained 12 ML algorithms to predict the outcomes. We assessed the performance of these models using various metrics such as accuracy, precision, sensitivity, recall, and area under the receiver operating characteristic curve (AUC). Among the 37 variables examined, 24 were found to be potential indicators of mortality, as determined by the chi-square test of independence. The Logistic Regression (LR) algorithm achieved the highest performance, with an accuracy of 92.5% and an AUC of 80.1%, on the optimized dataset. The Gradient Boosting Classifier (GBC) and Adaptive Boosting (ADA) algorithms closely followed the LR algorithm, producing similar results. Our study also revealed that baseline reduced oxygen saturation, presence of comorbidities, and older age were the most relevant factors in predicting mortality in hospitalized children and adolescents. The use of ML models can be an asset in making clinical decisions and implementing evidence-based patient management strategies, which can enhance patient outcomes and overall quality of medical care. LR, GBC, and ADA models have demonstrated efficiency in accurately predicting mortality in COVID-19 pediatric patients

Keywords: COVID-19; artificial intelligence; machine learning; child; death.

LISTA DE ABREVIATURAS E SIGLAS

ADA	AdaBoost Classifier
AI	Artificial Intelligence
AUC	Area Under the Curve
AUC/ROC	Area Under the Receiver Operating Characteristic Curve
AUPRC	Area Under the Precision-Recall Curve
DT	Decision Tree
EHR	Eletronic Health Record
ET	Extra Trees
GBC	Gradient Boosting Classifier
ICU	Intensive Care Unit
IQR	Interquartile Range
KNN	K-Nearest Neighbors
LDA	Linear Discriminant Analysis
LR	Logistic Regression
ML	Machine Learning
NB	Naïve Bayes
NN	Neural Network
PROBAST	Prediction model Risk of Bias Assessment Tool
PROBAST-AI	Prediction model Risk of Bias Assessment Tool based on AI
QDA	Quadratic Discriminant Analysis
RF	Random Forest
RT-PCR	Reverse Transcription Polymerase Chain Reaction
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
SHAP	Shapley Additive exPlanations
SIVEP	Sistema de Informação de Vigilância Epidemiológica da Gripe
SMOTE	Synthetic Minority Oversampling Technique
SMOTE ENN Neighbor	Synthetic Minority Oversampling Technique Edited Nearest
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency

TRIPOD	Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis
TRIPOD-AI	Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis for AI
UTI	Unidade de Terapia Intensiva
XAI	Explainable Artificial Intelligence
XGBoost	Extreme Gradient Boosting

SUMÁRIO

1. INTRODUÇÃO

Os profissionais de saúde têm lidado com uma quantidade crescente de informações em seus ambientes de trabalho, assim como sobrecarga dentro dos hospitais (1). Essa sobrecarga foi agravada pela pandemia de COVID-19, que aumentou o número de pacientes hospitalizados em todo o mundo. Estudos mostram que dezenas de milhares de pessoas morrem a cada ano devido a erros médicos, frequentemente causados por exaustão ou estresse que os médicos experimentam em suas vidas diárias (2-6).

Durante e após a pandemia de COVID-19, o mundo tem experimentado uma evolução tecnológica no campo da inteligência artificial. Com o poder dos algoritmos de aprendizado de máquina, tarefas rotineiras e administrativas foram transferidas para esses sistemas, possibilitando que pessoas se tornem mais produtivas com menos esforço (7,8). Na área da saúde, a IA tem o potencial de auxiliar os profissionais em suas diversas áreas de atuação tanto do ponto de vista pessoal como profissional. Por exemplo, no campo pessoal a IA pode aliviar a exaustão dos profissionais, simplificar as tarefas cotidianas e reduzir a carga de trabalho, automatizando deveres administrativos. No campo de atuação profissional, a IA pode contribuir com o desenvolvimento de uma medicina personalizada e de intervenção precoce por meio de desenvolvimento de modelos de análise preditiva (9).

Utilizando sistemas eficientes de aprendizado de máquina, profissionais de saúde podem triar pacientes com suporte computacional e avaliar probabilidades de diversos desfechos, como óbito, internação em UTI, necessidade de suporte ventilatório, gravidade de uma doença específica, entre outros. Isso otimiza o fluxo de trabalho do médico e da equipe de saúde, fornecendo indicações de quais pacientes necessitam de atenção imediata e têm maiores riscos de complicações (9).

Estudos na literatura já foram conduzidos utilizando metodologias de Machine Learning (ML) na medicina para prever diversos desfechos em diferentes domínios médicos, por meio de tipos variados de dados, como texto e imagens (10). No entanto, ainda há uma lacuna a ser preenchida quanto ao uso desses algoritmos em pediatria. Poucos estudos exploram o uso de algoritmos de ML para previsão de desfechos em crianças e adolescentes, especialmente no contexto da COVID-19 (11).

Dessa maneira, o objetivo deste estudo é explorar técnicas de aprendizado de máquina para melhorar a compreensão e previsão dos desfechos da COVID-19 em crianças e adolescentes. Dentre as principais metas, incluem-se o desenvolvimento de modelos preditivos robustos para identificar os fatores de risco de mortalidade em pacientes pediátricos hospitalizados com COVID-19. Ao analisar um conjunto de dados abrangente que engloba vários parâmetros clínicos e demográficos, este estudo busca determinar os preditores mais significativos de mortalidade, ajudando assim os profissionais de saúde a tomar decisões informadas para o manejo do paciente.

Além disso, esta tese tem como objetivo comparar o desempenho de diferentes algoritmos de aprendizado de máquina, como AdaBoost, CatBoost, Random Forest, Regressão Logística, entre outros, na previsão da mortalidade por COVID-19 em casos pediátricos. Por meio de um processo rigoroso de avaliação, utilizando métricas como precisão, sensibilidade e área sob a curva ROC, o estudo pretende identificar o modelo mais eficaz para a previsão de mortalidade. Esta análise comparativa visa destacar os pontos fortes e limitações de diversas abordagens de aprendizado de máquina no contexto dos desfechos da COVID-19 em pacientes pediátricos.

Finalmente, este estudo pretende contribuir com percepções importantes para as políticas de saúde pública, mostrando a utilidade dos algoritmos de aprendizado de máquina na análise de bancos de dados de domínio público e fornecendo informações para os tomadores de decisão em saúde. Ao identificar parâmetros-chave na previsão do risco de mortalidade, este estudo tem como objetivo aprimorar a qualidade do atendimento e os desfechos clínicos dos pacientes pediátricos com COVID-19.

Seguindo as normas do Programa de Pós-Graduação em Saúde da Criança e do Adolescente, esta tese é estruturada da seguinte forma: Introdução, Revisão da literatura sob o formato de artigo científico previamente publicado, seguida pelos objetivos. A seção de Métodos está incluída dentro de cada artigo, tanto a revisão da literatura quanto o artigo original. Resultados, Discussão e Conclusões serão apresentados no formato de um artigo original intitulado "Análise Comparativa de Algoritmos de Aprendizado de Máquina para Previsão de Mortalidade por COVID-19 em Crianças e Adolescentes Usando um Grande Conjunto de Dados Públicos no Brasil." Finalmente, a conclusão da tese e apêndices são fornecidos. As Referências Bibliográficas são listadas no final de cada artigo ou seção. As citações no texto seguem o sistema Vancouver *(Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Writing and Editing for Biomedical Publication* - www.icmje.org).

REFERÊNCIAS

1. Youssef D, Youssef J, Abou-Abbas L, Kawtharani M, Hassan H. Prevalence and correlates of burnout among physicians in a developing country facing multi-layered crises: a cross-sectional study. Sci Rep. 2022 Dec 1;12(1).

2. Ulutasdemir N, Luna A, Tusconi M, Ryan E. The relationship between physician burnout and depression, anxiety, suicidality and substance abuse: A mixed methods systematic review [Internet]. Available from: https://www.crd.york.ac.uk/prospero/display_re

3. Jung F, Bodendieck E, Bleckwenn M, Hussenoeder F, Luppa M, Riedel-Heller S. Burnout, work engagement and work hours – how physicians' decision to work less is associated with work-related factors. BMC Health Serv Res. 2023 Dec 1;23(1).

4. Patel RS, Bachu R, Adikey A, Malik M, Shah M. Factors related to physician burnout and its consequences: A review. Vol. 8, Behavioral Sciences. MDPI Multidisciplinary Digital Publishing Institute; 2018.

5. Marques-Pinto A, Moreira S, Costa-Lopes R, Zózimo N, Vala J. Predictors of Burnout Among Physicians: Evidence From a National Study in Portugal. Front Psychol. 2021 Oct 1;12.

6. Yates SW. Physician Stress and Burnout. Vol. 133, American Journal of Medicine. Elsevier Inc.; 2020. p. 160–4.

7. Verdonk C, Verdonk F, Dreyfus G. How machine learning could be used in clinical practice during an epidemic. Vol. 24, Critical Care. BioMed Central Ltd.; 2020.

8. Ahuja AS. The impact of artificial intelligence in medicine on the future role of the physician. PeerJ. 2019;2019(10).

9. Abramoff MD, Whitestone N, Patnaik JL, Rich E, Ahmed M, Husain L, Hassan MY, Tanjil MSH, Weitzman D, Dai T, Wagner BD, Cherwek DH, Congdon N, Islam K. Autonomous artificial intelligence increases real-world specialist clinic productivity in a cluster-randomized trial. NPJ Digit Med. 2023 Oct 4;6(1):184. doi: 10.1038/s41746-023-00931-7. PMID: 37794054; PMCID: PMC10550906.

10. Deo RC. Machine learning in medicine. Circulation. 2015 Nov 17;132(20):1920–30.

11. Lages A, Santos D, Pinhati C, Perdigão J, Galante S, Silva L, et al. Machine learning algorithms to predict outcomes in children and adolescents with COVID-19: A systematic review ☆. Artif Intell Med [Internet]. 2024;150:102824. Available from: https://doi.org/10.1016/j.artmed.2024.102824

2. REVISÃO DA LITERATURA (publicada em revista científica)

2.1 Artigo de revisão sistemática (artigo já publicado em revista científica)

Machine learning algorithms to predict outcomes in children and adolescents with COVID-19: a systematic review

Adriano Lages dos Santos[1,2], Clara Pinhati[1], Jonathan Perdigão[1], Stella Galante[1], Ludmilla Silva[1], Isadora Veloso[1], Ana Cristina Simões e Silva[1], Eduardo Araújo Oliveira[1]

Affiliations

1 - Department of Pediatrics, Health Sciences Postgraduate Program, School of Medicine, Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil.

2 - Federal Institute of Minas Gerais (IFMG)

Abstract

Background and Objectives: We aimed to analyze the study designs, modeling approaches, and performance evaluation metrics in studies using machine learning techniques to develop clinical prediction models for children and adolescents with COVID-19.

Methods: We searched four databases for articles published between 01/01/2020 and 10/25/2023, describing the development of multivariable prediction models using any machine learning technique for predicting several outcomes in children and adolescents who had COVID-19.

Results: We included ten articles, six (60% [95% confidence interval (CI) 0.31 - 0.83]) were predictive diagnostic models and four (40% [95% CI 0.17 - 0.69]) were prognostic models. All models were developed to predict a binary outcome (n=10/10, 100% [95% CI 0.72 - 1]. The most frequently predicted outcome was disease detection (n=3/10, 30% [95% CI 0.11 - 0.60]). The most used machine learning models in the studies were tree-based (n=12/33, 36.3% [95% CI 0.17 - 0.47]) and neural networks (n=9/33, 27.2% [95% CI 0.15 - 0.44]).

Conclusion: Our review revealed that attention is required to address problems including small sample sizes, inconsistent reporting practices on data preparation, biases in data sources, lack of reporting metrics such as calibration and discrimination, hyperparameters and other aspects that allow reproducibility by other researchers and might improve the methodology.

Systematic Review Registration: PROSPERO, CRD42023414699. OSF, https://doi.org/10.17605/OSF.IO/EW2JD

Introduction

The healthcare landscape has witnessed a significant transformation in recent years with the advent of predictive models powered by advanced machine learning algorithms (1). These models have played a role in the evidence-based medicine revolution, providing clinicians with tools to improve decision-making, ameliorate patient outcomes, and optimize healthcare (2). A prediction model can be defined as a computational tool that utilizes historical data and statistical techniques to forecast future events. The analysis of large amount of patient data, including demographics, clinical variables, and diagnostic information, has the potential to aid in early detection of diseases, risk assessment, treatment planning, and personalized medicine (3-5). As predictive modeling continues to evolve, its impact on healthcare continues to grow, enabling clinicians to make more informed decisions and ultimately leading to better outcomes and patient care. Clinical prediction models typically fall into one of two main categories: prognostic prediction models, which predict the likelihood of developing a particular health outcome over a specific period, and diagnostic prediction models, which determine an individual's likelihood of having a particular health condition (typically a disease) (6).

Machine learning techniques have been helping in the analysis of large-scale COVID-19 data, including in studies with children and adolescents. Several studies provide insights into the clinical outcomes, vaccine efficacy, and risk factors associated with COVID-19 in this specific population (7-9). These algorithms can assist clinicians and researchers in analyzing these datasets by identifying patterns, predicting outcomes, and finding relevant risk factors for severe illness or adverse events. Taking advantage of computational methods, machine learning can help uncover hidden relationships, identify early warning signs, and help clinical decision-making. In the context of pediatric patients, machine learning can provide a tool for extracting actionable insights from the complex and diverse data related to COVID-19 in children and adolescents, ultimately contributing to the development of targeted interventions.

Development and validation of prediction models for clinical settings rely on the use of appropriate study designs and modeling strategies. However, there is a lack of comprehensive information regarding the specific study designs, modeling approaches, and performance measures employed in studies that utilize machine learning for prediction modeling (10). Therefore, our objective was to conduct a systematic review to analyze and summarize the key characteristics related to study design, modeling techniques, and performance measures reported in studies focusing on clinical prediction models developed using supervised machine learning algorithms in pediatric patients with COVID-19.

Methods

We followed the PRISMA 2020 guidelines for systematic reviews (11). This systematic review was registered and approved in PROSPERO under the protocol CRD42023414699 and in OSF available at https://doi.org/10.17605/OSF.IO/EW2JD.

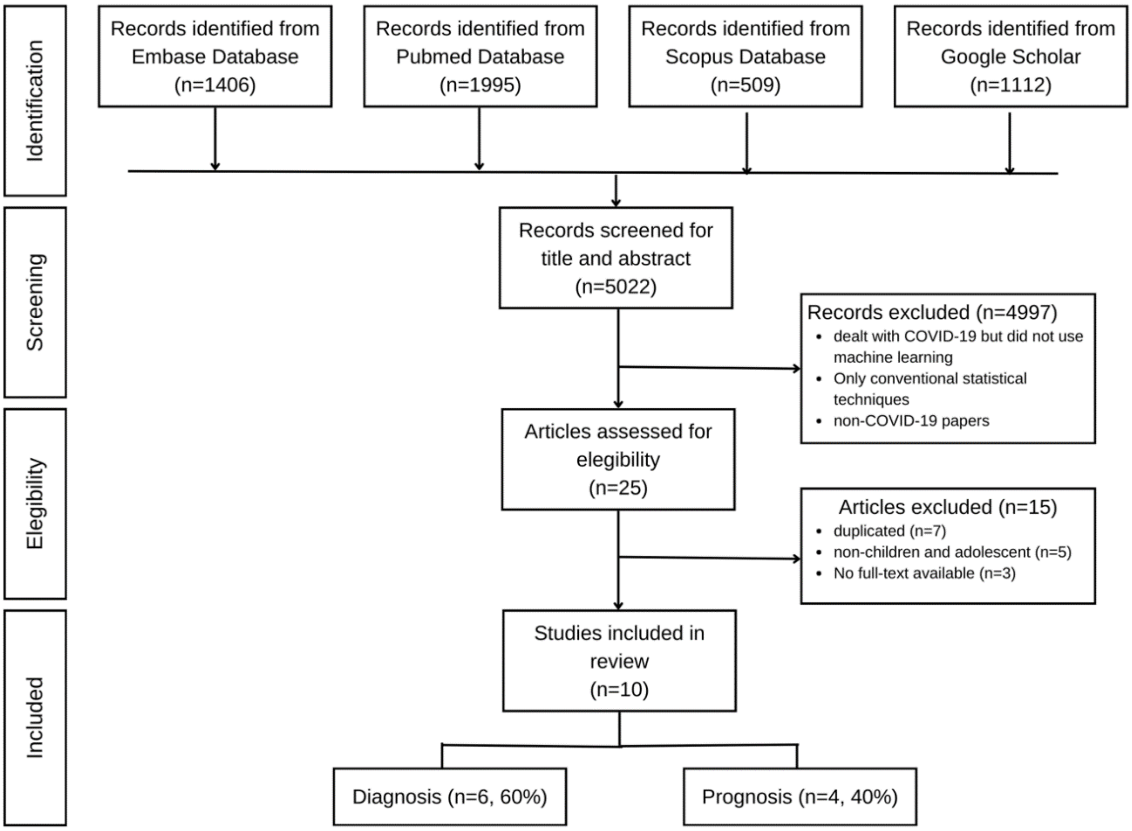

Identification
Screening
Elegibility
Included
Records identified from Embase Database (n=1406)
Records identified from Pubmed Database (n=1995)
Records identified from Scopus Database (n=509)
Records identified from Google Scholar (n=1112)
Records screened for title and abstract (n=5022)
Records excluded (n=4997)
dealt with COVID-19 but did not use machine learning
Only conventional statistical techniques
non-COVID-19 papers
Articles assessed for elegibility (n=25)
Articles excluded (n=15)
duplicated (n=7)
non-children and adolescent (n=5)
No full-text available (n=3)
Studies included in review (n=10)
Diagnosis (n=6, 60%)
Prognosis (n=4, 40%)

Figure 1. Flowchart of included studies

The systematic mapping was conducted following three adopted stages described below. (12)

Step 1 - Conduct searches: Based on the research questions, a replicable method for searching and retrieving articles in four selected scientific databases was defined and executed. The databases were Embase, Google Scholar, Pubmed, and Scopus Elsevier.

Step 2 - Selection of studies: A systematic method was defined and applied to select only the relevant articles for this study using inclusion and exclusion eligibility criteria. We used the open-source software Zotero (version 6.0.26) to exclude duplicate articles from the search results.

Step 3 - Data extraction and analysis: Finally, the relevant data from the primary studies were summarized and presented in this study. For each study, we collected the following information: study design characteristics (such as cohort, case-control, randomized trial), data source (such as routinely collected data, registries, administrative databases), study population details, outcome measures, setting information, patient characteristics, sample size (before and after participant exclusion), number of events, number of candidate and final predictors, handling of missing data, hyperparameter optimization, dataset splitting (such as train-validation-test), method for internal validation (such as bootstrapping, cross-validation), number of models developed and/or validated, and availability of code, data, and model. Country was defined based on the location of the first author's affiliation. For each model, we extracted information on the algorithm used, predictor selection methods, variable importance reporting, use of penalization techniques, hyperparameters reporting, and performance metrics (such as discrimination and calibration).

Step 1 - Search strategy for scientific articles

To identify possible primary studies relevant to data extraction, the search was based on (i) studies using keyword combinations derived from our objective and (ii) the execution of automatic searches on scientific databases using search terms. Initially, relevant keywords related to four main fields were selected: (a) COVID-19; (b) medicine; (c) early childhood, childhood, and adolescence; (d) Artificial Intelligence and Machine Learning.

The resulting keywords for each main field were:

COVID-19: COVID-19 OR SARS-COV-2

Medicine: outcomes OR outcome OR mortality OR death OR hospitalization OR hospitalized OR ICU OR ventilation

Population: Early childhood, childhood, adolescence: child OR "early childhood" OR children OR newborn OR adolescent OR adolescents

AI and Machine Learning: "machine learning" OR "artificial intelligence" OR algorithm OR algorithms OR dataset OR dimensions OR training OR sample OR samples OR prediction OR predict OR predicting OR forecast OR forecasting OR classification OR regression OR dimension OR models OR model OR predictive OR predictors OR bootstrapping OR bootstrap

Search terms were defined by grouping keywords in the same domain with the logical operator "OR" and grouping the three main concepts with the logical operator "AND". Then, automatic searches were executed on four scientific databases, including Embase, Google Scholar, Pubmed, and Scopus Elsevier. The search limited articles by year of publication (2019 to 2023).

Step 2 - Eligibility criteria (Selection of studies)

The studies retrieved from automatic searches were filtered to exclude articles not aligned with the study objectives. At this stage, three independent researchers defined and applied the following inclusion and exclusion criteria.

Inclusion criteria:

Studies whose main focus is on the use of machine learning algorithms to predict deaths and other outcomes in children or adolescents who had COVID-19.

The search period comprises 01/01/2020 to 10/25/2023. The year limit of 2019 was used because some databases did not allow filtering with monthly granularity. Thus, it was not possible to specify the month of March 2020 (the beginning of the pandemic).

To be included in the first selection, articles must address the topics of COVID-19 in children or adolescents and use machine learning algorithms to predict various outcomes in these patients. Although the outcome of death is highlighted in the search keywords in the Medicine domain, this search also considered other outcomes to increase the range of possible articles returned in the search. Only articles written in English were considered for the search. Only articles published in journals or conferences were considered for this search. Regarding articles published in conferences, we consider those papers presented at conferences and published in the conference proceedings.

Exclusion criteria:

Articles written in languages other than English. Articles that do not deal with COVID-19 in children and adolescents, articles that do not use machine learning algorithms in the prediction of various COVID-19 outcomes, duplicated articles, and articles that were selected in the databases but whose completed text files were not obtained even after demanding the corresponding authors.

The study selection process was carried out in two phases: (i) in the first selection phase, the titles and abstracts of the studies retrieved from the searches were read, and studies that did not meet the inclusion criteria were excluded; (ii) in the second selection phase, all articles were downloaded, and their introduction and conclusion were read to remove studies that met the exclusion criteria.

For this review, we did not use the "snowballing" technique, which involves checking if there are any articles in the references of the selected articles, after a complete reading, that were not found in the initial database search. If such articles are identified, they are then selected for inclusion in the review. Figure 1 presents the number of articles selected after each phase and the application of inclusion and exclusion criteria. And the table in the Supplemental File 1 also summarize the results after each phase.

Screening and selection process

The titles and abstracts were thoroughly examined by three researchers independently from a team of eight researchers to identify studies that potentially met the eligibility criteria. The group of researchers comprised two senior medical professors, a doctoral candidate, and five undergraduate medical students. The undergraduate medical students and the doctoral student were involved in research projects related to the effects of COVID-19 in children and adolescents. Subsequently, full-text articles were obtained, and three groups of two researchers independently evaluated all articles, while the same articles were collectively reviewed by four researchers to ensure agreement. In the event of any discrepancies during the screening and selection process, the primary reviewer of this study was consulted to assess the concerned article and resolve discrepancies carefully.

Step 3 - Data extraction

We selected several items from existing methodological guidelines for reporting and critical appraisal of prediction model studies to build our data extraction form (TRIPOD and PROBAST) (13-14). The following items were extracted in the selected studies based on the systematic review conducted by Navarro et al. (10), including the items described in step 3 of our methodology. One reviewer recorded all items, while the other reviewers collectively assessed all articles. Articles were randomly assigned to reviewers. Discrepancies in data extraction were discussed and solved between the pair of reviewers. No limitations were imposed on the number of models extracted per article.

Summary statistics and integration of findings

The findings were condensed into percentages (with confidence intervals calculated using the Wilson score interval and the Wilson score continuity-corrected interval, as appropriate), medians, and interquartile range (IQR), accompanied by a descriptive synthesis.

We reported only overall performance data from the studies, specifically the overall mean performance reported in the studies. We did not differentiate performance into corrected, external validation, or apparent validation segments. We did not report external validation data, even for studies that validated their models using different data from the model development and testing phase.

Rather than assessing the intricacies of each modeling approach and its performance, our evaluations remained at the study level. We refrained from conducting a quantitative synthesis of the models' performance, such as a meta-analysis, as it fell outside the scope of our review due to the reason that the available studies on the topic may have significant heterogeneity in terms of study design, patient populations, interventions, or outcomes, making it inappropriate or unreliable to combine their results quantitatively. All analyses were conducted using the software R version 4.1.0 (R Core Team, Vienna, Austria).

Results

The search in the selected databases for this review yielded 5022 articles. After assessing the titles and abstracts, 25 studies potentially met the eligibility criteria. Following a thorough reading of all 25 studies, ten articles were included in this review: 6 (60% [95% confidence interval (CI) 0.31 - 0.83]) were predictive diagnostic models and 4 (40% [95% CI 0.17 - 0.69]) were prognostic models (Figure 1).

We evaluated the quality of the articles regarding their adherence to the TRIPOD guidelines and also assessed the risk of bias in the selected studies using the PROBAST tool. Regarding the adherence to the TRIPOD guidelines, the selected studies showed an average adherence of 67.09%. TRIPOD is a checklist consisting of 31 items, and the selected studies, on average, fulfilled 20 items from this checklist. The results of the adherence assessment of each article to the TRIPOD guidelines can be found in Supplemental File 2.

Regarding the risk of bias assessment using the PROBAST tool, five studies showed a high risk of bias concerning their prediction models, four studies showed a low risk of bias, and one study had an unclear result regarding bias risk. The results of the assessment for each study in the dimensions evaluated by PROBAST (Participants, Predictors, Outcome, and Analysis) can be found in Supplemental File 3.

Among the 10 articles, 7 studies (70% [95% CI 0.40 - 0.89]) developed prediction models and assessed their performance using internal validation techniques, while 3 studies (30% [95% CI 0.11 - 0.60]) developed and externally validated the same machine learning predictive model. Six studies were published in 2022 (60% [95% CI 0.31 - 0.83]), three in 2021 (30% [95% CI 0.11 - 0.60]) and one study in 2023 (10% [95% CI 0.018 - 0.40]). The clinical fields involved in the selected articles were pediatrics (n=7/10, 70% [95% CI 0.40 - 0.89]), public health (n=2/10, 20% [95% CI 0.057 - 0.51]), and pulmonology (n=1/10, 10% [95% CI 0.018 - 0.40]). The retrieved articles originated from Europe (n=4/10, 40% [95% CI 0.17 - 0.69]), Asia (n=3/10, 30% [95% CI 0.11 - 0.60]), and North America (n=3/10, 30% [95% CI 0.11 - 0.60]). Other study characteristics are presented in Table 1.

Table 1 – General characteristics of the included studies

	Total (n = 10)
Key characteristics	**n (%) [95% CI]**
Study aim	
Diagnosis	6 (60) [0.31 - 0.83]
Prognosis	4 (40) [0.17 - 0.69]
Study Type	
Model development only	7 (70) [0.40 - 0.89]
Model development with external validation	3 (30) [0.11 - 0.60]
Outcome aim	
Classification	6 (40) [0.31 - 0.83]
Risk Probabilities	4 (40) [0.17 - 0.69]
Setting [a]	
General population	6 (60) [0.31 - 0.83]
Secondary care	1 (10) [0.018 - 0.40]

Tertiary care	3 (30) [0.11 - 0.60]
Outcome format	
Binary	10 (100) [0.72 - 1]
Type of outcome	
Death	1 (10) [0.018 - 0.40]
Severity prediction	1 (10) [0.018 - 0.40]
Hospitalization prediction	2 (20) [0.063 - 0.55]
Complications	2 (20) [0.063 - 0.55]
Need of ICU	1 (10) [0.018 - 0.4]
Disease detection	3 (30) [0.11 - 0.60]
Mentioning reporting guidelines (Tripod, Strobe, Charms, other)	
TRIPOD	1 (10) [0.018 - 0.4]
None	9 (90) [0.60 - 0.98]
Model availability [a]	
Repository for data	5 (50) [0.24 - 0.76]
Repository for code	2 (20) [0.057 - 0.51]
Model presentation	8 (80) [0.49 - 0.94]
None	2 (20) [0.057 - 0.51]

[a] Counts are absolute numbers with column percentages in parentheses. The percentages sometimes do not add up to 100% because studies reported more than one option. ICU = intensive care unit

In total, 33 prediction models were developed (Mean: 3 models per study, IQR: 4, Range: 1-5). We did not set a limit for extracting models per study, since were few articles included in this review. Thus, all 33 models found in the selected studies were evaluated. The most used machine learning models in the studies were tree-based (n=12/33, 36.36% [95% CI 0.17 - 0.47]) and neural networks (n=9/33, 27.27% [95% CI 0.15 - 0.44]). Other algorithms encountered are described in Table 2.

Table 2 - Algorithms used for modeling in all extracted models from the selected studies

	All extracted models (n = 33)
Modeling algorithm [a]	**n (%) [95% CI]**
Tree Based Models	12 (36.3) [0.17- 0.47]
Decision trees (for example, CART)	3 (25) [0.089- 0.53]
Random forest	2 (16.6) [0.047- 0.45]
Gradient boosting machine (Catboost)	3 (25) [0.089- 0.53]
XGBoost	4 (33.4) [0.14- 0.61]
Neural Network (incl. deep learning)	9 (27.2) [0.15 - 0.44]
Support Vector Machine	2 (6.06) [0.017 - 0.20]
Naive Bayes	1 (3.03) [0.0054 - 0.15]
Multiple logistic regression	1 (3.03) [0.0054 - 0.15]
Logistic regression	4 (12.1) [0.048 - 0.27]
Linear discriminant analysis	2 (6.06) [0.017 - 0.20]
Other (TabNet, AutoM, DeepFM, etc)	3 (9.09) [0.031 - 0.24]

[a] Counts are absolute numbers with column percentages in parentheses. The percentages sometimes do not add up to 100% because studies developed more than one model.

Participants

The participants included in the reviewed studies were recruited from the general population (n=6/10, 60% [95% CI 0.31 - 0.83]), tertiary care settings (n=3/10, 30% [95% CI 0.11 - 0.60]), and secondary care settings (n=1/10, 10% [95% CI 0.018 - 0.40]) (Table 1).

Data sources

The prediction models were predominantly developed using administrative databases (n=7/10, 70% [95% CI 0.40 - 0.89]). Prospective cohort data (n=1/10, 10% [0.018 - 0.40]) and retrospective cohort data (n=1/10, 10% [0.018 - 0.40]) were reported in one study each. The reviewed studies utilized electronic medical records and surveys. However, there was no information available in the selected articles regarding the time spent on data collection for the studies. Similarly, no studies reported the time horizon for the predictions (n=10/10, 100% [95% CI 0.72 - 1]).

Outcomes

All models were developed to predict a binary outcome (n=10/10, 100% [95% CI 0.72 - 1]). The most frequently predicted outcome was disease detection (n=3/10, 30% [95% CI 0.11 - 0.60]) followed by hospitalization prediction and complications both with two studies each (n=2/10, 20%, [95% CI 0.057 - 0.51]). Other outcomes of severity prediction are described in Table 1.

Table 3 - Study design of included studies

	Total (n = 10)
Key items [a]	**n (%) [95% CI]**
Data sources	
Prospective cohort	1 (10) [0.018 - 0.40]
Retrospective cohort	1 (10) [0.018 - 0.40]
Electronic health record	1 (10) [0.018 - 0.40]
Administrative databases	7 (70) [0.4 - 0.89]
Survey	1 (10) [0.018 - 0.40]
Predictor horizon	
None	10 (100) [0.72 - 1]
Sample size justification	
Size of existing/available data	7 (70) [0.40 - 0.89]
None	3 (30) [0.11 - 0.66]
Internal validation [a]	
Split sample with test set	9 (90) [0.60 - 0.98]
(Random) split	5 (50) [0.24 - 0.76]
(Nonrandom) split	2 (20) [0.018 - 0.59]
Split	1 (10) [0.022 - 0.40]
Bootstrapping	1 (10) [0.022 - 0.40]
With test set	1 (100) [0.21 - 1]
Cross-validation	5 (50) [0.24 - 0.76]
Nested	5 (100) [0.57 - 1]
External validation	3 (30) [0.11 - 0.60]
Independent dataset	3 (100) [0.44 - 1]

[a] Counts are absolute numbers with column percentages in parentheses. The percentages sometimes do not add up to 100% because studies reported more than one option.

Candidate Predictors

Candidate predictors extracted from the studies were clinical history (n=5/10, 50% [95% CI 0.24 - 0.76]), demographics including sex, gender, and ethnicity/race (n=5/10, 50% [95% CI 0.24 - 0.76]) and disease (the diagnosed disease) (n=5/10, 50% [95% CI 0.24 - 0.76]). Other predictors extracted (physical examination, blood or urine parameters, imaging, pathology, and questionnaires) are described in Table 4. None of the selected studies used treatment modalities as predictors for the developed models and for one study, treatment as a candidate predictor is not applicable, since the developed models are dealing with imaging data. Studies included a median of 15 candidate predictors (IQR: 6 - 14.5). Four studies included continuous variables as candidate predictors (40% [95% CI 0.17 - 0.69]), the other three studies did not use continuous variables as predictors (30% [95% CI 0.11 - 0.60]). Most studies did not report the methods to handle continuous predictors (60% [95% CI 0.31 - 0.83]).

Table 4 - Predictors in included studies

	Total (n = 10)
Key items	**n (%) [95% CI]**
Type of candidate predictors [a]	
Demography	5 (50) [0.24 - 0.76]
Clinical history	5 (50) [0.24 - 0.76]
Physical examination	3 (30) [0.11 - 0.6]
Disease	5 (50) [0.24 - 0.76]
Blood or urine parameters	3 (30) [0.11 - 0.6]
Imaging	1 (10) [0.018 - 0.40]
Pathology	3 (30) [0.11 - 0.60]
Questionnaires	1 (10) [0.018 - 0.40]

Scale Score	1 (10) [0.018 - 0.40]
Treatment as candidate predictor	
Yes	
No	9 (90) [0.60 - 0.98]
Not applicable	1 (10) [0.018 - 0.40]
Continuous variables as candidate predictors [b]	
Yes	4 (40) [0.17 - 0.69]
No	3 (30) [0.11 - 0.60]
Unclear	3 (30) [0.11 - 0.60]
A-priori selection of candidate predictors	
Yes	5 (50) [0.24 - 0.76]
No	5 (50) [0.24 - 0.76]
Methods to handle continuous predictors [a, b]	
Nonlinear (planned)	1 (10) [0.018 - 0.40]
Unclear	6 (60) [0.31 - 0.83]
Not applicable	3 (30) [0.11 - 0.60]
Categorization of continuous predictors [b]	
Not reported	10 (100) [0.72 - 1]

[a] Counts are absolute numbers with column percentages in parentheses. The percentages sometimes do not add up to 100% because studies reported more than one option.

[b] as data preparation

Sample size

Selected studies had a median sample size of 11,108 participants (IQR: 5,664 - 65,518). Most studies report a sample size justification or calculation rationale as the size of existing/available data used (n=7/10, 70% [95% CI 0.40 - 0.89]), and 3 studies did not report any rationale about sample size (n=3/10, 30% [95% CI 0.11 - 0.66]) (Table 3) (Table 5).

Table 5 - Sample size of included studies

	Total (n = 10)	
Key items	**n (%) [95% CI]**	**Median [IQR], range**
Initial sample size	10 (100)	11,108 [5664 - 65518] 105 to 23 million
Final sample size	10 (100)	7,801 [5664 - 65518] 99 to 23 million
Model development	10 (100)	6,955 [3000 - 65518] 99 to 20 million
Internal validation	9 (88.9)	7,139 [799 - 58188] 99 to 16 million
External validation [a]	3 (22.3)	Not significant
Number of candidate predictors	10 (100)	23 [14 - 33] 3 to 200
Number of included predictors	10 (100)	16 [7 - 21] 3 to 65

[a] Only three studies conducted external validation. For the IQR calculation to have significance, a minimum of four values is required.

Missing values

Missing values were an exclusion criterion of participants in three studies (30% [95% CI 0.11 - 0.60]). On the other hand, seven studies were unclear regarding missing data being a criterion for exclusion of participants, as we did not find this information (70%, [95% CI 0.40 - 0.89]). When a study did not explicitly mention that there are no missing data, we consider that the study was not clear about the existence of missing data. To handle missing data, most of the studies are unclear (n=4/10, 40% [95% CI 0.17 - 0.69]). One study used Bayesian optimization (n=1/10, 10% [95% CI 0.018 - 0.40]), and two studies did not make imputation of the missing data in the data source (n=2/10, 20% [95% CI 0.057 - 0.51]). Other information about how studies reported ways to handle missing data is presented in Table 6.

Table 6 – Methods used for missing values handling

	Total (n = 10)
Key items	**n (%) [95% CI]**
Missingness as exclusion criteria for participants	
Yes	3 (30) [0.11 - 0.6]
Unclear	7 (70) [0.4 - 0.89]
Number of patients excluded	
Median [IQR] (range)	1007 [303 - 6,247,840] (6 to 12,494,266)
Methods of handling missing data	
No missing data	3 (30) [0.11 - 0.6]
No imputation	2 (20) [0.057 - 0.51]
Bayesian optimization	1 (10) [0.018 - 0.4]
Unclear	4 (40) [0.17 - 0.69]
Presentation of missing data	
Not summarized	6 (60) [0.31 - 0.83]
By all candidate predictors	1 (10) [0.018 - 0.4]
Not applicable	3 (30) [0.11 - 0.6]

Class imbalance and dimensionality reduction techniques

Eight among 10 studies (80%, [95% CI 0.49 - 0.94] did not report unbalanced data or any strategy to deal with class imbalance like Synthetic Minority Oversampling Technique (SMOTE), Random Undersampling Boosting (RUSBoost), Random oversampling, random under sampling, or other techniques. For one study class imbalance is not applicable, since the study deals with imaging as a data source and one study report the use of SMOTE to deal with class imbalance. Regarding dimensionality reduction, most studies did not report any technique to reduce the dimension of data (n=8/10, 80% [95% CI 0.49 - 0.94]). One study used principal component analysis (PCA) to reduce the dimension of data (Table 7).

Table 7 - Machine learning aspects in the included studies

	Total (n = 10)
key items	**n (%) [95% CI]**
Data preparation [a]	
Cleaning	2 (20) [0.057 - 0.51]
Aggregation	1 (10) [0.018 - 0.40]
Augmentation	1 (10) [0.018 - 0.40]
Encoding	2 (20) [0.057 - 0.51]
Normalization	1 (10) [0.018 - 0.40]
Other	2 (20) [0.057 - 0.51]
Not reported	6 (60) [0.31 - 0.83]
Data splitting	
Train-test set	6 (60) [0.31 - 0.83]
Train-validation-test set	4 (40) [0.17 - 0.69]
Dimensionality reduction techniques	
Principal component analysis	1 (10) [0.018 - 0.40]
Not Reported	8 (80) [0.49 - 0.940]
Not applicable	1 (10) [0.018 - 0.40]
Class Imbalance	
SMOTE	1 (10) [0.018 - 0.40]
Not Reported	8 (80) [0.49 - 0.94]
Not applicable	1 (10) [0.018 - 0.40]
Strategy for hyperparameter optimization [a]	
Cross-validation	4 (40) [0.17 - 0.69]
Manual search	1 (10) [0.018 - 0.40]
Predefined values/default	1 (10) [0.018 - 0.40]
Done automatically by CatBoost	1 (10) [0.018 - 0.40]
Not Reported	7 (70) [0.40 - 0.89]

[a] Counts are absolute numbers with column percentages in parentheses. The percentages sometimes do not add up to 100% because studies reported more than one option. SMOTE = Synthetic Minority Oversampling TEchnique

Modeling Algorithms

Neural networks were used in 9 among 33 models (27.2% [95% CI 0.15 - 0.44]) extracted from the selected studies, including multilayer perceptron, convolutional neural network, and recurrent neural networks. Tree-based models were reported in 12 among 33 models (36.3% [95% CI 0.17 - 0.47]). Other models such as TabNet, AutoML, and DeepFM were also adopted in the selected studies (n=3/33, 9.09% [95% CI 0.031 - 0.24]). We did not find any study that reported penalized regression models. Support Vector Machine (SVM), a popular machine learning technique, was also reported two times (n=2/33 6.06% [95% CI 0.017 - 0.20]).

Selection of predictors

Regarding the strategy to build models, different methods of selection of predictors were reported as presented in Table 8. Some of the strategies found in the selected studies include term frequency-inverse document frequency (TF-IDF) embedding, frequency encoding, and embedding in the learning process (data-driven approach), decided by pediatricians and others. The most cited method for model building was Spearman Correlation (n=4/33, 12.12% [95% CI 0.048 - 0.27]).

Table 8 - Model building of all included studies

	Total (n = 33)
Key items	**n (%) [95% CI]**
Selection of predictors [a]	
Impurity Based Feature Importance	1 (3.03) [0.054 - 0.15]
TF-IDF Embedding	1 (3.03) [0.054 - 0.15]
Frequency Encoding/Count Encoding	1 (3.03) [0.054 - 0.15]
Spearman Correlation	4 (12.12) [0.048 - 0.27]
All predictors	2 (6.06) [0.017 - 0.20]
Decided by pediatricians	1 (3.03) [0.054 - 0.15]
Propensity Score	1 (3.03) [0.054 - 0.15]
Embedded in learning process	1 (3.03) [0.054 - 0.15]
Unclear	1 (3.03) [0.054 - 0.15]
Hyperparameter tunning reported	
Yes	2 (6.06) [0.017 - 0.2]0
No	7 (21.21) [0.11 - 0.38]
Unclear	1 (3.03) [0.054 - 0.15]
Variable importance reported [a]	
Shapley Value	2 (6.06) [0.017 - 0.20]
By Random Forest	2 (6.06) [0.017 - 0.20]
Weights/correlation	4 (12.12) [0.048 - 0.27]
Gain information	1 (3.03) [0.054 - 0.15]
None	3 (9.09) [0.031 - 0.24]
Penalization methods used	
Not reported	10 (30.3) [0.17 - 0.47]

Abbreviations: TF-IDF, term frequency-inverse document frequency.

[a] Counts are absolute numbers with column percentages in parentheses. The percentages sometimes do not add up to 100% because studies developed more than one model.

Variable Importance and hyperparameters

The variable importance scores provide valuable information on the extent to which each variable contributed to the prediction model (Probst et al, 2019). Despite our sample of studies being small, we found a heterogeneity of information about variable importance. Three studies did not provide any information about scores for variables (9.09% [95% CI 0.031 - 0.24]). For 4/33 (12.12% [95% CI 0.048 - 0.27]) the importance weights of variables/correlations were used to report variable importance to the models. Shapley values, another method to determine importance, were used in two studies (6.06% [95% CI 0.017 - 0.20]). Other methods informed by studies to determine variable importance are defined in Table 8. Hyperparameters (including default settings of models) were not reported in 7/10 (70% [95% CI 0.40 - 0.89]) studies. The most described strategy for hyperparameter optimization was cross-validation (n=4/10, 40% [95% CI 0.17 - 0.69]). Seven studies did not report any information about hyperparameter optimization (n=7/10, 70% [95% CI 0.40 - 0.89]), as shown in Table 7.

Performance metrics

The most used measure for the extracted models was the area under the Receiver Operating Characteristic curve (AUC/ROC) (n=15/33, 15.15% [95% CI 0.30 - 0.62]) to describe the discriminative ability of the proposed models (Table 9). Few methods for measuring agreement between predictions and observations (also called calibration) were used in the selected studies. Only four models used a calibration plot (12.12%, [95% CI 0.048 - 0.27]). Other measures of calibration used were calibration slope and calibration-in-the-large. General metrics were found in most studies for the developed models, such as accuracy (n=25/33, 75.75% [95% CI 0.59 - 0.87]) and F1-score (n=12/33, 36.36% [95% CI 0.22 - 0.53]).

Table 9 - Performance measures reported by included studies

	All extracted models (n = 33)
Key items	**n (%) [95% CI]**
Calibration [a]	
Calibration plot	4 (12.12) [0.048 - 0.27]
Calibration slope	1 (3.03) [0.0054 - 0.15]
Calibration in the large	1 (3.03) [0.0054 - 0.15]
None	5 (15.15) [0.067 - 0.31]
Discrimination	
AUC/AUC-ROC	15 (45.45) [0.30 - 0.62]
AUPRC	8 (24.24) [0.13 - 0.41]
Min(Re,Pr)	3 (9.09) [0.031- 0.24]
C-statistic	1 (3.03) [0.0054 - 0.15]
None	1 (3.03) [0.0054 - 0.15]
Classification	
Sensitivity	12 (36.36) [0.22- 0.53]
Specificity	12 (36.36) [0.22- 0.53]
Recall	9 (27.27) [0.15- 0.44]
Precision	8 (24.24) [0.13- 0.41]
Overall [a]	
Predictive values	1 (3.03) [0.0054 - 0.15]
AUC difference	2 (6.06) [0.017 - 0.2]
Accuracy	25 (75.75) [0.59 - 0.87]
F1-score	12 (36.36) [0.22 - 0.53]
Youden Index	1 (3.03) [0.0054 - 0.15]

Abbreviations: AUC/ROC, Area Under the Receiver Operating Characteristic Curve, AUPRC, Area Under the Precision-Recall Curve, Min (Re, Pr), Minimum value between Recall and Precision.

[a] Counts are absolute numbers with column percentages in parentheses. The percentages sometimes do not add up to 100% because studies developed more than one model.

Predictive performance

Studies that reported their discriminative abilities of the proposed models had solid results (AUC next to 1) with an internally validated median AUC of 0.91 (IQR 0.76-

0.98; range 0.68 - 0.98). For calibration and overall performance metrics, as shown in Table 10.

Table 10- Predictive performance of all extracted models [a]

	All extracted models (n = 33)	
Key items	**Reported, n (%)**	**Apparent performance**
		Median [IQR], range
Calibration		
Slope	2 (6.06)	Not significant
Calibration-in-the large	1 (3.03)	Not significant
Pearson chi-square	1 (3.03)	Not significant
Discrimination		
AUC	15 (45.45)	0.98 [0.84 - 0.98], 0.68 to 0.98
AUPRC	3 (9.09)	Not significant
AUROC	3 (9.09)	Not significant
Accuracy	22 (66.66)	0.81 [0.8 - 0.92], 0.79 to 0.96
F-Measure	11 (33.33)	0.84 [0.84 - 0.92], 0.45 to 0.92
Min(Re, Pr)	3 (9.09)	Not significant
Sensitivity	18 (54.54)	0.90 [0.69 - 0.93], 0.69 to 0.94
Specificity	18 (54.54)	0.89 [0.87 - 0.94], 0.87 to 0.99
Precision	10 (30.3)	0.83 [0.83 - 0.93], 0.77 to 0.99
Recall	7 (21.21)	0 [0.85 - 0.85], 0.82 to 0.92

Abbreviations: AUC/ROC, Area Under the Receiver Operating Characteristic Curve, AUPRC, Area Under the Precision-Recall Curve, Min (Re, Pr), Minimum value between Recall and Precision.

[a] Counts are absolute numbers with column percentages in parentheses. The percentages sometimes do not add up to 100% because studies reported more than one option.

Internal validation and external validation

Nine among 10 studies (88.9% [95% CI 0.60- 0.98]) internally validate their models, splitting samples into a training and test set. The train-test set was split randomly into

5/10 studies (50% [95% CI 0.24 - 0.76]) and 2/10 studies used a nonrandom split (20% [95% CI 0.057, 0.51]). One study reported bootstrapping on a test set without citing the number of iterations. Five studies that performed cross-validation (50% [95% CI 0.24 - 0.76]), all of them used nested cross-validation (100% [95% CI 0.57 - 1]). For further details, see Table 4. Only three studies performed an external validation of their models (30% [95% CI 0.11 - 0.60] by using independent datasets to validate their models (100% [95% CI 0.44 – 1]).

Model availability

We did not find any studies that created an online calculator or web system containing some way to use the developed models. We found a repository for data in five studies (n=5/10, 50% [95% CI 0.24 - 0.76]), and in two studies we did not find any information about data, code, and even a detailed description of model construction (n=2/10, 20% [95% CI 0.057 - 0.51]). The presentation of the models in detail with flowcharts or other images that convey the architecture of the solution proposed in the study was found in eight articles (80% [95% CI 0.49 - 0.94]). We found two studies that reported a repository for accessing and reading the source code of the developed model (Table 1).

Discussion

Principal findings

The present review aimed to identify and analyze predictive and prognostic models developed using machine learning techniques for children and adolescent who had COVID-19. Firstly, a notable finding was the low number of studies found that utilized machine learning models for predicting various outcomes in children and adolescents. This highlights the need for further studies of this nature in the field of pediatrics.

Despite obtaining a low number of studies in this review, the quantity of machine learning models found in the selected studies was diverse. The most used models were tree-based models, such as XGBoost, decision trees, and Categorial Boosting (CatBoost) (14-15). XGBoost is an optimized gradient boosting algorithm that excels in handling complex datasets and achieving high predictive accuracy. It utilizes a combination of gradient boosting and regularization techniques to produce strong predictive models. XGBoost is widely recognized for its scalability, speed, and effectiveness in a variety of machine learning tasks. In another spectrum of machine learning, neural network models were also utilized in the selected studies. An example of a neural network model is the multilayer perceptron. A multilayer perceptron is a type of artificial neural network consisting of multiple layers of interconnected neurons (16). It is commonly used for non-linear regression and classification tasks. The network utilizes forward propagation to process input data and backpropagation to adjust the weights and biases during the training process.

Our findings suggest that machine learning techniques have potential for developing accurate predictive models across various clinical fields. For instance, several studies demonstrated high accuracy rates for predicting outcomes including disease diagnosis or prognosis. These models could be used to improve patient care by identifying high-risk individuals who may benefit from early interventions or personalized treatment plans.

Despite the promising results of some studies, we found that there was a lack of consistency in reporting model development and validation procedures across the selected articles. For instance, some studies did not provide detailed information about data sources or model construction methods. This lack of transparency can hinder reproducibility and limit the generalizability of the models to other populations or settings.

Concerning data sources, there are several biases existing in datasets used for machine learning model constructions. Bias, a statistical term, denotes when a model fails to provide an accurate representation of the population. Some biases present in datasets include:

- **Selection Bias**: This bias arises when data from a specific part of the population is used, not representing the entire target population of the study.

To mitigate this bias, it is essential to audit the dataset, ensuring samples accurately represent the study's target population.

- **Overgeneralization Bias**: Researchers encounter this bias when assuming that observations in their dataset mirror those in any dataset aimed at assessing the same problem. To address this issue, external validation is crucial for evaluating model performance.

- **Automation Bias**: This bias occurs when researchers heavily rely on automation tools for data processing before model training. Complete trust in these tools is discouraged; it's vital to verify correct data transformation outcomes.

- **Sampling Bias**: This bias occurs when sampling techniques are not used to balance classes within the dataset. This may lead to models with high accuracy in classifying the most represented class in the dataset.

Still regarding to biases in the data, the most common inconsistency observed in the identified articles pertains to the failure to share modified training data. Researchers should elucidate the state of data post-modifications made for training, including the removal of erroneous features, handling of features with substantial missing data, categorical variable encoding, data sampling, and other relevant procedures. Mere mention of using data from a specific website is insufficient. Without this crucial information, the assessment of the actual data employed in the studies becomes challenging. This can lead to indications of biases in the models and render them less interpretable. Studies that do not disclose their data and source code make the research less transparent.

To address these issues, future studies should follow established guidelines for developing and reporting predictive models (e.g., TRIPOD statement) (17). Additionally, researchers should consider external validation of their models to assess their performance in independent datasets (18).

Another consideration is the ethical implications of using machine learning models in clinical practice. For instance, there is a risk of perpetuating bias or discrimination if the models are trained on biased data or if they are not validated across diverse populations (19). Therefore, it is crucial to ensure that these models are developed and used ethically and responsibly.

Another finding from our review is that the majority of the selected studies used administrative databases as their primary data source. This suggests that machine learning techniques may be particularly useful for analyzing large-scale administrative datasets to identify patterns and predict outcomes.

The machine learning models have a potential impact on clinical decision-making. These models have shown promise for improving patient outcomes by identifying high-risk individuals or predicting disease progression. However, the models should not be viewed as a replacement for clinical judgment or human expertise (20). Instead, they should be used as a tool to support clinical decision-making and improve patient care.

There is a deficiency in the way the selected studies reported data in the models. The limitations include inadequate reporting of sample sizes, missing information about hyperparameter tuning, lack of implementation details, and performance measures of the models. These issues are important for reproducibility purposes (37).

Few studies employed cross-validation techniques in model development. Cross-validation helps to prevent the phenomenon of overfitting (21), where the model achieves 100% accuracy on the test data, which represents the model's development data that has not been seen by the model before. However, if the test data happens to be identical to the training data, it is necessary to train and test the model using different folds of the data. Cross-validation divides the model development data into multiple folds, using each fold as both training and testing data. The lack of cross-validation can lead to inaccurate information regarding the performance of the models.

The most commonly used method for predictor selection in the selected studies was Spearman correlation. Few studies discussed techniques for dimensionality reduction of predictors, although most studies had a low number of features for model development. The selected studies did not provide clear information about missing data and how they handled it. Many methodological details in the majority of studies were unclear. Several studies did not make their code or data available in separate repositories for other researchers to read and reproduce the analysis. Many studies did not report information regarding the calibration and discrimination of the models. It is important to report data about the calibration and discrimination of a machine learning model because these metrics provide insights into the model's performance and reliability. Calibration measures the agreement between the predicted probabilities and the observed outcomes, indicating whether the model's predictions are well-calibrated and accurate. Discrimination, on the other hand, assesses the model's ability to distinguish between different outcomes or classes, indicating its predictive power. Reporting these metrics allows researchers and practitioners to evaluate the model's effectiveness, identify potential biases or limitations, and compare its performance against other models or benchmarks. Ultimately, it promotes transparency, reproducibility, and informed decision-making in utilizing machine learning models.

The studies did not provide a solid contribution to the medical community as they did not create any website or other means for physicians and other interested parties to test the model. There is a need for closer collaboration between this emerging field of evidence-based medicine and practicing clinicians. The availability of models is crucial for other physicians to provide feedback on the performance of the models developed for data specific to their regions.

No selected study provided information on the prediction horizon of the models. This type of information can be important for the clinical field to understand the validity of the predictions made.

It is worth noting the lack of external validation to effectively test the selected models with unseen data. However, obtaining external validation data can be challenging, and testing models with multiple sources requires time and effort to acquire and organize large databases for evaluation by machine learning models.

The nature of the data was not widely discussed in the majority of articles. As important as the model itself, the quality and preparation of the data used for training greatly influence the model's performance. If the data is not properly prepared before training, biases may be introduced, affecting the model's true performance. Few studies mentioned how the data were treated in terms of their nature (continuous, discrete, etc.) and how the data were encoded for evaluation by the developed models.

How models were externally validated

In the three studies that use external validation to validate their models, the procedure has been conducted to assess the model's real-world applicability. The studies conducted external validation, adapting to their specific dataset characteristics. For models with small sample sizes, the researchers in the first study employed data splitting, allocating a portion of the dataset for training and another for validation. Additionally, they acquired external data from independent sources to further validate the model's performance. Key performance metrics, such as accuracy and precision, were calculated and compared between the internal and external datasets, ensuring a comprehensive assessment of generalization of results. In the second and third study, addressing models with large sample sizes, adopted a similar approach, splitting their dataset into training and validation subsets. They emphasized the importance of external validation, even with large data, by obtaining an independent and unseen dataset. Performance metrics were evaluated on both the internal and external validation datasets. Data splitting was complemented by techniques such as k-fold cross-validation to maximize data utilization. Since all studies report good metric values with tests with the external validation datasets, this can exemplify the importance of external validation in machine learning research, contributing to the transparency and real-world applicability of their findings.

Traditional Statistical Models versus Machine Learning Models

Traditional statistics has greater transparency and interpretability of relationships between different variables in the data, clearly showing insights between dependent and independent variables. On the other hand, machine learning models can learn different relationships between data that were not detected by traditional statistical models, but this is not the focus. Until recently, developers paid little attention to the explainability of machine learning models. The models were seen as black boxes. This scenario has changed, and today's models are more explicit about their results. However, the aim of machine learning models is different from traditional statistical models. The aim of these models is that from a set of data that the model has never seen, it is able to classify that data correctly or predict something correctly as if it were a human being, machine learning models are oriented towards the result and the final performance of the prediction.

For example, when using a diagnostic system for a disease that uses machine learning, the aim is for the doctor to enter the patient's data into the system and it will tell them whether the patient is likely to have the disease, showing which variables contributed most to that outcome. These models often see different relationships between the data compared to traditional statistics, as the focus is on providing an answer with a higher degree of accuracy for the task proposed to the model. For the same set of data, machine learning models often find different relationships between the data than statistical models. This is because for the model to give the correct answers as to which classes the data belongs to, the variables that are important to it are different.

We will delve into a comparison of machine learning models and traditional statistical models regarding performance and utility, highlighting the strengths and limitations of both approaches (39-41).

Strengths and Limitations of Traditional Statistical Models:

- Statistical models are designed for inference about the relationships between variables. They are used to identify the underlying patterns and relationships in the data and establish both the scale and significance of the relationship.
- Statistical models explicitly specify a probabilistic model for the data and identify variables that are usually interpretable and of special interest, such as effects of predictor variables.
- Statistical models are best suited for small to medium-sized datasets.

- Statistical models require a lot of assumptions to identify the underlying relationships between variables.
- Statistical models presuppose that the input variables are not highly associated with one another and do not exhibit multicollinearity.
- Certain statistical models rely on the sample size being sufficiently big to guarantee precise parameter estimates.

Strengths and Limitations of Machine Learning Models:

- Machine learning models are designed to make the most accurate predictions possible. They are built for providing accurate predictions without explicit programming.
- Machine learning models can provide better predictions than statistical models.
- Machine learning models are more empirical and do not impose relationships between predictors and outcomes, nor isolate the effect of any single variable.
- Machine learning models are best suited for large datasets. Machine learning models are more difficult to understand and explain than statistical models.
- Machine learning models do not provide a level of interpretability that is possible with statistical models.

Choosing between machine learning models and traditional statistical models depends on the purpose of the analysis. If the goal is to find and explain the relationships between variables, statistical models are the better approach. If the goal is to make accurate predictions, machine learning models are the better approach.

Comparison to previous studies

To the best of our knowledge, at the present moment of writing the results of this study, we did not find another study that has conducted a systematic review to identify the methodological conduct and study design of research utilizing prediction models for outcomes in children and adolescents using machine learning algorithms. However, in other similar studies that evaluated machine learning models for adult patients, similar issues regarding methodological conduct and reporting have been identified in various reviews that have explored different machine learning techniques (22-24). Neglected aspects such as missing data, sample size, calibration, and model availability have been consistently observed (22, 24-26). In a review examining the trends of prediction models utilizing electronic health records (EHR), it was noted that the utilization of ensemble models increased from 6% to 19% (27). Another comprehensive review focusing on prediction models for hospital readmission revealed a substantial growth in the application of algorithms including Support Vector Machine (SVM), Random Forest (RF), and Neural Networks (NN), with an increase from none to 38% over the past 5 years (28). Additionally, the adoption of methods to address class imbalance in EHR datasets increased from 7% to 13% (27).

Limitations of this study

The information extracted in our study was solely based on the content reported in the articles. Regrettably, only a small number of articles provided the essential information required by reporting guidelines, making the process of data extraction challenging (29). Additionally, there was inconsistency in the terminology used across papers. For instance, the term "validation" was frequently used to describe both tuning and testing (i.e., internal validation), a concern previously identified in a review of studies on deep learning models (30). This highlights the necessity of a uniform terminology for the critical evaluation of machine learning models (31).

In our study, we encountered such limitations that prevented us from conducting a meta-analysis. The scarcity of studies refers to the limited number of relevant studies available, which may arise due to the novelty of the research area, ethical considerations, or limited research resources. Additionally, the heterogeneity among studies, in terms of study design, population characteristics, interventions, or outcome measures, the variation in methodologies and findings across studies may introduce substantial clinical and methodological heterogeneity, making it inappropriate to combine the results quantitatively.

Our data extraction form was primarily drawn based on the items and signaling questions from the TRIPOD and PROBAST tools. Although these tools were initially developed for regression-based prediction models, the majority of items and signaling questions were still applicable to studies on machine learning-based models.

Implications for future research

The extent to which the selected studies aimed to improve clinical care with the developed models or primarily sought to showcase promising results with the proposed models is questionable. There was limited emphasis on aspects including the study's objective, clinical workflow, outcome format, prediction horizon, and clinically relevant performance metrics. Guidelines and meta-epidemiological studies have strongly emphasized the importance of applying optimal methodology and transparent reporting in prediction model studies (32,35). The TRIPOD and PROBAST provide best practice recommendations for the design, conduct, and reporting of prediction models, regardless of the modeling technique employed (12,17,32,33). However, it is crucial to extend these recommendations to include areas such as data preparation, tunability, fairness, and data leakage.

Extensions of PROBAST and TRIPOD specifically designed for artificial intelligence (AI) or machine learning-based prediction models, namely PROBAST-AI and TRIPOD-AI, are currently being developed (31,34). As machine learning continues to gain importance in healthcare, it is highly recommended for future studies to reinforce the adoption of a minimum standard in methodological conduct and reporting to increase the generalizability and applicability of these models (12,17,32, 33).

Furthermore, the limited accessibility of the developed models poses a barrier to conducting independent validation, a crucial step before their integration into clinical practice. Openly sharing the source code and, ultimately, the clinical prediction model itself is a fundamental measure to establish trust and credibility in the application of AI and machine learning in the clinical setting (36).

Conclusion

Our study highlights important considerations when developing and using machine learning models in healthcare settings. Future research should focus on addressing limitations including small sample sizes, inconsistent reporting practices, biases in data sources, and ethical implications to ensure that these models are developed and used responsibly to improve patient care.

REFERENCES

1. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. New England Journal of Medicine. 2019;380(14):1347. doi:10.1056/NEJMra1814259
2. Obermeyer Z, Emanuel EJ. Predicting the future—Big data, machine learning, and clinical medicine. New England Journal of Medicine. 2016;375(13):1216. doi:10.1056/NEJMp1606181
3. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA. 2018;319(13):1317. doi:10.1001/jama.2017.18391
4. Wyatt JC. Clinical data systems: overcoming the barriers to their development. JAMIA. 1996;3(6):408. doi:10.1136/jamia.1996.97046762
5. Moons KGM, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98(9):691. doi:10.1136/heartjnl-2011-301247
6. Oliveira EA, Oliveira MCL, Silva ACS e, et al. Clinical Outcomes of Omicron Variant (B.1.1.529) Infection in Children and Adolescents Hospitalized With COVID-19 in Brazil with Observational Data on the Efficacy of the Vaccines in

Adolescents. The Pediatric Infectious Disease Journal. 2023;42(3):218. doi:10.1097/INF.0000000000003783

7. Oliveira EA, et al. Comparison of the First and Second Waves of the Coronavirus Disease 2019 Pandemic in Children and Adolescents in a Middle-Income Country: Clinical Impact Associated with Severe Acute Respiratory Syndrome Coronavirus 2 Gamma Lineage. The Journal of Pediatrics. 2023;244:178.

8. Vasconcelos MA, Mendonça ACQ, Colosimo EA, et al. Outcomes and risk factors for death among hospitalized children and adolescents with kidney diseases and COVID-19: an analysis of a nationwide database. Pediatr Nephrol. 2023;38:181. doi:10.1007/s00467-022-05588-0

9. Andaur Navarro CL, Damen JAA, van Smeden M, et al. Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models. Journal of clinical epidemiology. 2022.

10. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. The BMJ. 2021;372.

11. Tawfik GM, Dila KAS, Mohamed MYF, et al. A step-by-step guide for conducting a systematic review and meta-analysis with simulation data. Trop Med Health. 2019;47:46. doi:10.1186/s41182-019-0165-6

12. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Med. 2015;13:1. doi:10.1186/s12916-014-0241-z

13. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Annals of internal medicine. 2019;170(1):51. doi:10.7326/M18-1376

14. Darapaneni N, Srinivas P, Reddy KM, et al. Tree Based Models: A Comparative and Explainable Study for Credit Default Classification. 2022 IEEE 9th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON). 2022:1.

15. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016.

16. Car Z, Segota SB, Andelic N, et al. Modeling the Spread of COVID-19 Infection Using a Multilayer Perceptron. Computational and Mathematical Methods in Medicine. 2020.

17. Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162:W1e73.

18. Campagner A, Carobene A, Cabitza F. External validation of Machine Learning models for COVID-

19. Michelson KN, Klugman CM, Kho AN, Gerke S. Ethical Considerations Related to Using Machine Learning-Based Prediction of Mortality in the Pediatric Intensive Care Unit. The Journal of pediatrics. 2022;247:125–128. https://doi.org/10.1016/j.jpeds.2021.12.069

20. Xu Y, Liu X, Cao X, Huang C, Liu E, Qian S, Liu X, Wu Y, Dong F, Qiu CW, Qiu J, Hua K, Su W, Wu J, Xu H, Han Y, Fu C, Yin Z, Liu M, Roepman R, et al. Artificial intelligence: A powerful paradigm for scientific research. Innovation (Cambridge (Mass.)). 2021;2(4):100179. https://doi.org/10.1016/j.xinn.2021.100179

21. Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge, MA: MIT Press. 2016.

22. Dhiman P, Ma J, Andaur Navarro CL, Speich B, Bullock G, Damen JAA, et al. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. BMC Med Res Methodol. 2022;22:1e16.

23. Dhiman P, Ma J, Andaur Navarro C, Speich B, Bullock G, Damen JA, et al. Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. J Clin Epidemiol. 2021;138:60e72.

24. Collins GS, De Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol. 2014;14:40.

25. Damen JAAG, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ. 2016;353:i2416.

26. Bouwmeester W, Zuithoff NPA, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. PLoS Med. 2012;9(5):1e12.

27. Yang C, Kors JA, Ioannou S, John LH, Markus AF, Rekkas A, et al. Trends in the conduct and reporting of clinical prediction model development and validation: a systematic review. J Am Med Inform Assoc. 2022;29(5):983e9.

28. Artetxe A, Beristain A, Gra~na M. Predictive models for hospital readmission risk: a systematic review of methods. Comput Methods Programs Biomed. 2018;164:49e64.

29. Andaur Navarro CL, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review. BMC Med Res Methodol. 2022;22:12.

30. Kim DW, Jang HY, Ko Y, Ko Y, Son JH, Kim PH, et al. Inconsistency in the use of the term ‘‘validation’’ in studies reporting the performance of deep learning

algorithms in providing diagnosis from medical imaging. PLoS One. 2020;15:1e10.

31. Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open. 2021;11(7):e048008.

32. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170:51e8.

33. Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med. 2019;170:W1e33.

34. Collins GS, M Moons KG. Reporting of artificial intelligence prediction models. Lancet. 2019;393:1577e9.

35. Damen JAAG, Debray TPA, Pajouheshnia R, Reitsma JB, Scholten RJPM, Moons KGM, et al. Empirical evidence of the impact of study characteristics on the performance of prediction models: a meta-epidemiological study. BMJ Open. 2019;9(4):1e12.

36. Van Calster B, Wynants L, Timmerman D, Steyerberg EW, Collins GS. Predictive analytics in health care: how can we know it works? J Am Med Inform Assoc. 2019;26(12):1651e4.

37. Philipp Probst, Anne-Laure Boulesteix, and Bernd Bischl. 2019. Tunability: importance of hyperparameters of machine learning algorithms. J. Mach. Learn. Res. 20, 1 (January 2019), 1934–1965.

38. Van Smeden, M., Reitsma, J. B., Riley, R. D., Collins, G. S., & Moons, K. G. Clinical prediction models: diagnosis versus prognosis. Journal of clinical epidemiology.2021;132:142–145. https://doi.org/10.1016/j.jclinepi.2021.01.009

39. Rajula HS, Verlato G, Manchia M, Antonucci N, Fanos V. Comparison of Conventional Statistical Methods with Machine Learning in Medicine: Diagnosis, Drug Development, and Treatment. Medicina (Kaunas, Lithuania). 2020;56(9):455. https://doi.org/10.3390/medicina56090455.

40. Bennett M, Kleczyk EJ, Hayes K, Mehta R. Evaluating Similarities and Differences between Machine Learning and Traditional Statistical Modeling in Healthcare Analytics. IntechOpen. doi: 10.5772/intechopen.105116.

41. Ley C, Martin RK, Pareek A, et al. Machine learning and conventional statistics: making sense of the differences. Knee Surg Sports Traumatol Arthrosc. 2022;30:753–757. https://doi.org/10.1007/s00167-022-06896-6.

3. OBJETIVOS

3.1 OBJETIVO PRINCIPAL

Utilizar algoritmos de Aprendizado de Máquina para identificar as principais características clínicas e fatores de risco para mortalidade de crianças e adolescentes hospitalizados com COVID-19.

3.2 OBJETIVOS SECUNDÁRIOS

1 - Realizar uma revisão sistemática da literatura com o objetivo de analisar estudos que utilizaram modelos preditivos com Inteligência Artificial para estudar características clínicas e fatores de risco relacionados à COVID-19 em crianças e adolescentes.

2 - Utilizar algoritmos de Aprendizado de Máquina para identificar as principais características clínicas e fatores de risco preditivos da gravidade da COVID-19 em crianças e adolescentes hospitalizados com COVID-19.

3 - Comparar o desempenho e a precisão dos modelos para prever óbitos na população-alvo.

4. MÉTODOS

Desenho do estudo

O delineamento da pesquisa é de um estudo de coorte retrospectivo incluindo a análise de todos os casos de COVID-19 em crianças e adolescentes hospitalizados (pacientes com idade inferior a 18 anos) registrados no Sistema de Informação de Vigilância Epidemiológica da Influenza (SIVEP-Gripe) e não hospitalizados incluídos no sistema e-SUS Notifica do Ministério da Saúde-Brasil (MS).

Fonte dos dados

Sistema de vigilância e-SUS Notifica:

Em 27 de março de 2000, o Departamento de Informática do SUS – DATASUS disponibilizou o e-SUS Notifica, ferramenta online para registro de notificação de casos suspeitos e confirmados de síndrome gripal leve relacionada à COVID-19.

Sistema de vigilância SIVEP-gripe:

Sistema de registro de dados de abrangência nacional estabelecido pelo Ministério da Saúde em 2009 para manter vigilância de infecções respiratórias agudas graves no Brasil. O SIVEP-Gripe tem sido a principal fonte de informações sobre as admissões e óbitos hospitalares do COVID-19 no Brasil(1, 2). A notificação do COVID-19 é compulsória no Brasil e o SIVEP-Gripe recebe notificações de pacientes internados em hospitais públicos e privados(3). Para todos os pacientes cadastrados no sistema, os dados relativos às características demográficas e clínicas têm sido registrados sistematicamente.

O Ministério da Saúde do Brasil disponibiliza essas bases de dados na plataforma OpenSUS (https://opendatasus.saude.gov.br/dataset). Portanto, informações detalhadas sobre esses bancos de dados, incluindo formulário de relatório e dicionário de dados, códigos e todos os dados não identificados, como dados de participantes individuais, estão disponíveis publicamente neste site.

Para o presente estudo, baixamos a última versão disponível dos conjuntos de dados em abril de 2023. Para o propósito da presente análise, limitamos o período do estudo de 24 de fevereiro de 2020 a fevereiro de 2023.

O SIVEP-gripe é disponibilizado em arquivos únicos divididos por ano desde 2009. Já os arquivos e-SUS Notifica são divididos por ano e UF da Federação. O Brasil é um país continental com uma população de mais de 200 milhões de pessoas. Além disso, o e-SUS Notifica registra não apenas os casos confirmados de COVID-19, mas todos os casos sintomáticos com sintomas respiratórios ou outros suspeitos de infecção por SARS-CoV-2. No Brasil, o registro é obrigatório; portanto, prestadores de serviços de saúde públicos e privados devem notificar casos suspeitos de COVID-19 e internações.

Assim, devido à grande quantidade de dados disponíveis, o e-SUS Notifica é disponibilizado em arquivos de acordo com os 27 Estados da Federação. Além disso, para alguns Estados populosos, como São Paulo, foram 13 lotes com cerca de 800.000 indivíduos por arquivo. Assim, foram disponibilizados 144 lotes com informações de interesse para nossa análise.

Abordamos esses arquivos passo a passo para obter informações confiáveis pela seguinte metodologia. Primeiro, baixamos a última versão disponível dos conjuntos de dados em abril de 2023; Em seguida, retiramos todos os indivíduos cadastrados fora do período de interesse do nosso estudo; Retiramos indivíduos maiores de 18anos; Por fim, retiramos indivíduos sem informações sobre o teste de COVID-19 ou com testes indisponíveis no momento da análise. Após essas etapas, mesclamos sequencialmente os arquivos por Estado, por Regiões e para todo o país. Por fim, unimos as duas bases de dados (SIVEP-gripe e e-SUS Notifica), reunindo todos os dados em um único arquivo para análise. Antes de combinar os conjuntos de dados, tornamos as variáveis incluídas compatíveis para o processo de mesclagem. Também, antes da análise final, procuramos ativamente por indivíduos duplicados nos conjuntos de dados.

Participantes

Os critérios de inclusão e exclusão para o estudo são os seguintes:

a) Critérios de inclusão

Serão incluídos no estudo todos os pacientes registrados consecutivamente nestas bases de dados, com idade inferior a 18 anos, com um resultado positivo do teste RT-PCR quantitativo (RT-qPCR) ou de antígeno para SARS-CoV-2.

Para ser incluído no banco de dados *SIVEP-Gripe*, o caso deve apresentar quadro clínico de síndrome respiratória semelhante à gripe e pelo menos um dos seguintes critérios: dispneia ou dificuldade respiratória ou saturação de O_2 menor que 95% em ar ambiente ou cianose ou sintomas específicos para crianças (retrações intercostais, batimento de aletas nasais, desidratação e inapetência).

Para ser incluído no banco de dados do *e-SUS Notifica*, o caso deve apresentar uma síndrome gripal leve com a seguinte definição: Indivíduo com quadro respiratório agudo, caracterizado por pelo menos 2 (dois) dos seguintes sinais e sintomas: febre, calafrios, dor garganta, dor de cabeça, tosse, corrimento nasal, distúrbios do olfato ou distúrbios do paladar. Para crianças, além dos itens anteriores, a obstrução nasal também é considerada, na ausência de outro diagnóstico específico.

b) Critérios de exclusão

Pacientes com idade superior a 18 anos de idade. Casos sem confirmação laboratorial de COVID-19.

Após coletado e armazenado dos dados, conforme descrito acima, os dados foram preparados para análise. Esta etapa envolveu tarefas como limpeza dos dados, recodificação, integração de dados, transformação de dados, manipulação de valores ausentes, remoção de valores discrepantes e garantia da consistência dos dados. Após todas estas etapas, de um total de aproximadamente 135 milhões de indivíduos registrados nas bases de dados, foram selecionados 3,521,883 crianças e adolescentes elegíveis para a participação no estudo, como descrito nos fluxogramas abaixo na Figura 1 e Figura 2.

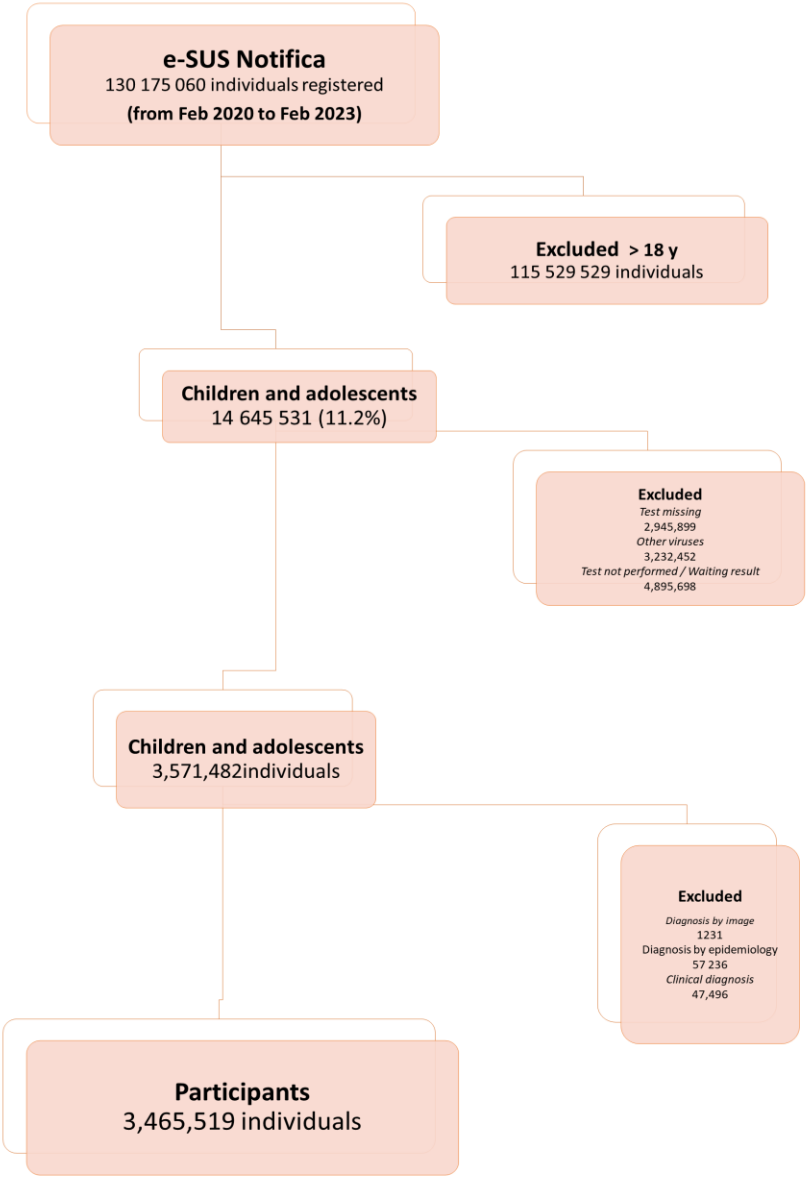


Figura 1 – Digrama mostrando os processos de inclusão e exclusão de participantes no estudo para a base de dados e-SUS Notifica.

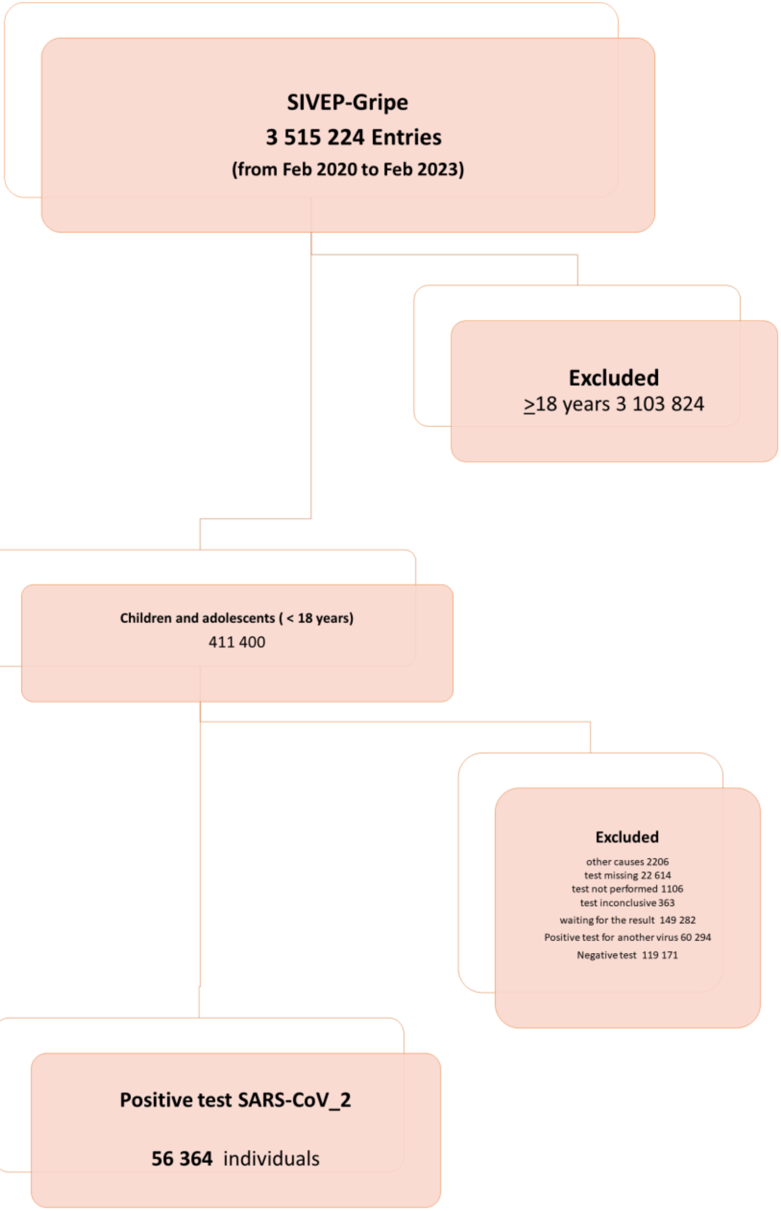


Figura 2 – Digrama mostrando os processos de inclusão e exclusão de participantes no estudo para a base de dados Sivep-Gripe.

Variáveis expositivas

Entre as variáveis expositivas forma incluídas dados clínicos e demográficos.

Dados demográficos: idade, sexo, etnia e regiões do país. O Brasil está geopoliticamente dividido em cinco macrorregiões: Norte, Nordeste, Centro-Oeste, Sudeste e Sul. Essas macrorregiões têm diferenças históricas na capacidade e cobertura social, econômica e do sistema de saúde(4, 5). O Instituto Brasileiro de Geografia e Estatística (IBGE) classifica racialmente a população brasileira em cinco categorias. Essa classificação do IBGE é baseada na cor e os indivíduos são solicitados a se auto identificarem como Branco, Preto, Pardo, Amarelo, ou Indígena(6).

Dados clínicos: data de início dos sintomas, definida como o dia em que o primeiro sintoma ou sinal ocorreu, e a data de admissão quando o paciente foi hospitalizado. Sinais e/ou sintomas de apresentação (febre, tosse, desconforto respiratório, gastrointestinal e saturação de oxigênio reduzida) e presença de comorbidades preexistentes (doença cardíaca, doença pulmonar, asma, doença renal, doença neurológica (incluindo atraso no desenvolvimento), doença hematológica, diabetes, obesidade, deficiência imunológica, malignidade, anormalidades pós-transplante, sindrômicas e cromossômicas) também foram registrados. Para análise, a presença de comorbidade será categorizada em dicotomizada (sim/não) e em quatro níveis (nenhuma, uma, duas e três ou mais condições médicas preexistentes). A evolução clínica foi relatada em termos de suporte respiratório (nenhum, suporte não invasivo de oxigênio e ventilação invasiva), admissão em unidade de terapia intensiva (UTI), recuperação, óbito e situação clínica em andamento. A data do óbito ou alta também foi registrada.

Desfechos

O desfecho primário será o tempo até o evento (mortalidade intra-hospitalar). O tempo de sobrevida foi definido desde o dia da admissão até o evento (óbito ou alta). Como desfechos secundários, também serão avaliados o uso de recursos assistenciais (internação na UTI e suporte respiratório, definidos como nenhum, não invasivo ou invasivo).

Definições dos desfechos

Os seguintes desfechos serão considerados para a análise dos casos com infecção comprovada por SARS-CoV-2:

1. necessidade de suporte respiratório: estratificado em três grupos nenhum, suporte não invasivo de oxigênio e ventilação mecânica invasiva,
2. UTI: Admissão em unidade de terapia intensiva, dicotômica (sim/não)
3. Gravidade da COVID-19: As categorias de gravidade clínica incluem leve [sem necessidade de suporte de oxigênio sem internação na unidade de terapia intensiva (UTI)], moderada (necessidade de suporte de oxigênio sem

ventilação mecânica invasiva) e grave (necessita de ventilação mecânica invasiva ou morte

4. Tempo até a morte (mortalidade intra-hospitalar): Criamos variáveis para as análises de sobrevida de risco competitivo. Por exemplo, a partir dos campos “data de admissão” e “data de alta ou óbito” criamos a variável time_event (em dias de internação). Alternativamente, por falta de informação nestes arquivos, foram utilizados, respectivamente, os campos “data de início dos sintomas” ao invés de “data de admissão” ou a data de fechamento do formulário ao invés de “data de alta ou óbito”. O tempo de sobrevida será definido desde o dia da admissão até o evento (óbito ou alta).

Análise Estatística

A amostra será composta por todos os pacientes pediátricos (idade < 18 anos) com COVID-19 cadastrados nos sistemas de vigilância do MS entre fevereiro de 2020 e fevereiro de 2023.

Para a análise descritiva, serão utilizadas medianas e interquartis ou médias e desvio-padrão para resumir variáveis contínuas e frequências calculadas e proporções para variáveis categóricas. Para comparação de medianas e proporções, serão utilizados, respectivamente, os testes qui-quadrado e teste de Mann-Whitney.

A mortalidade será avaliada por análise de riscos competitivos, utilizando a função de incidência cumulativa (CIF)(7). A alta foi analisada como evento concorrente na análise de riscos competitivos(8). Dados completos não estavam disponíveis para todas as variáveis, especialmente etnia, sintomas na apresentação e comorbidades. Realizamos imputação múltipla usando todos os preditores mais o CIF para o desfecho primário. Isso envolve a criação de várias cópias dos dados e a imputação dos valores ausentes para cada conjunto de dados com valores sensíveis selecionados aleatoriamente de sua distribuição prevista. Dez imputados serão gerados usando o pacote de equações da cadeia de imputação múltipla (MICE) do software R. Combinamos os resultados das análises de cada um dos valores imputados usando as regras de Rubin para produzir estimativas e intervalos de confiança que incorporam a incerteza dos valores imputados(9). Para aqueles casos com dados ausentes sobre um determinado sintoma ou comorbidade, assumimos que a condição clínica estava ausente. Informações detalhadas sobre o gerenciamento de dados ausentes são fornecidas no artigo apresentado na seção de resultados desta Tese.

Desenvolvimento do Modelo de Previsão de Risco (estatística convencional)

Desenvolvemos um modelo de predição clínica e um sistema de pontuação de risco baseado em pontos seguindo as diretrizes fornecidas por Austin et al.(10) para modelos na presença de riscos concorrentes. A coorte de desenvolvimento será derivada dos casos admitidos na primeira onda da COVID-19 no Brasil.

Os dados não disponíveis para variáveis, especialmente etnia e sintomas de apresentação serão imputados para as análises. Para comorbidades, consideramos os valores faltantes como ausência do quadro clínico. Serão usados medianas e intervalos interquartis ou médias e desvio-padrão para resumir as variáveis contínuas e será calculada frequências e proporções para as variáveis categóricas. Será examinado o desenvolvimento espacial e temporal da epidemia de COVID-19 (total de casos e mortes) em todo o país, dividindo nossa amostra em quartis. A mortalidade será avaliada por análise de riscos concorrentes, usando a função de incidência cumulativa (CIF). A alta hospitalar será analisada como um evento concorrente pela análise de riscos concorrentes. O modelo de sub-distribuição proporcional de riscos de Fine e Gray será ajustado para estimar o efeito das covariáveis na mortalidade. As covariáveis usadas para análises multivariadas serão selecionadas com base em sua significância na análise univariada (p <0,10). As variáveis do modelo final com valor de p <0,05 serão consideradas estatisticamente significativas. Os resultados serão expressos como taxas de risco ajustadas (HR) e seus intervalos de confiança de 95% (CI).

Desenvolvimento do Modelo de Previsão de Risco (inteligência artificial)

Outro ponto a ser abordado neste projeto é a utilização de algoritmos de aprendizado de máquina para comparar e predizer desfechos clínicos de crianças e adolescentes com COVID-19. Este conjunto de dados incluindo quase 4 milhões de pacientes pediátricos, seguramente configura uma das maiores bases de dados disponíveis sobre este tema em todo o mundo. Para a análise de banco de dados deste porte (na literatura de análise dados chamados de BIG DATA), as técnicas de inteligência artificial podem construir com relevantes informações(11). Esta etapa exige a aplicação de várias técnicas de análise de dados para obter insights clinicamente relevantes. Deve envolver análise descritiva, análise exploratória de dados, análise estatística convencional, mineração de dados, aprendizado de máquina ou modelagem preditiva(12). Devem ser usados algoritmos e metodologias apropriados para analisar os dados e testar suas hipóteses. A hipótese investigada é de que possa haver diferentes fatores relacionados aos desfechos clínicos durante a pandemia(13). Para entender esses fatores e, principalmente, lidar com possíveis mudanças na importância desses fatores em diferentes períodos, trabalharemos com modelos de causalidade em aprendizado de máquina. O conceito de causalidade em aprendizado de máquina vai além de simplesmente prever o desfecho baseado em um conjunto de fatores. Isso porque a previsão do desfecho por si só pode não ser tão relevante quanto seus efeitos na proposta de intervenções – seja através de mudanças nos protocolos de tratamentos ou mesmo em políticas públicas – para reduzir, no caso dessa proposta, mortalidade, a necessidade de UTI ou uso de respiradores na população de interesse.

Modelos de causalidade vão além de detectar simples correlações nos dados, e trabalham com um grafo causal(14). Um grafo causal pode ser aprendido automaticamente a partir de um conjunto de dados, e posteriormente validado e refinado por especialistas do domínio. Tendo o grafo e um modelo de aprendizado capaz de inferir os desfechos, é possível planejar intervenções e simular contrafactuais.

Mais especificamente, um modelo de aprendizado de máquina é associativo, ou seja, capaz de responder perguntas utilizando padrões encontrados nos dados, como: quais os fatores que levam um paciente a UTI? Já um modelo de intervenção, que depende tanto do modelo associativo quanto do grafo causal, consegue responder perguntas do tipo "e se?". Por exemplo, e se os pacientes tivessem sido vacinados contra COVID-19, isso diminuiria suas chances de ir para UTI? Por último, tendo esses dois elementos e o modelo causal, conseguimos gerar contrafactuais, ou seja, simular se realmente ao receber a vacina contra COVID-19, um menor número de pacientes necessitaria de UTI.

REFERÊNCIAS

28. Bastos LS, Ranzani OT, Souza TML, Hamacher S, Bozza FA. COVID-19 hospital admissions: Brazil's first and second waves compared. Lancet Respir Med. 2021;9(8):e82-e3.

29. Ranzani OT, Bastos LSL, Gelli JGM, Marchesi JF, Baiao F, Hamacher S, et al. Characterisation of the first 250,000 hospital admissions for COVID-19 in Brazil: a retrospective analysis of nationwide data. Lancet Respir Med. 2021;9(4):407-18.

30. Bastos LS, Niquini RP, Lana RM, Villela DAM, Cruz OG, Coelho FC, et al. COVID-19 and hospitalizations for SARI in Brazil: a comparison up to the 12th epidemiological week of 2020. Cad Saude Publica. 2020;36(4):e00070120.

31. Castro MC, Massuda A, Almeida G, Menezes-Filho NA, Andrade MV, de Souza Noronha KVM, et al. Brazil's unified health system: the first 30 years and prospects for the future. Lancet. 2019;394(10195):345-56.

32. Rocha R, Atun R, Massuda A, Rache B, Spinola P, Nunes L, et al. Effect of socioeconomic inequalities and vulnerabilities on health-system preparedness and response to COVID-19 in Brazil: a comprehensive analysis. Lancet Glob Health. 2021;9(6):e782-e92.

33. Baqui P, Bica I, Marra V, Ercole A, van der Schaar M. Ethnic and regional variations in hospital mortality from COVID-19 in Brazil: a cross-sectional observational study. Lancet Glob Health. 2020;8(8):e1018-e26.

34. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Stat Med. 2007;26(11):2389-430.

35. Fine JP, Gray RJ. A proportional hazards model for the sub-distribution of a competing risk. J Am Stat Assoc. 1999;94:496-509.

36. Schafer JL. Multiple imputation: a primer. Stat Methods Med Res. 1999;8(1):3-15.

37. Austin PC, Lee DS, D'Agostino RB, Fine JP. Developing points-based risk-scoring systems in the presence of competing risks. Stat Med. 2016;35(22):4056-72.

38. Marchezini GF, Lacerda AM, Pappa GL, Meira W, Jr., Miranda D, Romano-Silva MA, et al. Counterfactual inference with latent variable and its application in mental health care. Data Min Knowl Discov. 2022;36(2):811-40.

39. Andaur Navarro CL, Damen JAA, van Smeden M, Takada T, Nijman SWJ, Dhiman P, et al. Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models. J Clin Epidemiol. 2023;154:8-22.

40. Sankaranarayanan S, Balan J, Walsh JR, Wu Y, Minnich S, Piazza A, et al. COVID-19 Mortality Prediction From Deep Learning in a Large Multistate

Electronic Health Record and Laboratory Information System Data Set: Algorithm Development and Validation. J Med Internet Res. 2021;23(9):e30157.

41. Pearl J. Causal inference in the health sciences: a conceptual introduction 2001; 2: 189-220. Health Services and Outcomes Research Methodology. 2001;2:189-220.

5. RESULTADOS e DISCUSSÃO

5.1 Artigo Original

Clinical characteristics and risk factors that affect children mortality with COVID-19: a nationwide retrospective cohort study in Brazil using machine learning

Adriano Lages dos Santos[1,2,*], Maria Christina L. Oliveira, MD, PhD[2], Enrico A. Colosimo, PhD[3], Clara C. Pinhati, MD[2], Stella C. Galante, MD[2], Hercílio Martelli-Júnior, PhD[4], Robert H. Mak[5], Ana Cristina Simões e Silva, MD, PhD[2], Eduardo A. Oliveira, MD, PhD[2]

[1] Department of Engineering and Informatics, Federal Institute of Science and Technology of Minas Gerais (IFMG), Belo Horizonte, MG, Brazil.
[2] Department of Pediatrics, Health Sciences Postgraduate Program, School of Medicine, Federal University of Minas Gerais (UFMG), Belo Horizonte, MG, Brazil.
[3] Department of Statistics, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil.
[4] Health Science/Primary Care Postgraduate Program, State University of Montes Claros (Unimontes), Montes Claros, MG, Brazil.
[5] Department of Pediatrics, Rady Children's Hospital, University of California, San Diego, CA, USA.
[*] Corresponding author at: Federal Institute of Science and Technology of Minas Gerais (IFMG), Belo Horizonte, MG, Brazil. E-mail (adriano.santos@ifmg.edu.br)

Abstract

Background

The COVID-19 pandemic has led to the use of advanced digital technologies such as artificial intelligence (AI) to predict mortality in adult patients. Nevertheless, machine learning (ML) models capable of predicting outcomes in children and adolescents are scarce. The primary objective of this study was to develop several ML models for forecasting mortality in hospitalized children and adolescents with confirmed COVID-19, and to assess their practicality in relation to extensive databases.

Method

In this cohort study, we used the SIVEP-Gripe dataset, a public resource maintained by the Ministry of Health, to track severe acute respiratory syndrome (SARS) in Brazil. To create subsets for training and testing the machine learning (ML) models, we divided the primary dataset into three parts. Using these subsets, we developed and trained 12 ML algorithms to predict the outcomes. We assessed the performance of these models using various metrics such as accuracy, precision, sensitivity, recall, and area under the receiver operating characteristic curve (AUC).

Results

Among the 37 variables examined, 24 were found to be potential indicators of mortality, as determined by the chi-square test of independence. The LR algorithm achieved the highest performance, with an accuracy of 92.5% and an AUC of 80.1%, on the optimized dataset. GBC and ADA closely followed the LR algorithm, producing similar results. Our study also revealed that baseline reduced oxygen saturation, presence of comorbidities, and older age were the most relevant factors in predicting mortality in hospitalized children and adolescents.

Conclusions

The use of ML models can be an asset in making clinical decisions and implementing evidence-based patient management strategies, which can enhance patient outcomes and overall quality of medical care. LR, GBC, and ADA models have demonstrated efficiency in accurately predicting mortality in COVID-19 pediatric patients.

Keywords: COVID-19; artificial intelligence, machine learning; children; death prediction.

Introduction

Since the onset of the COVID-19 pandemic, the global community has witnessed remarkable progress in artificial intelligence (AI), particularly in machine learning (ML) algorithms, such as large language models (LLMs) [1,2]. These models have played a crucial role in assisting researchers globally in devising innovative solutions to the diverse challenges in the healthcare field. The utilization of generative AI to provide diagnoses and prognoses for various diseases across different medical specialties has experienced substantial growth in recent years [3]. This growth also encompasses the application of ML algorithms to predict various outcomes of COVID-19.

Despite the extensive use of ML algorithms in diagnostics and prognosis of COVID-19 in adults, there is a notable lack of studies specifically for children and adolescents. This significant gap needs to be addressed [4]. Early identification of high-risk patients is vital for reducing the strain on healthcare systems and formulating public policies aimed at minimizing death and mortality rates. The development of a predictive model that anticipates unfavorable outcomes in patients with COVID-19 could aid in the efficient allocation of scarce medical resources, improve healthcare quality, and optimize patient management strategies.

The objective of this study was to examine the ability of machine learning (ML) models to predict mortality or hospital discharge in a cohort of hospitalized children and adolescents with laboratory-confirmed COVID-19. To accomplish this, this study utilized data from a comprehensive nationwide dataset provided by the Brazilian government. This analysis aimed to determine the most critical predictors for ML models and the criteria used by these models when making predictions. Additionally, this study evaluated the effectiveness of these models in forecasting deaths resulting from COVID-19.

Methods

Study design and dataset description

In this retrospective cohort study, we used data from the Surveillance Information System (SIVEP-Gripe) to investigate COVID-19 cases among hospitalized individuals aged < 18 years. In 2009, the Ministry of Health established a nationwide database to register severe acute respiratory infections in Brazil. SIVEP-Gripe has served as the primary repository for information on COVID-19 hospitalizations in the country. The reporting of hospitalizations due to COVID-19 is mandatory in Brazil, with SIVEP-Gripe receiving notifications from both public and private hospitals. The database systematically recorded the demographic and clinical findings of all enrolled patients. Our analysis covered the period from epidemiological week 08 (commencing on February 16, 2020) to epidemiological week 08, 2023 (ending on February 19, 2023). We included all consecutively registered patients under the age of 18 years who tested positive for SARS-CoV-2 using quantitative RT-PCR (RT-qPCR) or antigen tests and had been admitted to a hospital.

Data preparation

Over the designated period, 56,330 patient records with verified RT-PCR test outcomes for SARS-CoV-2 infection were documented. After completing the required procedures for preprocessing data for the machine learning algorithms, 24,097 records were chosen for the training, validation, and testing stages of the models. The subset of data from the SIVEP-Gripe dataset, which includes information about children and adolescents, is hereafter referred to as the SIVEP-Kids dataset.

In the SIVEP-Kids dataset, there are 37 primary features in four main categories: patient demographics (four features), clinical features (12 features), personal disease/comorbidity history (14 features), virus strain information (one feature), vaccine information (two features), a feature indicating the number of different comorbidities a patient has, a feature indicating whether a patient has comorbidities or not, a feature categorizing the number of comorbidities a patient has, a feature indicating the time of the outcome, and an output variable (0: survived and 1: deceased) for COVID-19 patients. The primary features of the SIVEP-Kids dataset are presented in Supplementary Table 1.

Regarding the primary features presented in the SIVEP-Kids dataset, the ethnicity feature had five categories: Asian, Black, Brown, Indigenous, and White. Similarly, the region was divided into five regions: Central West, North, Northeast, South, and Southeast. The virus strain feature identified four types of strains in the dataset: ancestral, delta, gamma, and Omicron. For features 6 through 32, all are of the nominal type and have values of "Yes" or "No," indicating the presence or absence of a specific disease or clinical condition in the patient. The total comorbidity feature records the total number of comorbidities per patient in the SIVEP-Kids dataset. Feature 34 (number of vaccine doses) had valid values ranging from zero to three doses. Feature 38 is the target variable of this study, with three types of outcomes: discharge, death, and in-hospital, with the latter referring to cases in which the patient is still in the hospital in an ongoing clinical situation. In the present study, we considered only two types of outcomes in the target variable: death and discharge. This decision aimed to enhance the accuracy of machine learning algorithms, as multi-class problems (those with more than two classes in the target variable) are challenging and tend to reduce the accuracy of ML models because of the large number of decision boundaries to navigate, often failing to accurately separate instances across more than two classes [5,6]. Detailed information on the clinical, demographic, and epidemiological covariates recorded in the SIVEP-Gripe is described elsewhere [7, 8].

Data pre-processing

Data preprocessing is a critical step in addressing the influence of irrelevant, redundant, and unreliable data, ultimately improving data quality and resolving inconsistencies [9]. In this study, data preprocessing was conducted prior to training the machine learning models. Initially, the patient records with missing data were removed from the dataset. For example, records of sex, ethnicity, and reduced oxygen saturation were excluded if any missing values were detected. Missing values for the target variable were treated as the absence of the outcome of interest (death). Additionally, we utilized categorical encoding to transform nominal data into numerical representations. By applying one-hot encoding, we ensured that our analysis was guided by intrinsic relationships within the data rather than by the constraints of non-numerical representations [10].

After applying the criteria for excluding data in the pre-processing step, we obtained a final sample consisting of 24,097 records. The dataset comprised 22,586 and 1,511 cases in the discharge and death classes, respectively. An imbalanced input distribution can lead to a bias in the results towards the dominant class, potentially skewing model performance and reducing generalizability. To address the problem posed by an imbalanced dataset, we employed the Synthetic Minority Over-sampling Technique (SMOTE) method, as outlined in <https://imbalanced-learn.org/stable/>. The SMOTE algorithm, which is widely utilized for synthetic oversampling, generates artificial samples for the minority class by randomly selecting instances from the minority class and their k-nearest neighbors. In this approach, a random data instance along with its k-nearest neighbors is chosen. Subsequently, the second data instance was selected from this set of k-nearest neighbors [11]. The synthesis of a new sample occurred along the line connecting these two instances as a convex combination. This process was iterated until a balance was achieved between minority and majority classes. The SMOTE method mitigates the risk of overfitting, distinguishing it from the random oversampling technique, and it is recognized for its potential to produce better results [12, 13, 14].

Feature Selection

Chi-square tests were used to discern statistically significant differences between the outcomes of discharged and deceased patients. Feature importance scores derived from XGBoost and random forests (as detailed in Supplementary Figure 1) were utilized to identify the essential variables for forecasting COVID-19 mortality. This methodology aims to increase the interpretability and steadfastness of mortality prediction models.

Feature selection techniques exhibited elevated scores for robust predictors such as overall comorbidities, diminished oxygen saturation, and age. Nevertheless, some disparities were evident in the importance scores between XGBoost and random forest for specific parameters. XGBoost showed considerable importance in reducing oxygen saturation and overall comorbidities, whereas random forest allocated minimal importance. A statistically significant difference ($P < 0.01$) in oxygen saturation and total comorbidities was observed between patients who survived and those who died. Chi-square tests were applied to recognize crucial mortality predictors, demonstrating moderate to high importance in XGBoost and low importance in random forest.

Owing to the inconsistencies observed between the two methods, we opted to select the most pertinent features for training the models using the chi-squared test. Consequently, we developed three distinct datasets to train and validate the machine learning models. These datasets included a dataset with features selected using the chi-squared test, a dataset with features chosen by two pediatricians, and a dataset with all 37 features, according to Supplementary Table 1, except for the target variable. Our objective was to determine the dataset that yielded the most favorable results.

The dataset containing characteristics chosen by pediatricians comprised 17 features: sex, age, ethnicity, region, virus strain, dyspnea, fever, cough, odynophagia, abdominal pain, ageusia, anosmia, respiratory distress, reduced oxygen saturation, total comorbidities, vaccine doses, and nosocomial. The dataset selected by the chi-squared test comprised 24 features: age, ethnicity, region, viral strain, dyspnea, cough, respiratory distress, reduced oxygen saturation, cardiology, pulmonary disease, hypertension, immunosuppression, renal disease, asthma, total comorbidities, comorbidities, dichotomous comorbidities, time for outcome, vaccine doses, hematology, neurology, oncology, Down syndrome, and nosocomial infection.

For the purpose to conducting feature selection calculations using the Chi-square test, XGBoost, and random forest, the Scikit-learn library in its version 1.3.1 was used. The Pycaret library version 3.1.0 was employed for training and validating the models. Statistical significance was set at $P < 0.01$.

Model Development

In this study, a total of twelve machine learning algorithms were employed to develop predictive models. These algorithms included Gradient Boosting (GB), AdaBoost (Ada), CatBoost (Cat), Random Forest (RF), Extreme Gradient Boosting (XGBoost), Extra Trees (ET), Logistic Regression (LR), Linear Discriminant Analysis (LDA), Decision Tree (DT), Naïve Bayes (NB), k-nearest neighbors (KNN), and Quadratic Discriminant Analysis (QDA) [11]. The evaluation process involved the use of k-fold cross-validation, which is known to have low bias and variation. The optimized hyperparameters for the machine learning algorithms are provided in Supplementary Table 2, with constant values maintained across the three variations of the SIVEP-Kids dataset.

The performance of the predictive model was evaluated using various metrics, such as accuracy, precision, sensitivity, F1 score, and area under the ROC curve (AUC). A comprehensive analysis was conducted across all 12 machine learning algorithms to determine the best model for predicting mortality in COVID-19 patients.

Ethical aspects

We assessed data in SIVEP-Gripe, which are de-identified and publicly available. The study was approved by the Federal University of Minas Gerais institutional review board (register 6.127.414). The funding organizations had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript.

Results

Feature Selection

Twenty-four features, comprising demographic and clinical factors, were identified as the most relevant predictors using the chi-square independence test (Table 2). Additionally, Table 2 shows mean decreases in impurity and the importance scores of these variables calculated using the XGBoost and random forest algorithms. The descriptive statistics of these features are summarized in Supplementary Table 3

Table 2 - The significance levels, importance scores, and mean decreases in Gini for the key variables in COVID-19 mortality prediction were computed using XGBoost, Random Forest, and Chi-squared tests.

Nº	Feature Name	Chi-squared test		Random Forest	XGBoost
		X^2	P-value	Mean decrease impurity	Importance Score
1	Age	396.94	< 0.001	0.171	0.029
2	Region	17.02	< 0.001	0.084	0.035
3	Ethnicity	9.59	< 0.001	0.04	0.022
3	Virus Strain	34.25	< 0.001	0.048	0.023
4	Dyspnea	57.69	< 0.001	0.026	0.025
5	Cough	37.89	< 0.001	0.030	0.050
6	Respiratoy distress	79.53	< 0.001	0.025	0.035
7	Oxygen saturation reduced at admission	175.43	< 0.001	0.027	0.125
8	Obesity	66.27	< 0.001	0.005	0.020
9	Cardiology	212.09	< 0.001	0.009	0.029
10	Pulmonary	33.75	< 0.001	0.006	0.025
11	Hypertension	25.17	< 0.001	0.002	0.011
12	Immunosuppression	108.01	< 0.001	0.008	0.032
13	Renal	49.48	< 0.001	0.004	0.016

14	Asthma	18.74	< 0.001	0.007	0.040
15	Total Comorbidities	861.55	< 0.001	0.021	0.106
16	Comorbidities dichotomic	527.13	< 0.001	0.012	0.000[a]
17	Comorbidities categoric	830.74	< 0.001	0.019	0.000[a]
18	Time for Outcome	504.48	< 0.001	0.208	0.023
19	Hematology	27.33	< 0.001	0.004	0.014
20	Neurology	278.64	< 0.001	0.013	0.024
21	Oncology	52.08	< 0.001	0.003	0.020
22	Down Syndrome	79.18	< 0.001	0.006	0.023
23	Nosocomial	70.17	<0.001	0.013	0.024

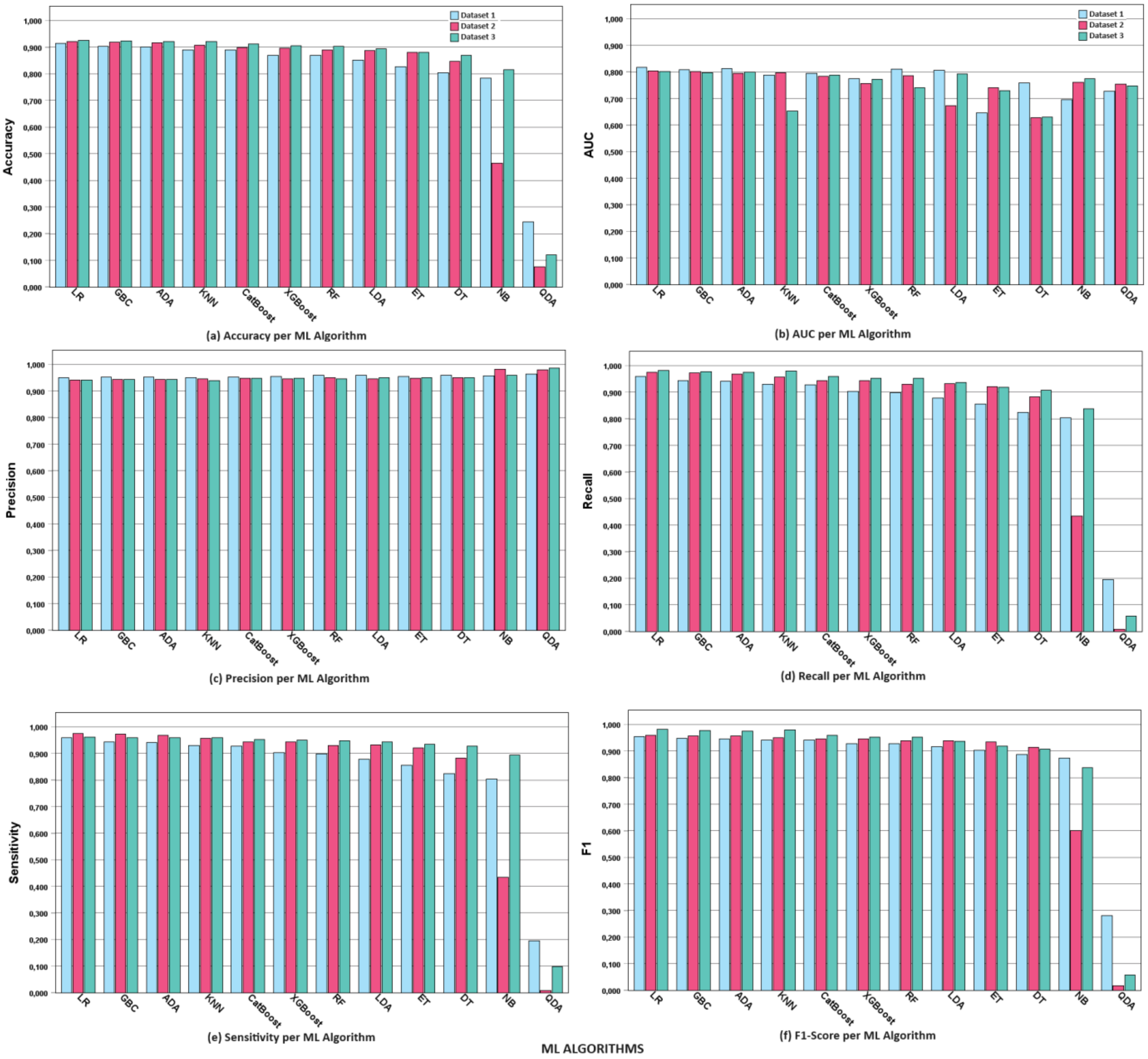
[a] Comorbidities dichotomic and comorbidities categoric had zero values for importance scores calculated with XGBoost. This is because the XGBoost algorithm detected multicollinearity between the two characteristics and total comorbidities. In this case, these two columns are ignored by the algorithm.

The findings in Table 2 suggest that the most important factors, as identified by the chi-square test, were age, cardiovascular disease, decreased oxygen saturation upon admission, total comorbidities, comorbidities as a binary feature, comorbidities as a categorical feature, and time to outcome. These factors demonstrated a higher level of statistical significance in distinguishing between the patients who experienced fatal outcomes and those who were discharged. This statistical significance is also apparent in the developed models and was of paramount importance in the training process.

In contrast, odynophagia, vaccine dose, abdominal pain, fever, vaccination, transplant, diabetes mellitus, vomiting, other syndromes, sex, diarrhea, and ageusia were identified as less relevant features in predicting COVID-19 mortality. Despite the clinical significance of these parameters in treatment efficacy and mortality prediction, a considerable number of them could be excluded from our machine learning analyses. Consequently, the execution of mortality prediction models could be achieved with a reduced set of factors while maintaining equivalent accuracy.

Assessment of the developed models

In this study, COVID-19 mortality prediction models were developed using 12 ML algorithms, namely, GBC, ADA, CatBoost, RF, XGBoost, ET, LR, LDA, DT, NB, KNN, and QDA. These models were trained on three feature datasets: dataset 1, containing all features; dataset 2, with features selected by pediatricians; and dataset 3, with features selected by the chi-squared independence test. The performance evaluation metrics used were accuracy, AUC, recall, precision, and sensitivity. The results are shown in Figure 1.



(a) Accuracy per ML Algorithm

(b) AUC per ML Algorithm

(c) Precision per ML Algorithm

(d) Recall per ML Algorithm

(e) Sensitivity per ML Algorithm

(f) F1-Score per ML Algorithm

Figure 1 – Metrics of ML algorithms per dataset. (a) Accuracy performances, (b) AUC, (c) Precision (d) Recall, (e) Sensitivity and (f) F1-score.

In general, the majority of the models demonstrated comparable levels of accuracy, displaying good to excellent performance across all three datasets. More specifically, numerically, the models performed best when trained on Dataset 3, which was selected using the chi-square method, followed by Datasets 2 and 1. However, Dataset 1 still exhibited commendable performance even when all features were included. For Dataset 3, the highest accuracies were achieved by LR (92.53%), GBC (92.34%), and ADA (92.19%). For Dataset 2, GB (92.08%), ADA (91.92%), and LR (91.73%) achieved the highest accuracy. For Dataset 1, GBC (91.41%), ADA (90.32%), and CatBoost (90.01%) were the best-performing models in terms of accuracy. Among the 12 algorithms analyzed, QDA consistently displayed the lowest performance across all datasets. Detailed comparison of the AUC for the top three models trained on Dataset 3, which achieved better results, is provided in Figure 2. Considering the reliability of the AUC metric for imbalanced datasets, particularly relevant in our study despite using SMOTE for balancing, is crucial. The AUC results are nearly identical across all three datasets, with a notable emphasis on dataset 1 containing all features.

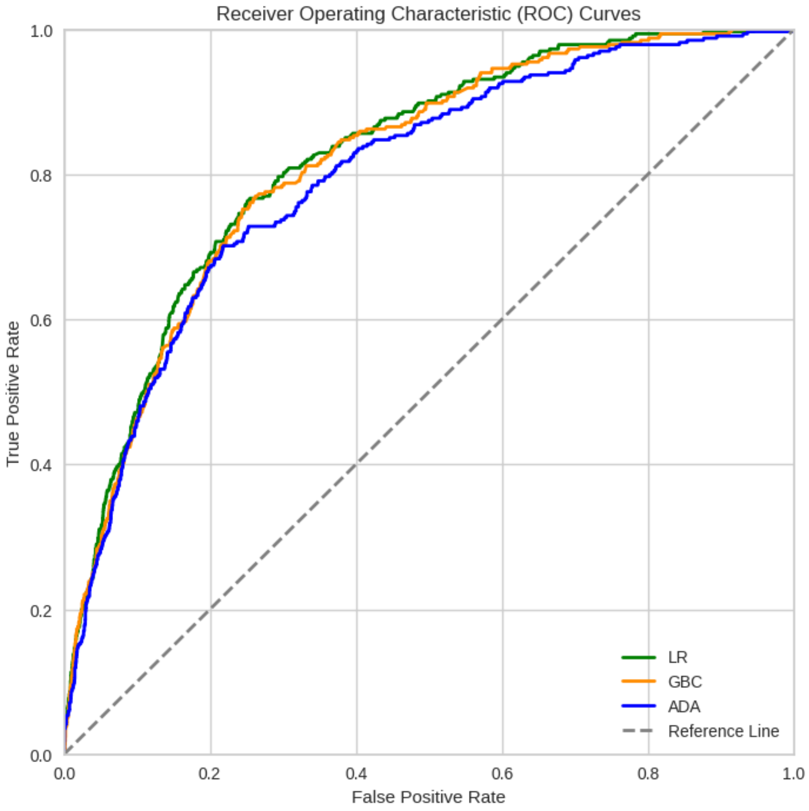


Figure 2 - ROC curves of the three best ML models for Dataset 3 that achieved better results.

Model Interpretation

We used SHAP summary and force plots to explain the decision-making process of the Gradient Boosting Classifier (GBC) model. GBC was chosen for its high ranking across the three datasets, and its decision-making process was analyzed to identify the most important features influencing its predictions. The force plot analysis is presented in Supplementary Figures 2 and 3.

In the summary plot for the SHAP values, the impact of each feature on the model's output is displayed as a dot on the horizontal axis. The position of the dot represents the SHAP value for that feature, indicating its contribution to prediction. The color of the dots corresponds to the value of the feature: red for higher values and blue for lower values, aiding in understanding the direction and magnitude of the impact on prediction. Figure 3 illustrates the contribution of feature values to the GBC decision. Features are plotted in the order of importance, with the most important characteristics at the top and the least important at the bottom.

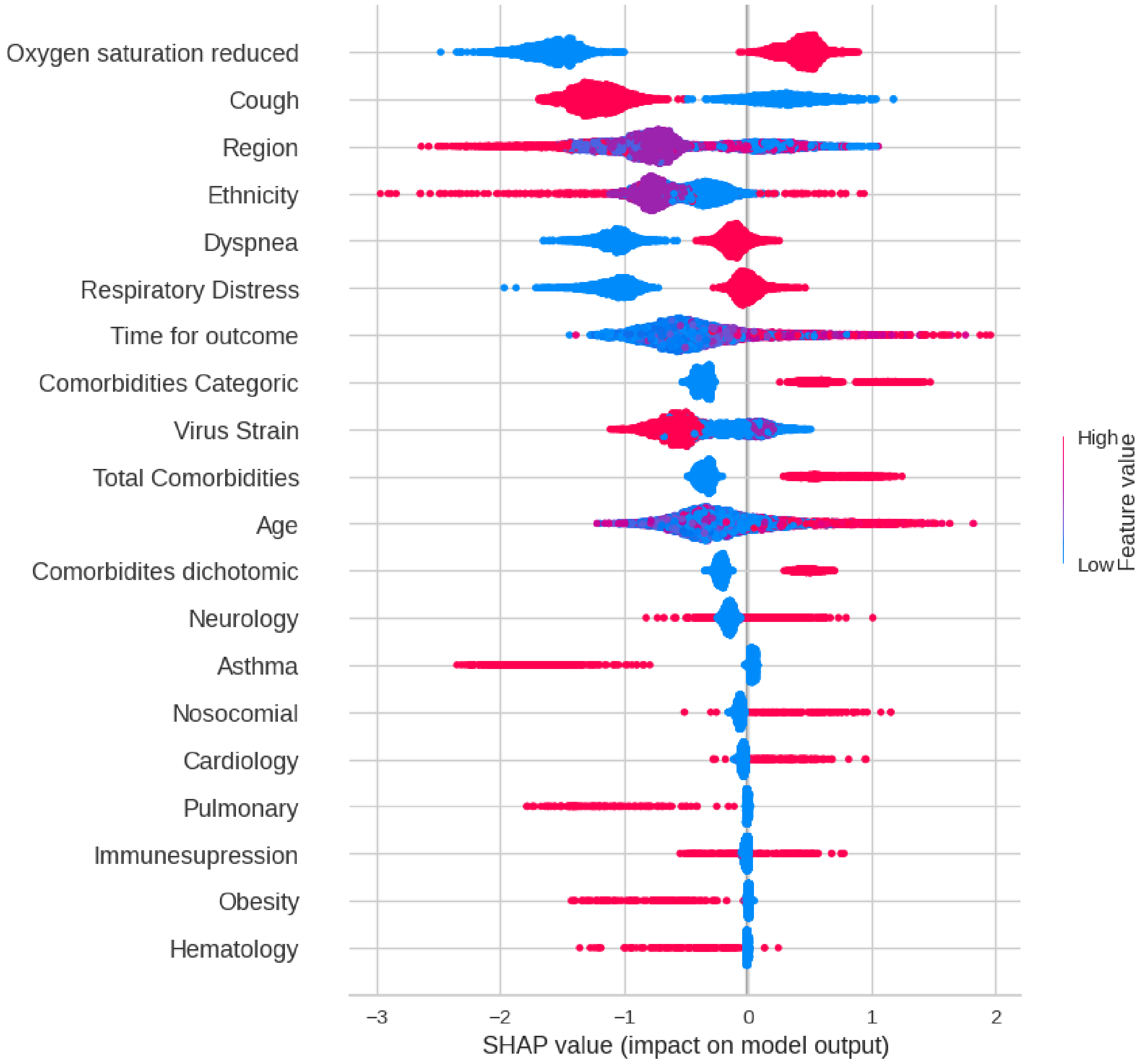


Figure 3 – A summary plot of SHAP values for mortality prediction on dataset 3 (features selected by Chi-squared test)

Figure 3 shows a summary plot of the SHAP values for each data point in the dataset. Each line corresponds to a point representing the impact of each feature. A greater separation of feature values indicates more effective variables for decision-making. In this context, the most crucial feature in the model's decision-making process is "oxygen saturation reduced." Blue points indicate low values of the feature, with "oxygen saturation reduced" taking on values of zero and one in our dataset. Zero indicates normal oxygen saturation and one indicates reduced saturation. Therefore, the graph illustrates that when "oxygen saturation reduced" is 0, it contributes to predicting a favorable outcome (patient discharge), whereas a value of 1 contributes to predicting a worse outcome (death). Similarly, the variable "comorbidities categoric" also exhibited a notable separation. Higher values in this variable, indicating the number of different comorbidities a patient may have ranged from zero to three or more, showed a significant influence. The graph reveals that for high values of this variable signifying patients with numerous comorbidities, the model tends to predict our target variable as 1 (death). Similar patterns were observed for dyspnea, respiratory distress, total comorbidities, and comorbidities.

Discussion

Key Points

In this study, we analyzed data from a large public dataset provided by the Brazilian government on patients hospitalized with COVID-19 in Brazil to develop and evaluate ML models predicting COVID-19 mortality risk in pediatric patients. Demographic information, risk factors, and clinical manifestations were evaluated to identify the key mortality predictors. We tested the ML models using three data subsets: (1) all dataset features, (2) features selected by pediatricians, and (3) statistically relevant features for predicting mortality. Our results show that ML models are robust and effective even without previous feature selection, which minimally improves model accuracy. However, we believe that feature selection is crucial for the model development. Dataset 3 (24 features) performed the best, followed by dataset 2 (17 features selected by medical experts). The findings from Dataset 1 may not apply to other data contexts. Models with fewer features are preferred if they achieve equal or better results. Therefore, for clinical use, models with fewer features, such as those trained on datasets 3 and 2, are preferable because they require less input from clinicians while providing accurate predictions.

Comparative analysis

Our findings are in agreement with other studies that have reported some important clinical predictors for COVID-19 patient mortality, the most relevant features included age[17, 18, 19, 20, 21], ethnicity[22], region [22], dyspnea[23], cough[17, 18, 21, 24, 25, 26], reduced oxygen saturation[26, 27, 28], cardiology disease[17, 19, 22, 26, 29, 30], pulmonary disease[17, 27], immunosuppression[22, 31, 32], renal[22, 33], asthma[22, 34, 35, 36], total comorbidities[17, 27, 34], hematology disease[37, 38, 39], neurology disease[17, 20, 21, 22], oncology disease[17, 20, 26, 40, 41], hypertension[17, 19, 20, 25], down syndrome[42], and total comorbidities [43].

In our analysis, 12 ML algorithms were tested to develop prediction models for hospitalized pediatric patients with COVID-19. The LR model performed the best, with 92.5% accuracy, 98.11% sensitivity, 94.13% precision, 96.07% F1-score, and 80.15% AUC. GBC and ADA models also showed good performance, with AUCs ≥ 79.6%. Other ML algorithms had acceptable performances, with AUCs ranging from 80.1 to 81.6%. The DT model had the weakest performance (AUC = 62.9%), and QDA had the lowest accuracy (7.9% to 24.3%). In addition, the importance and efficiency of multiple features in predicting COVID-19 mortality using XGBoost, random forest, and chi-squared tests were investigated. The results indicated that reduced oxygen saturation at admission, comorbidities, and older age were the most relevant predictors of mortality risk, as shown in the SHAP plots. These features are strong predictors of mortality risk in hospitalized pediatric COVID-19 patients. Integrating these with 23 other statistically relevant features improved the prognostic performance of the ML algorithms for mortality prediction in this group.

Models are often presented in the literature as black-box systems, lacking transparency regarding the contribution of each characteristic to their predictions. Machine learning models make decisions based on individual feature values, and it is crucial to understand these decisions, particularly in medical applications in which patient well-being is at stake. The concept of Explainable Artificial Intelligence (XAI) [44] enhances the interpretability and trustworthiness of these models. One XAI technique is the SHapley Additive extension (SHAP) values, which explain model outputs by attributing each feature's contribution to the prediction. Rooted in cooperative game theory, the SHAP values provide a unified measure of feature importance, considering all feature combinations. SHAP is a post-hoc interpretation technique that can be applied to any machine learning model. In this regard, our analysis revealed that reduced oxygen saturation, comorbidities (presented as numerical, binary, or ordinal features), dyspnea, and respiratory distress at admission were reliable predictors of mortality in pediatric patients with COVID-19.

Few studies have evaluated ML models for predicting deaths of children and adolescents with COVID-19. In this regard, we recently conducted a systematic review to analyze and summarize the key characteristics related to the study design, modeling techniques, and performance measures reported in studies focusing on clinical prediction models developed using supervised machine learning algorithms in pediatric patients with COVID-19 [4]. We found 10 studies (six predictive diagnostic models, and four were prognostic models). All models were developed to predict binary outcomes. The most frequently predicted outcome was disease detection. The most commonly used machine learning models in these studies were tree-based and neural networks. However, our systematic review revealed that most studies failed to address relevant issues, including small sample sizes, inconsistent reporting practices on data preparation, biases in data sources, lack of reporting metrics such as calibration and discrimination, hyperparameters, and other aspects that allow reproducibility by other researchers and might improve the methodology. Other studies have evaluated ML models for predicting various outcomes in the pediatric setting, but in contexts other than COVID-19. Detailed information regarding each of these studies is provided in Supplementary Material.

Public policies in data management, Information Systems and audit for government data

As shown in the methodology section the SIVEP-Kids dataset had a total of 56,330 records and after data pre-processing, 24,097 records were kept in the database. This loss of data in pre-processing was due to errors and inconsistencies in the data that arise from the process of generating these datasets. Currently, the Brazilian government does not have an entity to audit health data made available by the Ministry of Health. Furthermore, there is no concern in the development of the systems that feed this data. Many systems have important fields that are not mandatory. In a hospital, during periods of high demand, the tendency is for healthcare professionals who are working to fill out data in these systems to only provide the data that is mandatory, or those that they consider most important. In this way, a lot of data is lost, or important information is not reported in these systems.

It is important that the Brazilian government creates public policies for data management and auditing, as well as modernizing monitoring systems, constantly updating them and investing in staff training so that they can correctly fill in data in government systems.

Strengths and limitations of this study

The strength of this study lies in the use of a nationwide database to provide comprehensive data on COVID-19 in Brazilian pediatric patients. With a large sample size of lab-confirmed cases, this study details the clinical features, risk factors, and outcomes of hospitalized children. Another important finding of our study is that ML algorithms are robust for large databases, which may provide valuable insights for public health policies. Additionally, ML models may assist in clinical decision-making and evidence-based patient management, enhancing outcomes and medical care quality.

However, its limitations include a lack of generalizability to other regions, inclusion of only hospitalized (likely severe) cases, absence of hospital record data, missing data issues, and lack of a national audit system for data consistency.

Conclusions

In this study, we compared various machine learning (ML) algorithms to predict the mortality of hospitalized children and adolescents with COVID-19. The LR, GBC, and ADA models were particularly effective in accurately predicting mortality in hospitalized pediatric COVID-19 patients, potentially optimizing hospital resources, and improving patient survival chances.

Our findings revealed that characteristics such as reduced oxygen saturation levels at the time of admission and the presence of comorbidities are crucial factors for decision-making in ML models. By employing a Logistic Regression (LR) predictive model that incorporated a set of predictors, we were able to effectively identify high-risk patients upon admission, thereby improving the likelihood of patient survival. Further studies are required to explore different feature sets for classifier training and validation. For instance, this study focused on predicting short-term adverse outcomes, such as mortality or discharge, rather than long-term effects or protective public health measurements, such as the vaccination program.

Acknowledgments

The authors would like to thank State Research Foundation- FAPEMIG, Minas Gerais, Brazil, the National Council for Scientific and Technological Development - CNPq, Brazil, and the Coordination for the Improvement of Higher Education Personnel, CAPES, Brazil.

Data availability

The data used in this study is public and can be found on this website: https://opendatasus.saude.gov.br/dataset/. The data is available for download in .csv formats. All source code used in the project will be made available by the authors upon reasonable request.

References

1 - Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K. et al. Large language models in medicine. Nat Med 29, 1930–1940 (2023). https://doi.org/10.1038/s41591-023-02448-8

2 - Yu, Ping, Hua Xu, Xia Hu, and Chao Deng. 2023. "Leveraging Generative AI and Large Language Models: A Comprehensive Roadmap for Healthcare Integration" *Healthcare* 11, no. 20: 2776. https://doi.org/10.3390/healthcare11202776

3 - Bohr A, Memarzadeh K. The rise of artificial intelligence in healthcare applications. Artificial Intelligence in Healthcare. 2020:25–60. doi: 10.1016/B978-0-12-818438-7.00002-2. Epub 2020 Jun 26. PMCID: PMC7325854.

4 - dos Santos, A. L., Pinhati, C., Perdigão, J., Galante, S., Silva, L., Veloso, I., Simões e Silva, A. C., & Oliveira, E. A. (2024, February). Machine learning algorithms to predict outcomes in children and adolescents with COVID-19: A systematic review. Artificial Intelligence in Medicine, 102824. https://doi.org/10.1016/j.artmed.2024.102824

5 - S. Bengio, J. Weston, and D. Grangier, ‘‘Label embedding trees for large multi-class tasks,’’ in Proc. Adv. Neural Inf. Process. Syst., 2010, vol. 23, no. 1, pp. 163–171.

6 - Del Moral, P., Nowaczyk, S., Pashami, S. (2022) Why Is Multiclass Classification Hard? IEEE Access, 10: 80448-80462 https://doi.org/10.1109/access.2022.3192514

7 - Oliveira EA, Colosimo EA, Simões E Silva AC, Mak RH, Martelli DB, Silva LR, Martelli-Júnior H, Oliveira MCL. Clinical characteristics and risk factors for death among hospitalised children and adolescents with COVID-19 in Brazil: an analysis of a nationwide database. Lancet Child Adolesc Health. 2021 Aug;5(8):559-568. doi: 10.1016/S2352-4642(21)00134-6. Epub 2021 Jun 11. PMID: 34119027; PMCID: PMC8192298.

8 - Oliveira EA, Oliveira MCL, Silva ACSE, Colosimo EA, Mak RH, Vasconcelos MA, Silva LR, Martelli DB, Pinhati CC, Martelli-Júnior H. Effectiveness of BNT162b2 and CoronaVac vaccines against omicron in children aged 5 to 11 years. World J Pediatr. 2023 Oct;19(10):949-960. doi: 10.1007/s12519-023-00699-6. Epub 2023 Mar 13. PMID: 36914907; PMCID: PMC10010648.

9 - García, S., Luengo, J. & Herrera, F. Data Preprocessing in Data Mining Vol. 72 (Springer, 2015)

10 - Xiayu Xiang, Shaoming Duan, Hezhong Pan, Peiyi Han, Jiahao Cao, and Chuanyi Liu. 2021. From One-hot Encoding to Privacy-preserving Synthetic Electronic Health Records Embedding. In Proceedings of the 2020 International Conference on Cyberspace Innovation of Advanced Technologies (CIAT 2020).

Association for Computing Machinery, New York, NY, USA, 407–413. https://doi.org/10.1145/3444370.3444605

11 - Dorn, M. et al. Comparison of machine learning techniques to handle imbalanced COVID-19 CBC datasets. PeerJ Comput. Sci. 7, 1–34 (2021)

12 - Erol, G., Uzbaş, B., Yücelbaş, C. & Yücelbaş, Ş. Analyzing the effect of data preprocessing techniques using machine learning algorithms on the diagnosis of COVID-19. Concurr. Comput. 34(28), 1–16 (2022).

13 - Wang K, Tian J, Zheng C, Yang H, Ren J, Li C, Han Q, Zhang Y. Improving Risk Identification of Adverse Outcomes in Chronic Heart Failure Using SMOTE+ENN and Machine Learning. Risk Manag Healthc Policy. 2021;14:2453-2463 https://doi.org/10.2147/RMHP.S310295

14 - Wongvorachan, T.; He, S.; Bulut, O. A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. *Information* 2023, *14*, 54. https://doi.org/10.3390/info14010054

15 - Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

16 – Moez Ali (2020). PyCaret: An open-source, low-code machine learning library in Python. PyCaret version 3.1.0, https://www.pycaret.org.

17 - Zakariaee, S. S., Abdi, A. I., Naderi, N. & Babashahi, M. Prognostic significance of chest CT severity score in mortality prediction of COVID-19 patients, a machine learning study. Egypt J. Radiol. Nucl. Med. 54(73), 1–9 (2023).

18 - Wu, G. et al. Development of a clinical decision support system for severity risk prediction and triage of COVID-19 patients at hospital admission: An international multicentre study. Eur. Respir. J. 56(2), 1–11 (2020).

19 - Yadaw, A. S. et al. Clinical features of COVID-19 mortality: Development and validation of a clinical prediction model. Lancet Digit. Health 2(10), 516–525 (2020).

20 - Moulaei, K., Ghasemian, F., Bahaadinbeigy, K., Sarbi, R. E. & Taghiabad, Z. M. Predicting mortality of COVID-19 patients based on data mining techniques. J. Biomed. Phys. Eng. 11(5), 653–662 (2021).

21 - Moulaei, K., Shanbehzadeh, M., Mohammadi-Taghiabad, Z. & Kazemi-Arpanahi, H. Comparing machine learning algorithms for predicting COVID-19 mortality. BMC Med. Inform. Decis. Mak. 22(1), 1–12 (2022).

22 - Baqui, P., Marra, V., Alaa, A.M. et al. Comparing COVID-19 risk factors in Brazil using machine learning: the importance of socioeconomic, demographic and structural factors. Sci Rep 11, 15591 (2021). https://doi.org/10.1038/s41598-021-95004-8

23 - Shi L, Wang Y, Wang Y, Duan G, Yang H. Dyspnea rather than fever is a risk factor for predicting mortality in patients with COVID-19. J Infect. 2020 Oct;81(4):647-679. doi: 10.1016/j.jinf.2020.05.013. Epub 2020 May 15. PMID: 32417316; PMCID: PMC7228739.

24 - Gao, Y. et al. Machine learning-based early warning system enables accurate mortality risk prediction for COVID-19. Nat. Commun. 11(1), 1–10 (2020).

25 - Das, A. K., Mishra, S. & Gopalan, S. S. Predicting CoVID-19 community mortality risk using machine learning and development of an online prognostic tool. PeerJ 8, 1–17 (2020).

26 - Assaf, D., Gutman, Y., Neuman, Y. et al. Utilization of machine-learning models to accurately predict the risk for critical COVID-19. Intern Emerg Med 15, 1435–1443 (2020). https://doi.org/10.1007/s11739-020-02475-0

27 - Banoei, M.M., Dinparastisaleh, R., Zadeh, A.V. et al. Machine-learning-based COVID-19 mortality prediction model and identification of patients at low and high risk of dying. Crit Care 25, 328 (2021). https://doi.org/10.1186/s13054-021-03749-5

28 - Kar, S., Chawla, R., Haranath, S.P. et al. Multivariable mortality risk prediction using machine learning for COVID-19 patients at admission (AICOVID). Sci Rep 11, 12801 (2021). https://doi.org/10.1038/s41598-021-92146-7

29 - Das, A. K., Mishra, S. & Gopalan, S. S. Predicting CoVID-19 community mortality risk using machine learning and development of an online prognostic tool. PeerJ 8, 1–17 (2020).

30 - Allenbach, Y. et al. Development of a multivariate prediction model of intensive care unit transfer or death: A French prospective cohort study of hospitalized COVID-19 patients. PLoS ONE 15(10), 1–12 (2020).

31 - Gao, Y., Chen, L., Chi, J., et al. Development and validation of an online model to predict critical COVID-19 with immune-inflammatory parameters. j intensive care 9, 19 (2021). https://doi.org/10.1186/s40560-021-00531-1

32 - Xu, W., Sun, NN., Gao, HN. et al. Risk factors analysis of COVID-19 patients with ARDS and prediction based on machine learning. Sci Rep 11, 2933 (2021). https://doi.org/10.1038/s41598-021-82492-x

33 - Jianhong Kang, Ting Chen, Honghe Luo, Yifeng Luo, Guipeng Du, Mia Jiming-Yang, Machine learning predictive model for severe COVID-19, Infection, Genetics and Evolution, Volume 90, 2021, 104737, ISSN 1567-1348, https://doi.org/10.1016/j.meegid.2021.104737.

34 - Aktar, S.; Talukder, A.; Ahamad, M.M.; Kamal, A.H.M.; Khan, J.R.; Protikuzzaman, M.; Hossain, N.; Azad, A.K.M.; Quinn, J.M.W.; Summers, M.A.; et al. Machine Learning Approaches to Identify Patient Comorbidities and Symptoms That Increased Risk of Mortality in COVID-19. *Diagnostics* 2021, *11*, 1383. https://doi.org/10.3390/diagnostics11081383

35 - An, C., Lim, H., Kim, DW. et al. Machine learning prediction for mortality of patients diagnosed with COVID-19: a nationwide Korean cohort study. Sci Rep 10, 18716 (2020). https://doi.org/10.1038/s41598-020-75767-2

36 - Chimbunde E, Sigwadhi LN, Tamuzi JL, Okango EL, Daramola O, Ngah VD, Nyasulu PS. Machine learning algorithms for predicting determinants of COVID-19 mortality in South Africa. Front Artif Intell. 2023 Oct 10;6:1171256. doi: 10.3389/frai.2023.1171256. PMID: 37899965; PMCID: PMC10600470.

37 - Wallace Duarte de Holanda, Lenardo Chaves e Silva, Álvaro Alvares de Carvalho César Sobrinho,
Machine learning models for predicting hospitalization and mortality risks of COVID-19 patients,
Expert Systems with Applications, Volume 240, 2024, 122670, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2023.122670.

38 - Huyut, M.T.; Velichko, A.; Belyaev, M. Detection of Risk Predictors of COVID-19 Mortality with Classifier Machine Learning Models Operated with Routine Laboratory Biomarkers. *Appl. Sci.* 2022, *12*, 12180. https://doi.org/10.3390/app122312180

39 - Kamel FO, Magadmi R, Qutub S, Badawi M, Badawi M, Madani TA, Alhothali A, Abozinadah EA, Bakhshwin DM, Jamal MH, Burzangi AS, Bazuhair M, Alqutub H, Alqutub A, Felemban SM, Al-Sayes F, Adam S. Machine Learning-Based Prediction of COVID-19 Prognosis Using Clinical and Hematologic Data. Cureus. 2023 Dec 9;15(12):e50212. doi: 10.7759/cureus.50212. PMID: 38089943; PMCID: PMC10710934.

40 - Hu, H., Yao, N. & Qiu, Y. Comparing rapid scoring systems in mortality prediction of critically ill patients with novel coronavirus disease. Acad. Emerg. Med. 27(6), 461–468 (2020).

41 - Chin, V. et al. A case study in model failure? COVID-19 daily deaths and ICU bed utilization predictions in New York state. Eur. J. Epidemiol. 35(8), 733–742 (2020).

42 - Landes SD, Turk MA, Damiani MR, Proctor P, Baier S. Risk Factors Associated With COVID-19 Outcomes Among People With Intellectual and Developmental Disabilities Receiving Residential Services. *JAMA Netw Open.* 2021;4(6):e2112862. doi:10.1001/jamanetworkopen.2021.12862

43 - Alballa N, Al-Turaiki I. Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review. Inform Med Unlocked. 2021;24:100564. doi: 10.1016/j.imu.2021.100564. Epub 2021 Apr 3. PMID: 33842685; PMCID: PMC8018906.

44 - Lundberg, S.M., Erion, G., Chen, H. et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2, 56–67 (2020). https://doi.org/10.1038/s42256-019-0138-9

6. RESULTADOS ADICIONAIS PRELIMINARES

Esta tese também utilizou modelos de aprendizado de máquina para predição de outros desfechos, que foram: gravidade da COVID-19, necessidade de UTI (Unidade de Terapia Intensiva) e necessidade de suporte ventilatório. Esses desfechos estavam presentes na base de dados do SIVEP -Gripe e, portanto, foram analisados pelos algoritmos de aprendizado de máquina que também avaliaram o desfecho óbito. Para todos os desfechos existentes na base de dados do SIVEP-Gripe, o melhor método de seleção de características da base de dados foi o método Chi-quadrado. Nas seções a seguir apresentaremos os resultados dos modelos para esses desfechos. Todos os procedimentos metodológicos adotados para o desfecho óbito e que foram apresentados na seção 5 (artigo original) também foram adotados para os desfechos apresentados nessa seção.

6.1. Desfecho gravidade da COVID-19

O desfecho gravidade diz respeito à evolução da COVID-19 nos pacientes. A evolução da doença está relacionada com os outros desfechos, mas é importante analisarmos esse desfecho isoladamente para saber o comportamento dos modelos em relação as características e a predição de gravidade para um determinado paciente. A Tabela 1 mostra o resultado dos modelos de aprendizado de máquina para o desfecho gravidade.

Tabela 1 – Desempenho dos algoritmos de aprendizado de máquina no conjunto de dados 3 (características selecionadas pelo teste de independência de Chi-quadrado) para predição de gravidade da COVID-19.

Algoritmo	Acurácia	AUC	Recall	Precisão	F1	Sensitividade
NB	0.7040	0.6811	0.3311	0.4888	0.3946	0.3311
XGBOOST	0.6959	0.7328	0.6232	0.4836	0.5444	0.6232
RF	0.6875	0.7241	0.6308	0.4733	0.5406	0.6308
GBC	0.6854	0.7429	0.6643	0.4720	0.5517	0.6643
ET	0.6839	0.7135	0.6187	0.4686	0.5330	0.6187
ADA	0.6684	0.7360	0.6839	0.4544	0.5459	0.6839
DT	0.6469	0.6363	0.6109	0.4262	0.5019	0.6109
QDA	0.6216	0.6455	0.4657	0.4498	0.4067	0.4657
LDA	0.6084	0.7112	0.7499	0.4070	0.5276	0.7499
KNN	0.6063	0.6692	0.6966	0.3995	0.5077	0.6966
LR	0.6053	0.7111	0.7522	0.4049	0.5263	0.7522

Como podemos ver na Tabela 1, os três algoritmos que tiveram melhor desempenho na predição de gravidade da COVID-19 em pacientes pediátricos hospitalizados foram: Gradient Boosting Classifier (GBC), Adaboost (ADA) e Extreme Gradient Boosting (XGBOOST). Lembrando que para conjuntos de dados que precisaram de ser balanceados com SMOTE como foi o caso do nosso conjunto de dados, a AUC é a nossa métrica principal. A AUC é uma métrica mais confiável para dados desbalanceados e que passaram por processos de imputação.

Os modelos apresentaram uma baixa sensibilidade na predição da gravidade da doença. Isso significa que, para o desfecho gravidade, os modelos tenderam a classificar um número maior de falsos negativos, dessa forma identificando um paciente que teve maior gravidade da doença como sendo um paciente de baixo risco para este evento. A acurácia dos modelos para o conjunto de dados foi mediana, apresentando em média 65% de acertos em relação aos dados do conjunto.

A Figura 3 mostra o gráfico de resumo de contribuições de características para a decisão do melhor modelo na predição de gravidade por COVID-19, o modelo GBC. As características com maior discriminação para que o modelo classifique um paciente com gravidade acentuada, moderada ou leve por COVID-19 foram: saturação de oxigênio reduzida, total de comorbidades, tosse, desconforto respiratório, problemas cardiológicos e diabetes. Em relação à saturação de oxigênio reduzida, a presença deste quadro clínico no paciente faz com que o modelo tenha tendência para classificar o paciente como quadro de gravidade acentuada ou moderada de COVID-19 e a não presença faz com que o modelo classifique o paciente como um quadro que não vai apresentar gravidade acentuada ou moderada. Essa é a característica mais importante do modelo. A mesma lógica vale para o total de comorbidades e para pacientes que tem problemas cardíacos e diabetes.

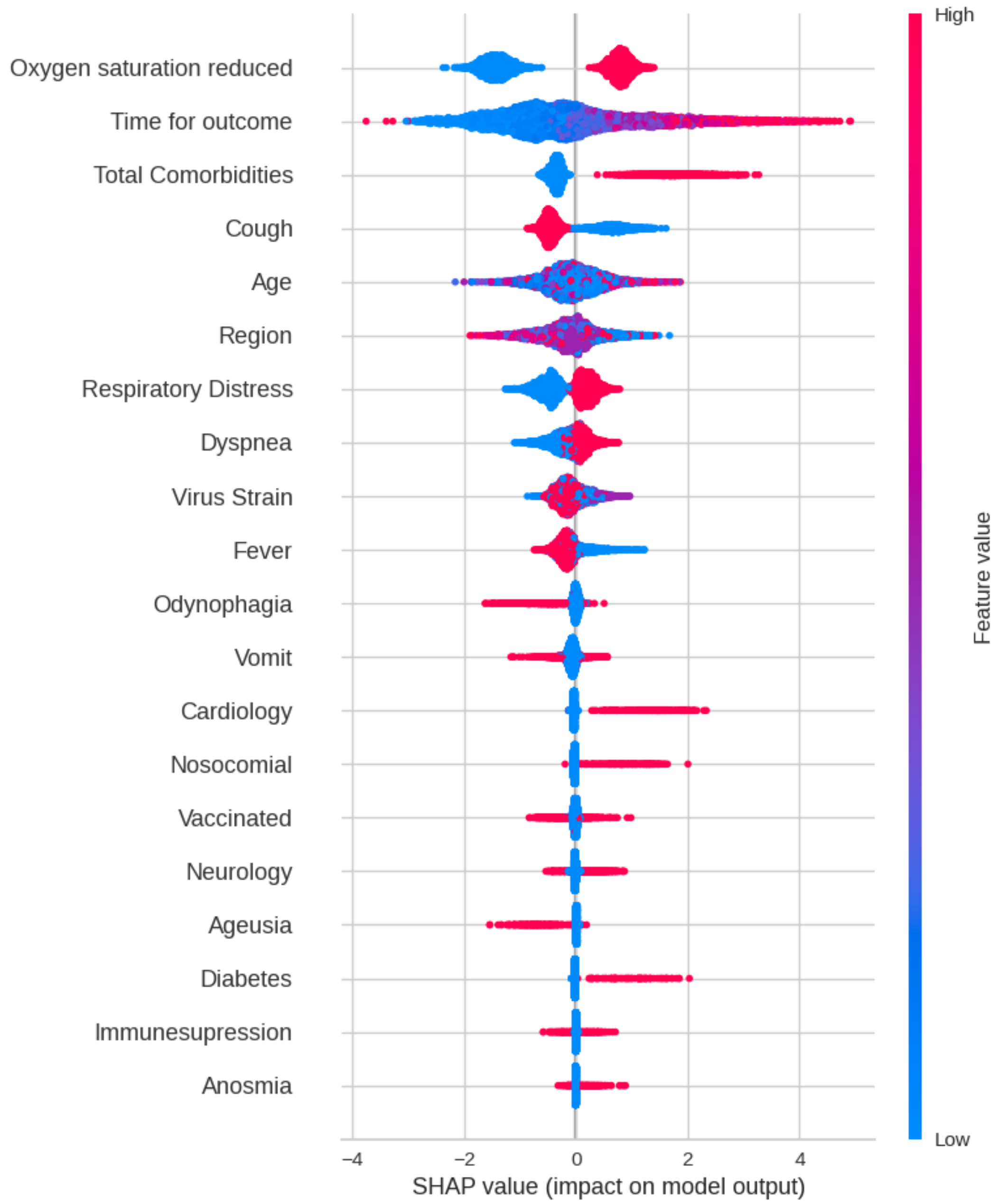


Figura 3. Gráfico de resumo de contribuições para tomada de decisão do algoritmo GBC para predição do desfecho gravidade.

6.2. Desfecho Necessidade de Suporte Ventilatório

O desfecho suporte ventilatório diz respeito a necessidade ou não de ventilação mecânica para um paciente. A Tabela 2 mostra o resultado dos modelos de aprendizado de máquina para esse desfecho presente no conjunto de dados do SIVEP-Kids.

Tabela 2 – Desempenho dos algoritmos de aprendizado de máquina no conjunto de dados 3 (características selecionadas pelo teste de independência de Chi-quadrado) para predição de necessidade de suporte ventilatório.

Algoritmo	Acurácia	AUC	Recall	Precisão	F1	Sensitividade
NB	0.5767	0.7029	0.5767	0.6217	0.5922	0.5910
ET	0.5618	0.7374	0.5618	0.6370	0.5842	0.4622
XGBOOST	0.5592	0.7461	0.5592	0.6448	0.5840	0.4608
RF	0.5565	0.7445	0.5565	0.6441	0.5807	0.4402
GBC	0.5438	0.7613	0.5438	0.6522	0.5703	0.3951
ADA	0.5255	0.7597	0.5255	0.6579	0.5521	0.3331
DT	0.5236	0.6538	0.5236	0.6151	0.5512	0.4418
LDA	0.4681	0.7546	0.4681	0.6695	0.4468	0.0779
LR	0.4414	0.7447	0.4414	0.6611	0.4555	0.1370
KNN	0.4274	0.6579	0.4274	0.6113	0.4652	0.2821
QDA	0.3671	0.6091	0.3671	0.3315	0.2980	0.2225

Os modelos não apresentaram bons resultados para o desfecho necessidade de suporte ventilatório, sendo os três melhores: GBC, ADA e LDA. Entretanto, as acurácias para a base de dados do SIVEP-Gripe não foram satisfatórias, ou seja, os modelos erram mais do que acertam para esse tipo de desfecho. A Figura 4 mostra o gráfico de resumo de contribuições de características para tomada de decisão pelo modelo GBC. As características mais importantes para que o modelo classifique que o paciente pediátrico terá necessidade de suporte ventilatório são: saturação de oxigênio reduzida, total de comorbidades, tosse, desconforto respiratório, infecção nosocomial, diabetes e ageusia.

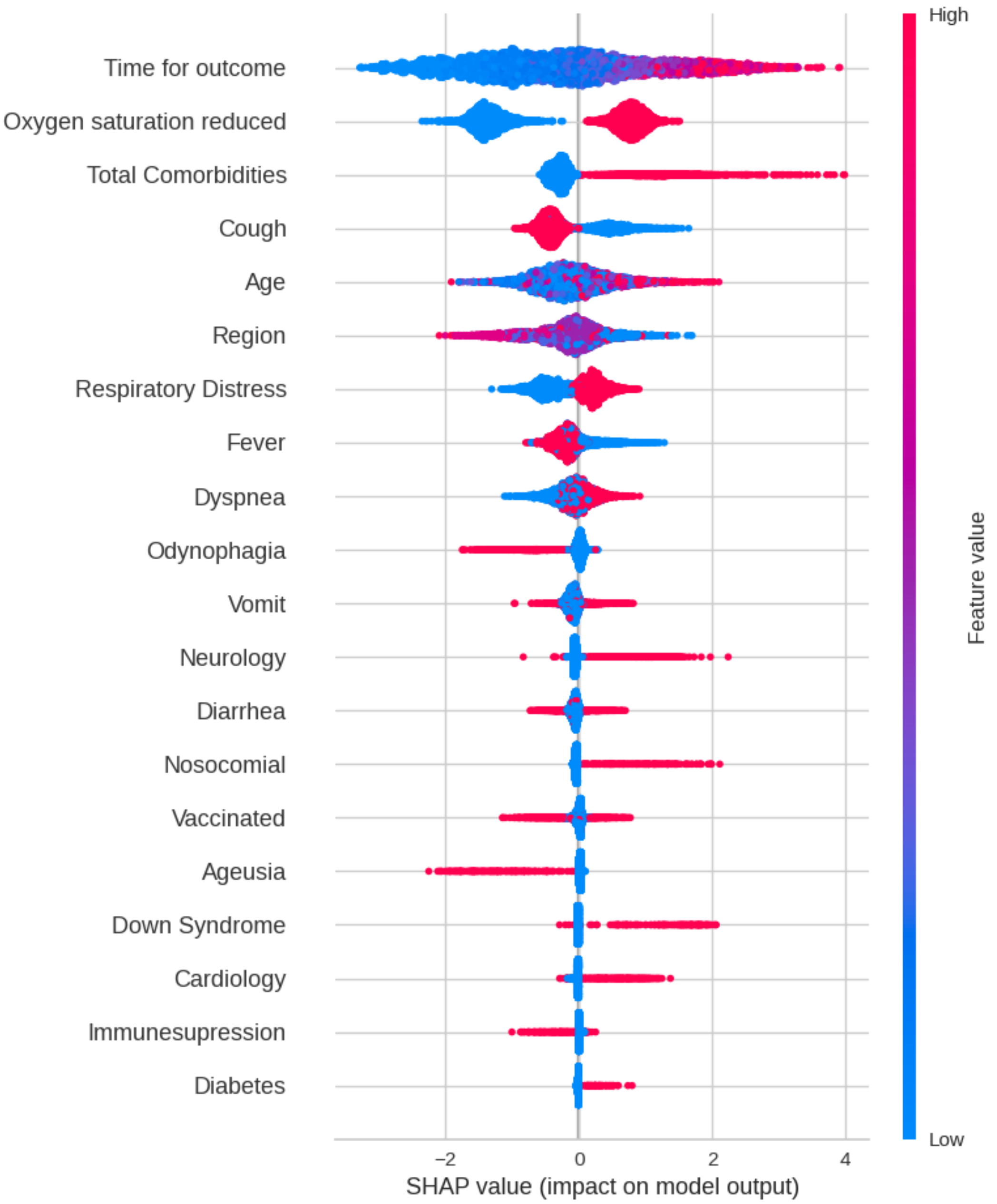


Figura 4. Gráfico de resumo de contribuições para tomada de decisão do algoritmo GBC para predição do desfecho suporte ventilatório.

6.3. Desfecho admissão em Unidade de Terapia Intensiva (UTI)

Nesse desfecho, os modelos tentam prever se o paciente pediátrico será internado na Unidade de Terapia Intensiva - UTI a partir dos dados presentes no conjunto. O desfecho UTI se aproxima muito do desfecho suporte ventilatório em termos práticos, pois pacientes com COVID-19 que foram para UTI fizeram a utilização de suporte ventilatório na maioria das vezes. A Tabela 3 mostra o desempenho dos modelos para esse desfecho.

Tabela 3 – Desempenho dos algoritmos de aprendizado de máquina no conjunto de dados 3 (características selecionadas pelo teste de independência de Chi-quadrado) para predição de necessidade de UTI.

Algoritmo	Acurácia	AUC	Recall	Precisão	F1	Sensitividade
NB	0.7155	0.6727	0.3215	0.4261	0.3663	0.3215
QDA	0.7151	0.6699	0.3229	0.4254	0.3670	0.3229
XGBOOST	0.6991	0.7215	0.5795	0.4339	0.4961	0.5795
GBC	0.6940	0.7334	0.6403	0.4336	0.5169	0.6403
RF	0.6875	0.7096	0.6006	0.4221	0.4956	0.6006
ET	0.6823	0.6916	0.5887	0.4147	0.4865	0.5887
ADA	0.6688	0.7263	0.6705	0.4098	0.5086	0.6705
DT	0.6537	0.6330	0.5905	0.3846	0.4658	0.5905
LDA	0.5968	0.7011	0.7437	0.3604	0.4855	0.7437
KNN	0.5954	0.6607	0.6937	0.3523	0.4672	0.6937
LR	0.5936	0.6997	0.7437	0.3582	0.4835	0.7437

De forma similar aos outros dois desfechos, nos resultados para necessidade de UTI o algoritmo GBC apresentou melhor performance, considerando AUC como métrica principal, seguido pelos algoritmos ADA e XGBOOST, respectivamente. Por outro lado, os resultados de acurácia foram um pouco melhores que os outros dois desfechos com valores de AUC um pouco mais elevados. Em relação à sensibilidade, os algoritmos continuam produzindo falsos negativos e, nesse caso, deixando de prever casos que foram positivos. Em relação a Figura 5, o gráfico de resumo de contribuições, mostra as principais características clínicas utilizadas pelo modelo GBC para realizar a tomada de decisão entre classificar um paciente que vai para a UTI e um paciente que não vai. A principal característica para que o modelo classifique um paciente para ir para UTI é o tempo para o desfecho, ou seja o tempo que o paciente está no hospital, isso quer dizer que quanto maior o tempo do paciente no hospital maiores as chances de ir para a UTI. Também foram considerados saturação reduzida de oxigênio, desconforto respiratório e total de comorbidades do paciente.

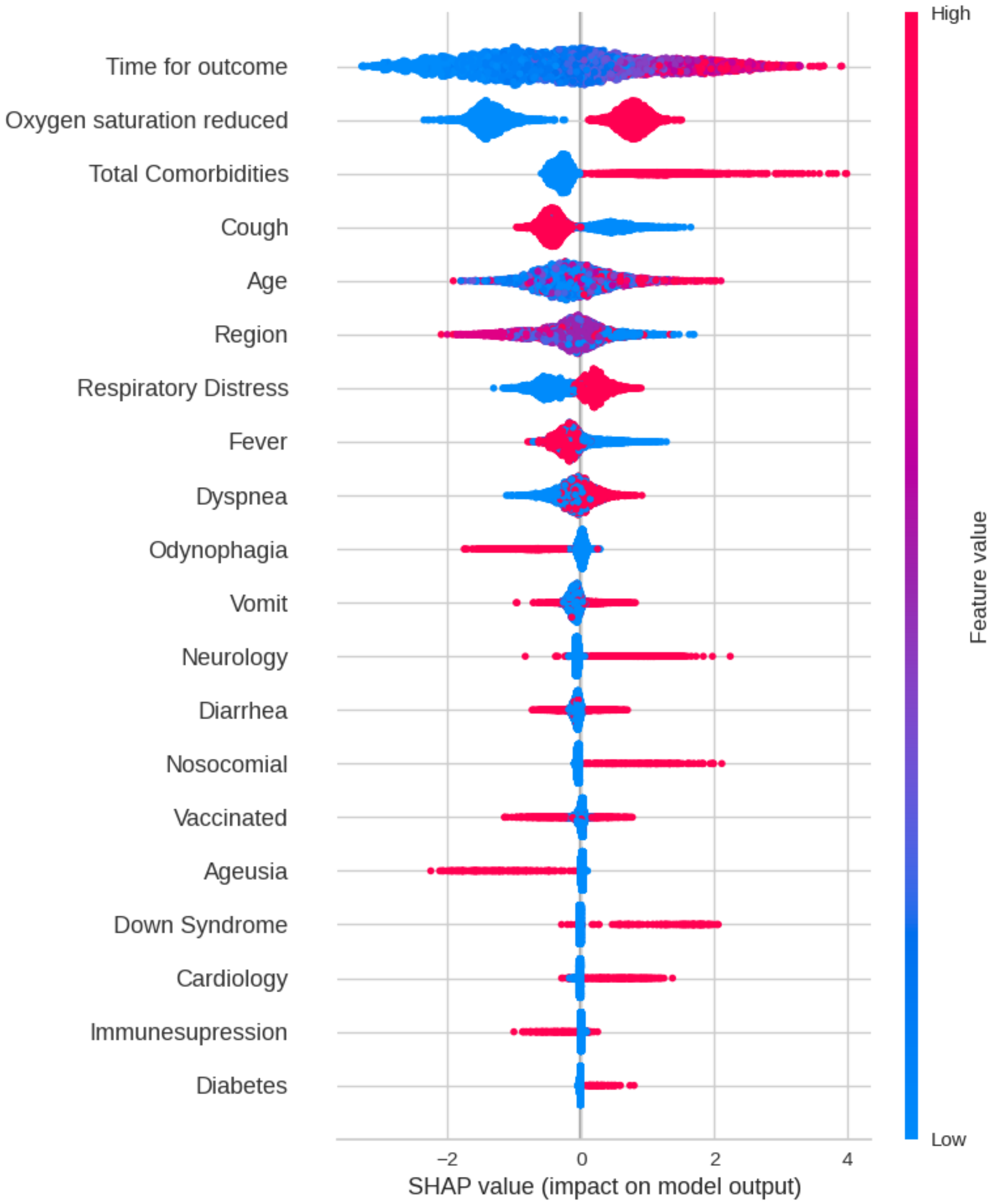
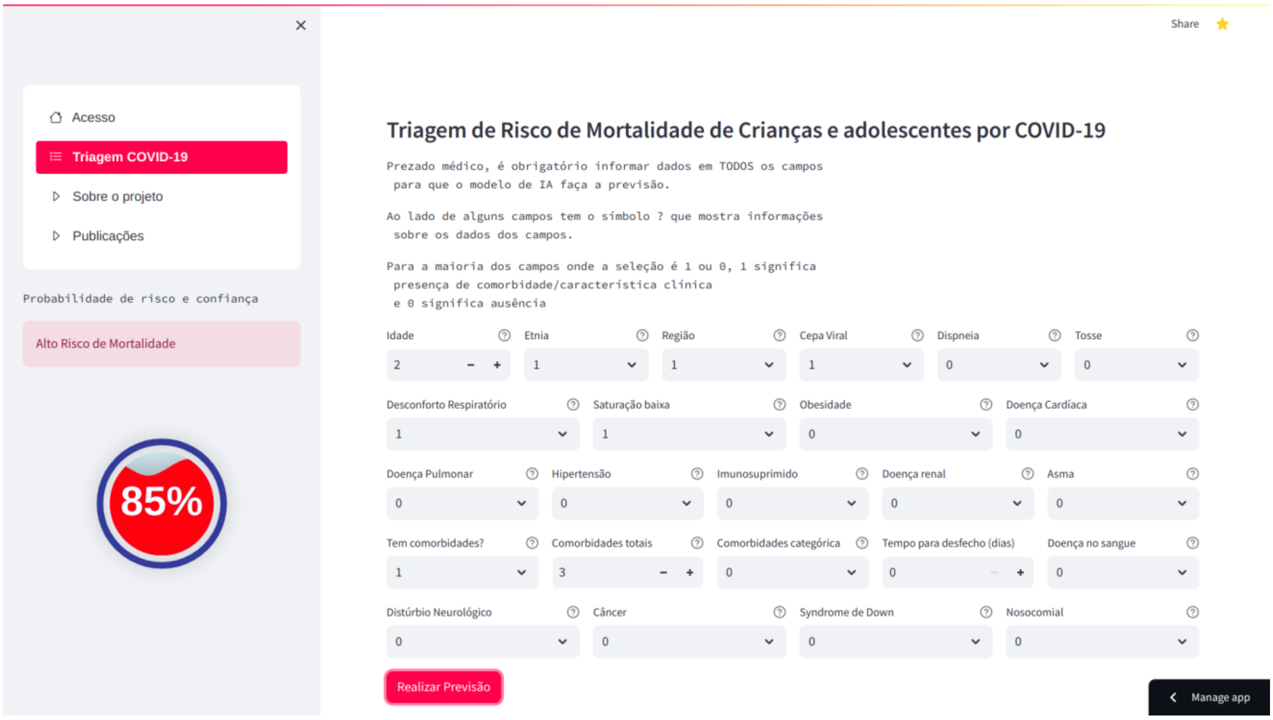


Figura 5. Gráfico de resumo de contribuições para tomada de decisão do algoritmo GBC para predição do desfecho UTI.

6.4. Sistema de suporte a decisão na triagem de risco de mortalidade para crianças e adolescentes por COVID-19

O trabalho apresentado nessa tese também quis ampliar os horizontes de aplicação de aprendizado de máquina e levar o contexto dos algoritmos para um sistema prático, permitindo que profissionais de saúde possam utilizar um sistema de suporte à tomada de decisão em um ambiente real. Dessa forma, desenvolvemos um sistema que utiliza o melhor algoritmo de aprendizado de máquina para o desfecho óbito e disponibilizamos o sistema para utilização neste link: https://covidriskcalculator.streamlit.app/. O sistema foi totalmente desenvolvido utilizando-se a linguagem de programação Python.

A Figura 6 apresenta a tela principal do sistema que contém as características selecionadas para treinamento dos modelos, como campos de entradas de dados pelo profissional de saúde. O profissional informa as características clínicas e comorbidades do paciente e após inserir os dados desejados ele seleciona o botão realizar previsão. Na parte esquerda da imagem, o sistema informa o risco de mortalidade do paciente que pode ser alto risco de mortalidade ou baixo risco de mortalidade e informa também o percentual de confiança que o modelo tem naquela previsão de mortalidade informada.



Probabilidade de risco e confiança

Alto Risco de Mortalidade

Triagem de Risco de Mortalidade de Crianças e adolescentes por COVID-19

```
Prezado médico, é obrigatório informar dados em TODOS os campos
 para que o modelo de IA faça a previsão.

Ao lado de alguns campos tem o símbolo ? que mostra informações
 sobre os dados dos campos.

Para a maioria dos campos onde a seleção é 1 ou 0, 1 significa
 presença de comorbidade/característica clínica
 e 0 significa ausência
```

Idade	Etnia	Região	Cepa Viral	Dispneia	Tosse
2	1	1	1	0	0

Desconforto Respiratório	Saturação baixa	Obesidade	Doença Cardíaca
1	1	0	0

Doença Pulmonar	Hipertensão	Imunosuprimido	Doença renal	Asma
0	0	0	0	0

Tem comorbidades?	Comorbidades totais	Comorbidades categórica	Tempo para desfecho (dias)	Doença no sangue
1	3	0	0	0

Distúrbio Neurológico	Câncer	Syndrome de Down	Nosocomial
0	0	0	0

Figura 6 – Tela principal do sistema de suporte a decisão para triagem de risco de mortalidade de crianças e adolescentes por COVID-19

Com o presente sistema esperamos que as aplicações de aprendizado de máquina para saúde saiam do âmbito exclusivamente teórico e possam também contribuir para o contexto prático, ou seja, hospitalar. Isso possibilitará que profissionais de saúde tenham uma ferramenta com inteligência artificial à disposição como suporte para a tomada de decisão. É importante ressaltar que um profissional de saúde não pode definir o destino de um paciente somente baseado no que é informado pelo sistema. O sistema serve apenas como um apontador de indícios. Quaisquer decisões devem ser tomadas pelo médico de acordo com as evidências clínicas e laboratoriais. É importante destacar que este sistema foi desenvolvido apenas para ilustrar as potencialidades das aplicações de IA utilizadas em nosso estudo. É imperativo comentar que a implementação na prática clínica de um sistema desta natureza deve seguir os princípios científicos mais rigorosos e os trâmites legais e éticos previstos nas legislações dos diversos órgãos regulatórios.

7. PERSPECTIVAS

O presente estudo se insere no contexto da utilização da IA e do chamado Big Data em Medicina e de suas promissoras aplicações na área. Essa nova era, propiciada pelo acelerado desenvolvimento tecnológico das últimas décadas, abre um leque de oportunidades inéditas para o avanço do conhecimento científico e para a prática médica. As aplicações de IA podem gerar inovações em diversos aspectos da Medicina, desde a identificação de novos fármacos e o desenvolvimento de terapias personalizadas até a compreensão mais profunda dos complexos mecanismos fisiopatológicos, sociais, e epidemiológicos que determinam a saúde e a doença. Essa vasta gama de dados, proveniente de diversas fontes como bancos de dados públicos, prontuários eletrônicos, pesquisas genômicas e dispositivos médicos de monitoramento contínuos, oferece uma riqueza de informações sem precedentes para a pesquisa médica. Em nossa opinião, para extrair o máximo de conhecimento do Big Data na Medicina, é fundamental a sinergia entre métodos estatísticos tradicionais e técnicas de IA. A combinação da robustez e flexibilidade da estatística com o poder de aprendizado de máquina da IA permite aos pesquisadores analisar conjuntos de dados complexos e identificar padrões sutis que podem ter um impacto significativo na compreensão de doenças e no desenvolvimento de novas terapias.

Contudo, a utilização do Big Data na Medicina não está isenta de desafios. A garantia da privacidade e segurança dos dados, a integração de diferentes fontes de informação e a interpretação dos resultados complexos gerados pelas análises de Big Data exigem soluções inovadoras e colaboração interdisciplinar. No entanto, as perspectivas para o futuro são promissoras. Acredita-se que a IA aliada ao Big Data têm o potencial de revolucionar a Medicina, levando a diagnósticos mais precisos, tratamentos mais eficazes e uma melhor compreensão da saúde humana em sua totalidade.

Ao se inserir nessa área da pesquisa médica, julgamos que este estudo pode contribuir para o avanço do conhecimento e a busca por soluções inovadoras para os desafios da prática clínica em Pediatria. Além disso, a extensiva revisão da literatura indica um amplo espaço para pesquisas com aprendizado de máquina na Pediatria. Para futuros estudos, pretendemos utilizar modelos de aprendizado de máquina para verificação da efetividade das vacinas na prevenção de óbitos por COVID-19. Também pretendemos utilizar esses modelos para diagnósticos e prognósticos de outras doenças em crianças e adolescentes, integrando dados clínicos, laboratoriais e de exames de imagem. Outra área de investigação para trabalhos futuros é sobre a utilização de grandes modelos de linguagem como ChatGPT, Llama e Gemini para auxiliar estudantes, residentes e profissionais de saúde no treinamento do diagnóstico clínico e na atuação nos mais diversos cenários da prática pediátrica.

8. CONCLUSÃO

Esta tese apresentou modelos de aprendizado de máquina para prever a mortalidade em crianças e adolescentes com COVID-19 e seus principais fatores clínicos de risco. Os resultados foram satisfatórios, mostrando que os modelos de aprendizado de máquina podem auxiliar os médicos em um processo de triagem para identificar pacientes com uma probabilidade maior de mortalidade. Os principais fatores preditivos da mortalidade de crianças e adolescentes hospitalizados com COVID-19 de acordo com os algoritmos de aprendizado de máquina utilizados foram: baixa saturação de oxigênio na admissão, dispneia, desconforto respiratório, total de comorbidades apresentadas pelo paciente ou se paciente tem alguma comorbidade. Os resultados dos algoritmos de aprendizado de máquina para os desfechos UTI, suporte ventilatório e gravidade da COVID-19 não tiveram o mesmo desempenho em comparação aos resultados para o desfecho óbito.

REFERÊNCIAS

Youssef D, Youssef J, Abou-Abbas L, Kawtharani M, Hassan H. Prevalence and correlates of burnout among physicians in a developing country facing multi-layered crises: a cross-sectional study. Sci Rep. 2022 Dec 1;12(1).

Ulutasdemir N, Luna A, Tusconi M, Ryan E. The relationship between physician burnout and depression, anxiety, suicidality and substance abuse: A mixed methods systematic review [Internet]. Available from: https://www.crd.york.ac.uk/prospero/display_re

Jung F, Bodendieck E, Bleckwenn M, Hussenoeder F, Luppa M, Riedel-Heller S. Burnout, work engagement and work hours – how physicians' decision to work less is associated with work-related factors. BMC Health Serv Res. 2023 Dec 1;23(1).

Patel RS, Bachu R, Adikey A, Malik M, Shah M. Factors related to physician burnout and its consequences: A review. Vol. 8, Behavioral Sciences. MDPI Multidisciplinary Digital Publishing Institute; 2018.

Marques-Pinto A, Moreira S, Costa-Lopes R, Zózimo N, Vala J. Predictors of Burnout Among Physicians: Evidence From a National Study in Portugal. Front Psychol. 2021 Oct 1;12.

Yates SW. Physician Stress and Burnout. Vol. 133, American Journal of Medicine. Elsevier Inc.; 2020. p. 160–4.

Verdonk C, Verdonk F, Dreyfus G. How machine learning could be used in clinical practice during an epidemic. Vol. 24, Critical Care. BioMed Central Ltd.; 2020.

Ahuja AS. The impact of artificial intelligence in medicine on the future role of the physician. PeerJ. 2019;2019(10).

Abramoff MD, Whitestone N, Patnaik JL, Rich E, Ahmed M, Husain L, Hassan MY, Tanjil MSH, Weitzman D, Dai T, Wagner BD, Cherwek DH, Congdon N, Islam K. Autonomous artificial intelligence increases real-world specialist clinic productivity in a cluster-randomized trial. NPJ Digit Med. 2023 Oct 4;6(1):184. doi: 10.1038/s41746-023-00931-7. PMID: 37794054; PMCID: PMC10550906.

Deo RC. Machine learning in medicine. Circulation. 2015 Nov 17;132(20):1920–30.

Lages A, Santos D, Pinhati C, Perdigão J, Galante S, Silva L, et al. Machine learning algorithms to predict outcomes in children and adolescents with COVID-19: A systematic review ☆. Artif Intell Med [Internet]. 2024;150:102824. Available from: https://doi.org/10.1016/j.artmed.2024.102824

Bastos LS, Ranzani OT, Souza TML, Hamacher S, Bozza FA. COVID-19 hospital admissions: Brazil's first and second waves compared. Lancet Respir Med. 2021;9(8):e82-e3.

Ranzani OT, Bastos LSL, Gelli JGM, Marchesi JF, Baiao F, Hamacher S, et al. Characterisation of the first 250,000 hospital admissions for COVID-19 in Brazil: a retrospective analysis of nationwide data. Lancet Respir Med. 2021;9(4):407-18.

Bastos LS, Niquini RP, Lana RM, Villela DAM, Cruz OG, Coelho FC, et al. COVID-19 and hospitalizations for SARI in Brazil: a comparison up to the 12th epidemiological week of 2020. Cad Saude Publica. 2020;36(4):e00070120.

Castro MC, Massuda A, Almeida G, Menezes-Filho NA, Andrade MV, de Souza Noronha KVM, et al. Brazil's unified health system: the first 30 years and prospects for the future. Lancet. 2019;394(10195):345-56.

Rocha R, Atun R, Massuda A, Rache B, Spinola P, Nunes L, et al. Effect of socioeconomic inequalities and vulnerabilities on health-system preparedness and response to COVID-19 in Brazil: a comprehensive analysis. Lancet Glob Health. 2021;9(6):e782-e92.

Baqui P, Bica I, Marra V, Ercole A, van der Schaar M. Ethnic and regional variations in hospital mortality from COVID-19 in Brazil: a cross-sectional observational study. Lancet Glob Health. 2020;8(8):e1018-e26.

Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Stat Med. 2007;26(11):2389-430.

Fine JP, Gray RJ. A proportional hazards model for the sub-distribution of a competing risk. J Am Stat Assoc. 1999;94:496-509.

Schafer JL. Multiple imputation: a primer. Stat Methods Med Res. 1999;8(1):3-15.

Austin PC, Lee DS, D'Agostino RB, Fine JP. Developing points-based risk-scoring systems in the presence of competing risks. Stat Med. 2016;35(22):4056-72.

Marchezini GF, Lacerda AM, Pappa GL, Meira W, Jr., Miranda D, Romano-Silva MA, et al. Counterfactual inference with latent variable and its application in mental health care. Data Min Knowl Discov. 2022;36(2):811-40.

Andaur Navarro CL, Damen JAA, van Smeden M, Takada T, Nijman SWJ, Dhiman P, et al. Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models. J Clin Epidemiol. 2023;154:8-22.

Sankaranarayanan S, Balan J, Walsh JR, Wu Y, Minnich S, Piazza A, et al. COVID-19 Mortality Prediction From Deep Learning in a Large Multistate Electronic Health Record and Laboratory Information System Data Set: Algorithm Development and Validation. J Med Internet Res. 2021;23(9):e30157.

Pearl J. Causal inference in the health sciences: a conceptual introduction 2001; 2: 189-220. Health Services and Outcomes Research Methodology. 2001;2:189-220.

APÊNDICE

Supplementary Material Systematic Literature Review

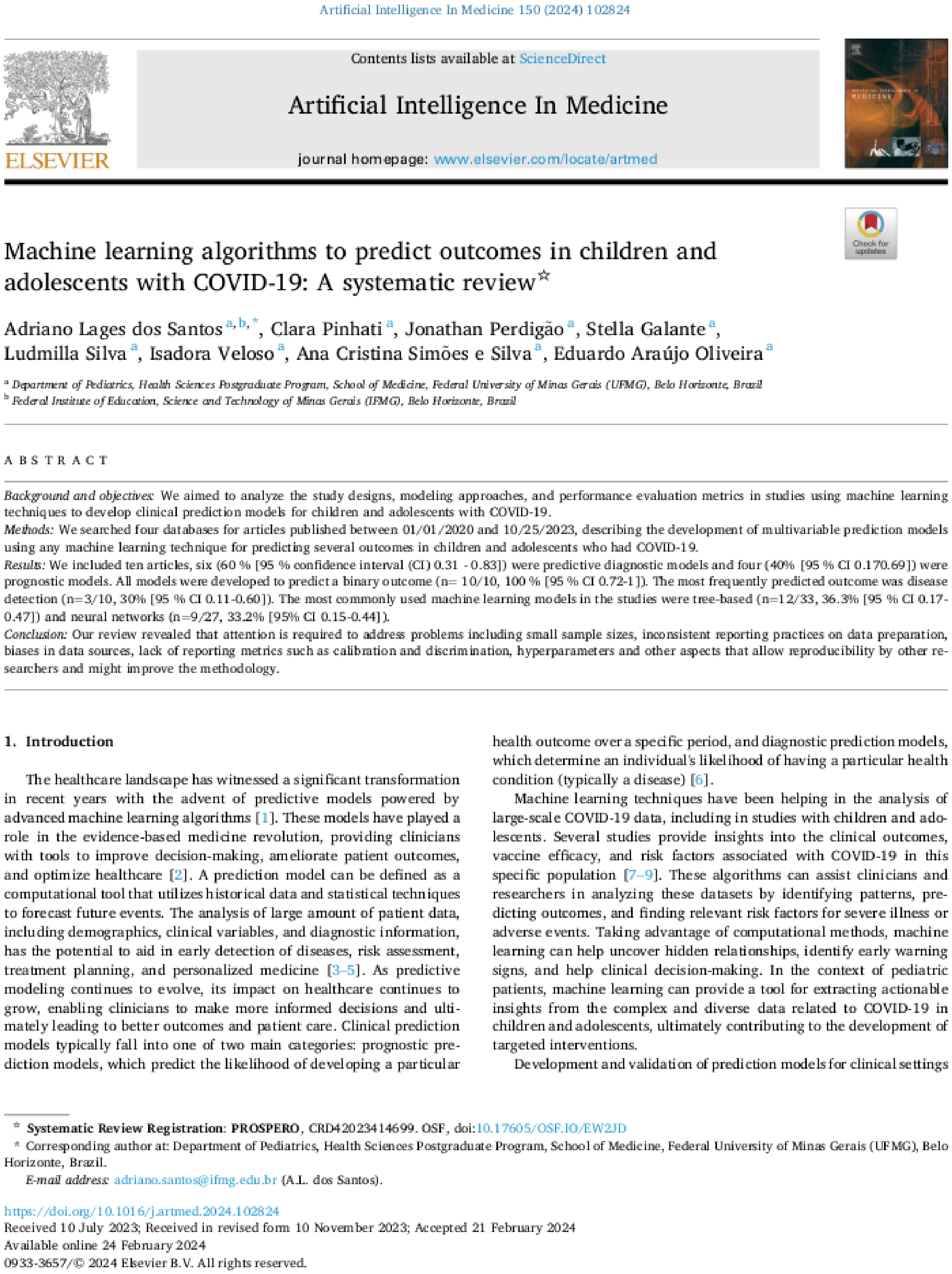
Supplemental File 1 - Selected studies after selection process

	retrieved studies	Studies selected by abstract and title	Duplicate studies	Studies that do not address COVID-19 in children or adolescents	Studies where the PDF was not found	Selected studies after complete screening
Elsevier Scopus	509	2	0	0	0	2
Embase	1406	7	0	4	0	4
Pubmed	1995	7	6	0	0	1
Google Scholar	1112	8	1	1	3	3
Total	**5022**	**25**	7	5	3	**10**

Supplemental File 2 - Tripod adherence score per study

Study	Tripod adherence score per study (Total Tripod checklist items: 31)
Byeon2022	24 (77,41%)
Cetin et al 2022	26 (83,87%)
Gao 2022	21 (67,74%)
Liu2022	26 (83,87%)
Ma2021	25 (80,64%)
Magrelli2021	14 (45,16%)
Mamlook2021	15 (48,38%)
Nugawela2022	16 (51,61%)
Pavliuk2022	16 (51,61%)
zhang2023	25(80,64%)
Mean	20,8 (67,09%)

Artigo publicado de revisão sistemática da literatura



Artificial Intelligence In Medicine 150 (2024) 102824

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

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

Machine learning algorithms to predict outcomes in children and adolescents with COVID-19: A systematic review[☆]

Adriano Lages dos Santos[a,b,*], Clara Pinhati[a], Jonathan Perdigão[a], Stella Galante[a], Ludmilla Silva[a], Isadora Veloso[a], Ana Cristina Simões e Silva[a], Eduardo Araújo Oliveira[a]

[a] *Department of Pediatrics, Health Sciences Postgraduate Program, School of Medicine, Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil*
[b] *Federal Institute of Education, Science and Technology of Minas Gerais (IFMG), Belo Horizonte, Brazil*

ABSTRACT

Background and objectives: We aimed to analyze the study designs, modeling approaches, and performance evaluation metrics in studies using machine learning techniques to develop clinical prediction models for children and adolescents with COVID-19.
Methods: We searched four databases for articles published between 01/01/2020 and 10/25/2023, describing the development of multivariable prediction models using any machine learning technique for predicting several outcomes in children and adolescents who had COVID-19.
Results: We included ten articles, six (60 % [95 % confidence interval (CI) 0.31 - 0.83]) were predictive diagnostic models and four (40% [95 % CI 0.170.69]) were prognostic models. All models were developed to predict a binary outcome (n= 10/10, 100 % [95 % CI 0.72-1]). The most frequently predicted outcome was disease detection (n=3/10, 30% [95 % CI 0.11-0.60]). The most commonly used machine learning models in the studies were tree-based (n=12/33, 36.3% [95 % CI 0.17-0.47]) and neural networks (n=9/27, 33.2% [95% CI 0.15-0.44]).
Conclusion: Our review revealed that attention is required to address problems including small sample sizes, inconsistent reporting practices on data preparation, biases in data sources, lack of reporting metrics such as calibration and discrimination, hyperparameters and other aspects that allow reproducibility by other researchers and might improve the methodology.

1. Introduction

The healthcare landscape has witnessed a significant transformation in recent years with the advent of predictive models powered by advanced machine learning algorithms [1]. These models have played a role in the evidence-based medicine revolution, providing clinicians with tools to improve decision-making, ameliorate patient outcomes, and optimize healthcare [2]. A prediction model can be defined as a computational tool that utilizes historical data and statistical techniques to forecast future events. The analysis of large amount of patient data, including demographics, clinical variables, and diagnostic information, has the potential to aid in early detection of diseases, risk assessment, treatment planning, and personalized medicine [3–5]. As predictive modeling continues to evolve, its impact on healthcare continues to grow, enabling clinicians to make more informed decisions and ultimately leading to better outcomes and patient care. Clinical prediction models typically fall into one of two main categories: prognostic prediction models, which predict the likelihood of developing a particular health outcome over a specific period, and diagnostic prediction models, which determine an individual's likelihood of having a particular health condition (typically a disease) [6].

Machine learning techniques have been helping in the analysis of large-scale COVID-19 data, including in studies with children and adolescents. Several studies provide insights into the clinical outcomes, vaccine efficacy, and risk factors associated with COVID-19 in this specific population [7–9]. These algorithms can assist clinicians and researchers in analyzing these datasets by identifying patterns, predicting outcomes, and finding relevant risk factors for severe illness or adverse events. Taking advantage of computational methods, machine learning can help uncover hidden relationships, identify early warning signs, and help clinical decision-making. In the context of pediatric patients, machine learning can provide a tool for extracting actionable insights from the complex and diverse data related to COVID-19 in children and adolescents, ultimately contributing to the development of targeted interventions.

Development and validation of prediction models for clinical settings

[☆] **Systematic Review Registration**: **PROSPERO**, CRD42023414699. OSF, doi:10.17605/OSF.IO/EW2JD

[*] Corresponding author at: Department of Pediatrics, Health Sciences Postgraduate Program, School of Medicine, Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil.

E-mail address: adriano.santos@ifmg.edu.br (A.L. dos Santos).

https://doi.org/10.1016/j.artmed.2024.102824
Received 10 July 2023; Received in revised form 10 November 2023; Accepted 21 February 2024
Available online 24 February 2024

rely on the use of appropriate study designs and modeling strategies. However, there is a lack of comprehensive information regarding the specific study designs, modeling approaches, and performance measures employed in studies that utilize machine learning for prediction modeling [10]. Therefore, our objective was to conduct a systematic review to analyze and summarize the key characteristics related to study design, modeling techniques, and performance measures reported in studies focusing on clinical prediction models developed using supervised machine learning algorithms in pediatric patients with COVID-19.

2. Methods

We followed the PRISMA 2020 guidelines for systematic reviews [11]. This systematic review was registered and approved in PROSPERO under the protocol CRD42023414699 and in OSF available at doi: 10.17605/OSF.IO/EW2JD.

The systematic mapping was conducted following three adopted stages described below [12].

Step 1 - Conduct searches: Based on the research questions, a replicable method for searching and retrieving articles in four selected scientific databases was defined and executed. The databases were Embase, Google Scholar, Pubmed, and Scopus Elsevier.

Step 2 - Selection of studies: A systematic method was defined and applied to select only the relevant articles for this study using inclusion and exclusion eligibility criteria. We used the open-source software Zotero (version 6.0. 26) to exclude duplicate articles from the search results.

Step 3 - Data extraction and analysis: Finally, the relevant data from the primary studies were summarized and presented in this study. For each study, we collected the following information: study design characteristics (such as cohort, case-control, randomized trial), data source (such as routinely collected data, registries, administrative databases), study population details, outcome measures, setting information, patient characteristics, sample size (before and after participant exclusion), number of events, number of candidate and final predictors, handling of missing data, hyperparameter optimization, dataset splitting (such as train-validation-test), method for internal validation (such as bootstrapping, cross-validation), number of models developed and/or validated, and availability of code, data, and model. Country was defined based on the location of the first author's affiliation. For each model, we extracted information on the algorithm used, predictor selection methods, variable importance reporting, use of penalization techniques, hyperparameters reporting, and performance metrics (such as discrimination and calibration).

2.1. Step 1 - Search strategy for scientific articles

To identify possible primary studies relevant to data extraction, the search was based on (i) studies using keyword combinations derived from our objective and (ii) the execution of automatic searches on scientific databases using search terms. Initially, relevant keywords related to four main fields were selected: (a) COVID-19; (b) medicine; (c) early childhood, childhood, and adolescence; (d) Artificial Intelligence and Machine Learning.

The resulting keywords for each main field were:

COVID-19: COVID-19 OR SARS-COV-2.

Medicine: outcomes OR outcome OR mortality OR death OR hospitalization OR hospitalized OR ICU OR ventilation.

Population: Early childhood, childhood, adolescence: child OR "early childhood" OR children OR newborn OR adolescent OR adolescents.

AI and Machine Learning: "machine learning" OR "artificial intelligence" OR algorithm OR algorithms OR dataset OR dimensions OR training OR sample OR samples OR prediction OR predict OR predicting OR forecast OR forecasting OR classification OR regression OR dimension OR models OR model OR predictive OR predictors OR bootstrapping OR bootstrap.

Search terms were defined by grouping keywords in the same domain with the logical operator "OR" and grouping the three main concepts with the logical operator "AND". Then, automatic searches were executed on four scientific databases, including Embase, Google Scholar, Pubmed, and Scopus Elsevier. The search limited articles by year of publication (2019 to 2023).

2.2. Step 2 - Eligibility criteria (selection of studies)

The studies retrieved from automatic searches were filtered to exclude articles not aligned with the study objectives. At this stage, three independent researchers defined and applied the following inclusion and exclusion criteria.

2.2.1. Inclusion criteria

Studies whose main focus is on the use of machine learning algorithms to predict deaths and other outcomes in children or adolescents who had COVID-19.

The search period comprises 01/01/2020 to 10/25/2023. The year limit of 2019 was used because some databases did not allow filtering with monthly granularity. Thus, it was not possible to specify the month of March 2020 (the beginning of the pandemic).

To be included in the first selection, articles must address the topics of COVID-19 in children or adolescents and use machine learning algorithms to predict various outcomes in these patients. Although the outcome of death is highlighted in the search keywords in the Medicine domain, this search also considered other outcomes to increase the range of possible articles returned in the search. Only articles written in English were considered for the search. Only articles published in journals or conferences were considered for this search. Regarding articles published in conferences, we consider those papers presented at conferences and published in the conference proceedings.

2.2.2. Exclusion criteria

Articles written in languages other than English. Articles that do not deal with COVID-19 in children and adolescents, articles that do not use machine learning algorithms in the prediction of various COVID-19 outcomes, duplicated articles, and articles that were selected in the databases but whose completed text files were not obtained even after demanding the corresponding authors.

The study selection process was carried out in two phases: (i) in the first selection phase, the titles and abstracts of the studies retrieved from the searches were read, and studies that did not meet the inclusion criteria were excluded; (ii) in the second selection phase, all articles were downloaded, and their introduction and conclusion were read to remove studies that met the exclusion criteria.

For this review, we did not use the "snowballing" technique, which involves checking if there are any articles in the references of the selected articles, after a complete reading, that were not found in the initial database search. If such articles are identified, they are then selected for inclusion in the review. Fig. 1 presents the number of articles selected after each phase and the application of inclusion and exclusion criteria. And the table in the Supplemental File 1 also summarize the results after each phase.

2.3. Screening and selection process

The titles and abstracts were thoroughly examined by three researchers independently from a team of eight researchers to identify studies that potentially met the eligibility criteria. The group of researchers comprised two senior medical professors, a doctoral candidate, and five undergraduate medical students. The undergraduate medical students and the doctoral student were involved in research projects related to the effects of COVID-19 in children and adolescents. Subsequently, full-text articles were obtained, and three groups of two

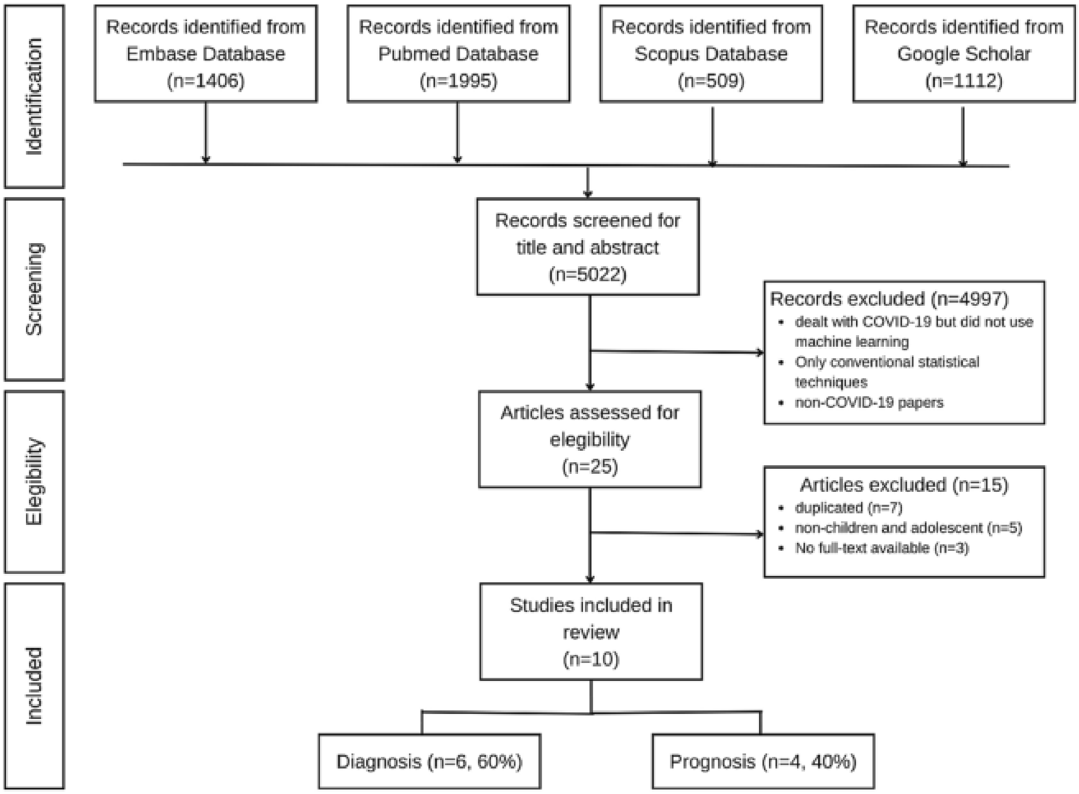

Identification
Screening
Elegibility
Included
Records identified from Embase Database (n=1406)
Records identified from Pubmed Database (n=1995)
Records identified from Scopus Database (n=509)
Records identified from Google Scholar (n=1112)
Records screened for title and abstract (n=5022)
Records excluded (n=4997)
• dealt with COVID-19 but did not use machine learning
• Only conventional statistical techniques
• non-COVID-19 papers
Articles assessed for elegibility (n=25)
Articles excluded (n=15)
• duplicated (n=7)
• non-children and adolescent (n=5)
• No full-text available (n=3)
Studies included in review (n=10)
Diagnosis (n=6, 60%)
Prognosis (n=4, 40%)

Fig. 1. Flowchart of included studies.

researchers independently evaluated all articles, while the same articles were collectively reviewed by four researchers to ensure agreement. In the event of any discrepancies during the screening and selection process, the primary reviewer of this study was consulted to assess the concerned article and resolve discrepancies carefully.

2.4. Step 3 - Data extraction

We selected several items from existing methodological guidelines for reporting and critical appraisal of prediction model studies to build our data extraction form (TRIPOD and PROBAST) [13–14]. The following items were extracted in the selected studies based on the systematic review conducted by Navarro et al. [10], including the items described in step 3 of our methodology. One reviewer recorded all items, while the other reviewers collectively assessed all articles. Articles were randomly assigned to reviewers. Discrepancies in data extraction were discussed and solved between the pair of reviewers. No limitations were imposed on the number of models extracted per article.

2.5. Summary statistics and integration of findings

The findings were condensed into percentages (with confidence intervals calculated using the Wilson score interval and the Wilson score continuity-corrected interval, as appropriate), medians, and interquartile range (IQR), accompanied by a descriptive synthesis.

We reported only overall performance data from the studies, specifically the overall mean performance reported in the studies. We did not differentiate performance into corrected, external validation, or apparent validation segments. We did not report external validation data, even for studies that validated their models using different data from the model development and testing phase.

Rather than assessing the intricacies of each modeling approach and its performance, our evaluations remained at the study level. We refrained from conducting a quantitative synthesis of the models' performance, such as a meta-analysis, as it fell outside the scope of our review due to the reason that the available studies on the topic may have significant heterogeneity in terms of study design, patient populations, interventions, or outcomes, making it inappropriate or unreliable to combine their results quantitatively. All analyses were conducted using the software R version 4. 1.0 (R Core Team, Vienna, Austria).

3. Results

The search in the selected databases for this review yielded 5022 articles. After assessing the titles and abstracts, 25 studies potentially met the eligibility criteria. Following a thorough reading of all 25 studies, ten articles were included in this review: 6 (60 % [95 % confidence interval (CI) 0. 31-0.83]) were predictive diagnostic models and 4 (40% [95 % CI 0. 17-0.69]) were prognostic models (Fig. 1).

We evaluated the quality of the articles regarding their adherence to the TRIPOD guidelines and also assessed the risk of bias in the selected studies using the PROBAST tool. Regarding the adherence to the TRIPOD guidelines, the selected studies showed an average adherence of 67.09 %. TRIPOD is a checklist consisting of 31 items, and the selected studies, on average, fulfilled 20 items from this checklist. The results of the adherence assessment of each article to the TRIPOD guidelines can be found in Supplemental File 2.

Regarding the risk of bias assessment using the PROBAST tool, five studies showed a high risk of bias concerning their prediction models, four studies showed a low risk of bias, and one study had an unclear result regarding bias risk. The results of the assessment for each study in the dimensions evaluated by PROBAST (Participants, Predictors, Outcome, and Analysis) can be found in Supplemental File 3.

Among the 10 articles, 7 studies (70 % [95 % CI 0.40 - 0.89]) developed prediction models and assessed their performance using internal validation techniques, while 3 studies (30% [95 % CI 0.11 - 0.60]) developed and externally validated the same machine learning predictive model. Six studies were published in 2022 (60 % [95 % CI 0.31 - 0.83]), three in 2021 (30% [95 % CI 0.11 - 0.60]) and one study in 2023 (10% [95 % CI 0.018–0. 40]). The clinical fields involved in the selected articles were pediatrics (n=7/10, 70% [95% CI 0.40-0.89]), public health (n=2/10, 20 % [95 % CI 0.057–0.51]), and pulmonology (n=1/10 % [95 % CI 0.018-0.40]). The retrieved articles originated from Europe (n=4/10, 40% [95 % CI 0.17-0.69]), Asia (n=3/10, 30% [95 % CI 0.11 - 0.60]), and North America (n=3/10, 30% [95 % CI 0.11-0.60]). Other study characteristics are presented in Table 1.

In total, 33 prediction models were developed (Mean: 3 models per study, IQR: 4, Range: 1-5). We did not set a limit for extracting models per study, since were few articles included in this review. Thus, all 33 models found in the selected studies were evaluated. The most commonly used machine learning models in the studies were tree-based (n=12/33, 36.36 % [95 % CI 0.17-0.47]) and neural networks (n=9/27, 33.27 % [95 % CI 0.15-0.44]). Other algorithms encountered are described in Table 2.

Table 1
General characteristics of the included studies.

Key characteristics	Total (n = 10)
	n (%) [95 % CI]
Study aim	
Diagnosis	6 (60) [0.31–0.83]
Prognosis	4 (40) [0.17–0.69]
Study type	
Model development only	7 (70) [0.40–0.89]
Model development with external validation	3 (30) [0.11–0.60]
Outcome aim	
Classification	6 (40) [0.31–0.83]
Risk probabilities	4 (40) [0.17–0.69]
Setting[a]	
General population	6 (60) [0.31–0.83]
Secondary care	1 (10) [0.018–0.40]
Tertiary care	3 (30) [0.11–0.60]
Outcome format	
Binary	10 (100) [0.72–1]
Type of outcome	
Death	1 (10) [0.018–0.40]
Severity prediction	1 (10) [0.018–0.40]
Hospitalization prediction	2 (20) [0.063–0.55]
Complications	2 (20) [0.063–0.55]
Need of ICU	1 (10) [0.018–0.4]
Disease detection	3 (30) [0.11–0.60]
Mentioning reporting guidelines (Tripod, Strobe, Charms, other)	
TRIPOD	1 (10) [0.018–0.4]
None	9 (90) [0.60–0.98]
Model availability[a]	
Repository for data	5 (50) [0.24–0.76]
Repository for code	2 (20) [0.057–0.51]
Model presentation	8 (80) [0.49–0.94]
None	2 (20) [0.057–0.51]

[a] Counts are absolute numbers with column percentages in parentheses. The percentages sometimes do not add up to 100 % because studies reported more than one option. ICU = intensive care unit.

3.1. *Participants*

The participants included in the reviewed studies were recruited from the general population (n=6/10, 60 % [95 % CI 0.31 - 0.83]), tertiary care settings (n=3/10, 30 % [95 % CI 0.11-0.60]), and secondary care settings (n=1/10 % [95 % CI 0.018-0.40]) (Table 1).

3.2. *Data sources*

The prediction models were predominantly developed using administrative databases (n=7/10, 70 % [95 % CI 0.40-0.89]). Prospective cohort data (n=1/10 % [0.018 - 0.40]) and retrospective cohort data (n=1/10 % [0.018-0.40]) were reported in one study each. The reviewed studies utilized electronic medical records and surveys. However, no information was available in the selected articles regarding the time spent on data collection for the studies. Similarly, no studies reported the time horizon for the predictions (n=10/10, 100 % [95 % CI 0.72-1]).

3.3. *Outcomes*

All models were developed to predict a binary outcome (n=10/10, 100 % [95 % CI 0.72-1]). The most frequently predicted outcome was disease detection (n=3/10, 30 % [95 % CI 0.11-0.60]) followed by hospitalization prediction and complications both with two studies each (n=2/10, 20 %, [95 % CI 0.057-0.51]). Other outcomes of severity prediction are described in Table 1.

3.4. *Candidate predictors*

Candidate predictors extracted from the studies were clinical history (n=5/10, 50 % [95 % CI 0.24 - 0.76]), demographics including sex, gender, and ethnicity/race (n=5/10, 50 % [95 % CI 0.24 - 0.76]) and disease (the diagnosed disease) (n=5/10, 50 % [95 % CI 0.24-0.76]). Other predictors extracted (physical examination, blood or urine parameters, imaging, pathology, and questionnaires) are described in Table 4. None of the selected studies used treatment modalities as predictors for the developed models and for one study, treatment as a candidate predictor is not applicable, since the developed models are dealing with imaging data. Studies included a median of 15 candidate predictors (IQR: 6-14. 5). Four studies included continuous variables as candidate predictors (40% [95 % CI 0.17-0.69]), the other three studies did not use continuous variables as predictors (30% [95 % CI 0.11 -

Table 2
Algorithms used for modeling in all extracted models from the selected studies.

Modeling algorithm[a]	All extracted models (n = 33)
	n (%) [95 % CI]
Tree Based Models	12 (36.36) [0.17–0.47]
Decision trees (for example, CART)	3 (25) [0.089–0.53]
Random forest	2 (16.57) [0.047–0.45]
Gradient boosting machine (Catboost)	3 (25) [0.089–0.53]
XGBoost	4 (33.43) [0.14–0.61]
Neural Network (incl. deep learning)	9 (27.27) [0.15–0.44]
Support Vector Machine	2 (6.06) [0.017–0.20]
Naive Bayes	1 (3.03) [0.0054–0.15]
Multiple logistic regression	1 (3.03) [0.0054–0.15]
Logistic regression	4 (12.12) [0.048–0.27]
Linear discriminant analysis	2 (6.06) [0.017–0.20]
Other (TabNet, AutoM, DeepFM, etc)	3 (9.09) [0.031–0.24]

[a] Counts are absolute numbers with column percentages in parentheses. The percentages sometimes do not add up to 100 % because studies developed more than one model.

0.60]). Most studies did not report the methods to handle continuous predictors (60 % [95 % CI 0.31-0.83]).

3.5. *Sample size*

Selected studies had a median sample size of 11,108 participants (IQR: 5,664–65,518). Most studies report a sample size justification or calculation rationale as the size of existing/available data used (n=7/10, 70 % [95 % CI 0.40-0.89]), and three studies did not report any rationale about sample size (n=3/10, 30 % [95 % CI 0.11-0.66]) (Table 3, Table 5).

3.6. *Missing values*

Missing values were an exclusion criterion of participants in three studies (30% [95 % CI 0.11-0.60]). On the other hand, seven studies were unclear regarding missing data being a criterion for excluding participants, as we did not find this information (70 %, [95 % CI 0.40-0.89]). When a study did not explicitly mention that there are no missing data, we consider that the study was not clear about the existence of missing data. To handle missing data, most of the studies are unclear (n=4/10, 40 % [95 % CI 0.17 - 0.69]). One study used Bayesian optimization (n=1/10 % [95 % CI 0.018-0.40]), and two studies did not make imputation of the missing data in the data source (n=2/10, 20 % [95 % CI 0.057-0.51]). Other information about how studies reported ways to handle missing data is presented in Table 6.

3.7. *Class imbalance and dimensionality reduction techniques*

Eight among 10 studies (80 %, [95 % CI 0.49–0.94]) did not report unbalanced data or any strategy to deal with class imbalance like Synthetic Minority Oversampling Technique (SMOTE), Random Undersampling Boosting (RUSBoost), Random oversampling, random under sampling, or other techniques. For one study class imbalance is not applicable, since the study deals with imaging as a data source and one study report the use of SMOTE to deal with class imbalance. Regarding dimensionality reduction, most studies did not report any technique to reduce the dimension of data (n=8/10, 80 % [95 % CI 0.49-0.94]). One study used principal component analysis (PCA) to reduce the dimension of data (Table 7).

Table 3
Study design of included studies.

Key items[a]	Total (n = 10)
	n (%) [95 % CI]
Data sources	
Prospective cohort	1 (10) [0.018–0.40]
Retrospective cohort	1 (10) [0.018–0.40]
Electronic health record	1 (10) [0.018–0.40]
Administrative databases	7 (70) [0.4–0.89]
Survey	1 (10) [0.018–0.40]
Predictor horizon	
None	10 (100) [0.72–1]
Sample size justification	
Size of existing/available data	7 (70) [0.40–0.89]
None	3 (30) [0.11–0.66]
Internal validation[a]	
Split sample with test set	9 (90) [0.60–0.98]
(Random) split	5 (50) [0.24–0.76]
(Nonrandom) split	2 (20) [0.018–0.59]
Split	1 (10) [0.022–0.40]
Bootstrapping	1 (10) [0.022–0.40]
With test set	1 (100) [0.21–1]
Cross-validation	5 (50) [0.24–0.76]
Nested	5 (100) [0.57–1]
External validation	3 (30) [0.11–0.60]
Independent dataset	3 (100) [0.44–1]

[a] Counts are absolute numbers with column percentages in parentheses. The percentages sometimes do not add up to 100 % because studies reported more than one option.

3.8. *Modeling algorithms*

Neural networks were used in 9 out of 33 models (27. 27 % [95 % CI 0.15-0.44]) extracted from the selected studies, including multilayer perceptron, convolutional neural network, and recurrent neural networks. Tree-based models were reported in 12 of 33 models (36. 36 % [95 % CI 0.17 - 0.47]). Other models such as TabNet, AutoML, and DeepFM were also adopted in the selected studies (n=3/9, 33.09 % [95 % CI 0.031-0.24]). We did not find any study that reported penalized regression models. Support Vector Machine (SVM), a popular machine learning technique, was also reported two times (n=2/33 6.06 % [95 % CI 0.017-0.20]).

3.9. *Selection of predictors*

Regarding the strategy to build models, different methods of selection of predictors were reported as presented in Table 8. Some of the strategies found in the selected studies include term frequency-inverse document frequency (TF-IDF) embedding, frequency encoding, and embedding in the learning process (data-driven approach), decided by pediatricians and others. The most cited method for model building was Spearman Correlation (n=4/12, 33.12 % [95 % CI 0.048-0.27]).

3.10. *Variable importance and hyperparameters*

The variable importance scores provide valuable information on how much each variable contributed to the prediction model (Probst et al., 2019). Despite our small sample of studies, we found a heterogeneity of information about variable importance. Three studies did not provide any information about scores for variables (9.09 % [95 % CI 0.031-0.24]). For 4/33 (12.12 % [95 % CI 0.048-0.27]) the importance weights of variables/correlations were used to report variable importance to the models. Shapley values, another method to determine importance, were used in two studies (6.06 % [95 % CI 0.017 - 0.20]). Other methods informed by studies to determine variable importance are defined in Table 8. Hyperparameters (including default settings of models) were not reported in 7/10 (70 % [95 % CI 0.40–0.89]) studies. Cross Validation was the most described strategy for hyperparameter optimization (n=4/10, 40 % [95 % CI 0.17-0.69]). Seven studies did not report any information about hyperparameter optimization (n=7/10, 70 % [95 % CI 0.40-0.89]), as shown in Table 7.

3.11. *Performance metrics*

The most used measure for the extracted models was the area under the Receiver Operating Characteristic curve (AUC/ROC) (n=15/15, 33.15 % [95 % CI 0.30-0.62]) to describe the discriminative ability of the proposed models (Table 9). Few methods for measuring agreement between predictions and observations (also called calibration) were used in the selected studies. Only four models used a calibration plot (12. 12%, [95% CI 0.048-0.27]). Other measures of calibration used were calibration slope and calibration-in-the-large. General metrics were found in most studies for the developed models, such as accuracy (*n*=25/33, 75.75 % [95 % CI 0.59-0.87]) and F1-score (n=12/33, 36.36 % [95 % CI 0.22-0.53]).

3.12. *Predictive performance*

Studies that reported their discriminative abilities of the proposed models had solid results (AUC next to 1) with an internally validated median AUC of 0.91 (IQR 0.76–0.98; range 0.68–0.98). For calibration and overall performance metrics, as shown in Table 10.

3.13. Internal validation and external validation

Nine out of 10 studies (88. 9% [95 % CI 0.60–0.98]) internally validate their models, splitting samples into a training and test set. The train-test set was split randomly into 5/ 10 studies (50 % [95 % CI 0.24 - 0.76]) and 2/10 studies used a nonrandom split (20% [95 % CI 0.057, 0.51]). One study reported bootstrapping on a test set without citing the number of iterations. Five studies that performed cross-validation (50 % [95 % CI 0.24-0.76]), all of them used nested cross-validation (100 % [95 % CI 0.57–1]). For further details, see Table 4. Only three studies performed an external validation of their models (30% [95 % CI 0.11-0.60]) by using independent datasets to validate their models (100 % [95 % CI 0.44 – 1]).

3.14. Model availability

We did not find any studies that created an online calculator or web system containing some way to use the developed models. We found a repository for data in five studies (n=5/10, 50 % [95 % CI 0.24-0.76]), and in two studies we did not find any information about data, code, and even a detailed description of model construction (n=2/10, 20 % [95 % CI 0.057-0.51]). The presentation of the models in detail with flowcharts or other images that convey the architecture of the solution proposed in the study was found in eight articles (80 % [95 % CI 0.49–0.94]). We found two studies that reported a repository for accessing and reading the source code of the developed model (Table 1).

4. Discussion

4.1. Principal findings

Table 4
Predictors in included studies.

Key items	Total (n = 10)
	n (%) [95 % CI]
Type of candidate predictors[a]	
Demography	5 (50) [0.24–0.76]
Clinical history	5 (50) [0.24–0.76]
Physical examination	3 (30) [0.11–0.6]
Disease	5 (50) [0.24–0.76]
Blood or urine parameters	3 (30) [0.11–0.6]
Imaging	1 (10) [0.018–0.40]
Pathology	3 (30) [0.11–0.60]
Questionnaires	1 (10) [0.018–0.40]
Scale Score	1 (10) [0.018–0.40]
Treatment as candidate predictor	
Yes	
No	9 (90) [0.60–0.98]
Not applicable	1 (10) [0.018–0.40]
Continuous variables as candidate predictors[b]	
Yes	4 (40) [0.17–0.69]
No	3 (30) [0.11–0.60]
Unclear	3 (30) [0.11–0.60]
A-priori selection of candidate predictors	
Yes	5 (50) [0.24–0.76]
No	5 (50) [0.24–0.76]
Methods to handle continuous predictors[a,b]	
Nonlinear (planned)	1 (10) [0.018–0.40]
Unclear	6 (60) [0.31–0.83]
Not applicable	3 (30) [0.11–0.60]
Categorization of continuous predictors[b]	
Not reported	10 (100) [0.72–1]

[a] Counts are absolute numbers with column percentages in parentheses. The percentages sometimes do not add up to 100 % because studies reported more than one option.

[b] As data preparation.

Table 5
Sample size of included studies.

Key items	Total (n = 10)	
	n (%) [95 % CI]	Median [IQR], range
Initial sample size	10 (100)	11,108 [5664–65,518] 105 to 23 million
Final sample size	10 (100)	7801 [5664–65,518] 99 to 23 million
Model development	10 (100)	6955 [3000–65,518] 99 to 20 million
Internal validation	9 (88. 9)	7139 [799–58,188] 99 to 16 million
External validation[a]	3 (22. 3)	Not significant
Number of candidate predictors	10 (100)	23 [14–33] 3 to 200
Number of included predictors	10 (100)	16 [7–21] 3 to 65

[a] Only three studies conducted external validation. In order for the IQR calculation to have significance, a minimum of four values is required.

Table 6
Methods used for missing values handling.

Key items	Total (n = 10)
	n (%) [95 % CI]
Missingness as exclusion criteria for participants	
Yes	3 (30) [0.11–0.6]
Unclear	7 (70) [0.4–0.89]
Number of patients excluded	
Median [IQR] (range)	1007 [303–6,247,840] (6 to 12,494,266)
Methods of handling missing data	
No missing data	3 (30) [0.11–0.6]
No imputation	2 (20) [0.057–0.51]
Bayesian optimization	1 (10) [0.018–0.4]
Unclear	4 (40) [0.17–0.69]
Presentation of missing data	
Not summarized	6 (60) [0.31–0.83]
By all candidate predictors	1 (10) [0.018–0.4]
Not applicable	3 (30) [0.11–0.6]

The present review aimed to identify and analyze predictive and prognostic models developed using machine learning techniques for children and adolescents who had COVID-19. Firstly, a notable finding was the few number of studies that utilized machine learning models for predicting various outcomes in children and adolescents. This fact highlights the need for further studies of this nature in the field of pediatrics.

Despite obtaining a low number of studies in this review, the quantity of machine learning models found in the selected studies was diverse. The most commonly used were tree-based models, such as XGBoost, decision trees, and Categorial Boosting (CatBoost) [15–16]. XGBoost is an optimized gradient-boosting algorithm that handles complex datasets and achieves high predictive accuracy. It combines gradient boosting and regularization techniques to produce strong predictive models. XGBoost is widely recognized for its scalability, speed, and effectiveness in various machine learning tasks. In another spectrum of machine learning, neural network models were also utilized in the selected studies. An example of a neural network model is the multilayer perceptron. A multilayer perceptron is an artificial neural network consisting of multiple layers of interconnected neurons [17]. It is commonly used for non-linear regression and classification tasks. The network utilizes forward propagation to process input data and backpropagation to adjust the weights and biases during the training process.

Our findings suggest that machine learning techniques can potentially develop accurate predictive models across various clinical fields. For instance, several studies demonstrated high accuracy rates for predicting outcomes, including disease diagnosis or prognosis. These models could be used to improve patient care by identifying high-risk

Table 7
Machine learning aspects in the included studies.

Key items	Total (n = 10)
	n (%) [95 % CI]
Data preparation[a]	
Cleaning	2 (20) [0.057–0.51]
Aggregation	1 (10) [0.018–0.40]
Augmentation	1 (10) [0.018–0.40]
Encoding	2 (20) [0.057–0.51]
Normalization	1 (10) [0.018–0.40]
Other	2 (20) [0.057–0.51]
Not reported	6 (60) [0.31–0.83]
Data splitting	
Train-test set	6 (60) [0.31–0.83]
Train-validation-test set	4 (40) [0.17–0.69]
Dimensionality reduction techniques	
Principal component analysis	1 (10) [0.018–0.40]
Not Reported	8 (80) [0.49–0.940]
Not applicable	1 (10) [0.018–0.40]
Class Imbalance	
SMOTE	1 (10) [0.018–0.40]
Not Reported	8 (80) [0.49–0.94]
Not applicable	1 (10) [0.018–0.40]
Strategy for hyperparameter optimization[a]	
Cross-validation	4 (40) [0.17–0.69]
Manual search	1 (10) [0.018–0.40]
Predefined values/default	1 (10) [0.018–0.40]
Done automatically by CatBoost	1 (10) [0.018–0.40]
Not Reported	7 (70) [0.40–0.89]

[a] Counts are absolute numbers with column percentages in parentheses. The percentages sometimes do not add up to 100 % because studies reported more than one option. SMOTE = Synthetic Minority Oversampling TEchnique.

Table 8
Model building of all included studies.

Key items	Total (n = 33)
	n (%) [95 % CI]
Selection of predictors[a]	
Impurity Based Feature Importance	1 (3.03) [0.054–0.15]
TF-IDF Embedding	1 (3.03) [0.054–0.15]
Frequency Encoding/Count Encoding	1 (3.03) [0.054–0.15]
Spearman Correlation	4 (12.12) [0.048–0.27]
All predictors	2 (6.06) [0.017–0.20]
Decided by pediatricians	1 (3.03) [0.054–0.15]
Propensity Score	1 (3.03) [0.054–0.15]
Embedded in learning process	1 (3.03) [0.054–0.15]
Unclear	1 (3.03) [0.054–0.15]
Hyperparameter tunning reported	
Yes	2 (6.06) [0.017–0.20]
No	7 (21.21) [0.11–0.38]
Unclear	1 (3.03) [0.054–0.15]
Variable importance reported [a]	
Shapley Value	2 (6.06) [0.017–0.20]
By Random Forest	2 (6.06) [0.017–0.20]
Weights/correlation	4 (12.12) [0.048–0.27]
Gain information	1 (3.03) [0.054–0.15]
None	3 (9.09) [0.031–0.24]
Penalization methods used	
Not reported	10 (30.3) [0.17–0.47]

Abbreviations: TF-IDF, term frequency-inverse document frequency.

[a] Counts are absolute numbers with column percentages in parentheses. The percentages sometimes do not add up to 100 % because studies developed more than one model.

Table 9
Performance measures reported by included studies.

Key items	All extracted models (n = 33)
	n (%) [95 % CI]
Calibration[a]	
Calibration plot	4 (12.12) [0.048–0.27]
Calibration slope	1 (3.03) [0.0054–0.15]
Calibration in the large	1 (3.03) [0.0054–0.15]
None	5 (15.15) [0.067–0.31]
Discrimination	
AUC/AUC-ROC	15 (45.45) [0.30–0.62]
AUPRC	8 (24.24) [0.13–0.41]
Min(Re,Pr)	3 (9.09) [0.031–0.24]
C-statistic	1 (3.03) [0.0054–0.15]
None	1 (3.03) [0.0054–0.15]
Classification	
Sensitivity	12 (36.36) [0.22–0.53]
Specificity	12 (36.36) [0.22–0.53]
Recall	9 (27.27) [0.15–0.44]
Precision	8 (24.24) [0.13–0.41]
Overall [a]	
Predictive values	1 (3.03) [0.0054–0.15]
AUC difference	2 (6.06) [0.017–0.2]
Accuracy	25 (75.75) [0.59–0.87]
F1-score	12 (36.36) [0.22–0.53]
Youden Index	1 (3.03) [0.0054–0.15]

Abbreviations: AUC/ROC, Area Under the Receiver Operating Characteristic Curve, AUPRC, Area Under the Precision-Recall Curve, Min (Re, Pr), Minimum value between Recall and Precision.

[a] Counts are absolute numbers with column percentages in parentheses. The percentages sometimes do not add up to 100 % because studies developed more than one model.

Table 10
Predictive performance of all extracted models.[a]

Key items	All extracted models (n = 33)	
	Reported, n (%)	Apparent performance
		Median [IQR], range
Calibration		
Slope	2 (6.06)	Not significant
Calibration-in-the large	1 (3.03)	Not significant
Pearson chi-square	1 (3.03)	Not significant
Discrimination		
AUC	15 (45.45)	0.98 [0.84–0.98], 0.68 to 0.98
AUPRC	3 (9.09)	Not significant
AUROC	3 (9.09)	Not significant
Accuracy	22 (66.66)	0.81 [0.8–0.92], 0.79 to 0.96
F-Measure	11 (33.33)	0.84 [0.84–0.92], 0. 45 to 0.92
Min(Re, Pr)	3 (9.09)	Not significant
Sensitivity	18 (54.54)	0.90 [0.69–0.93], 0.69 to 0.94
Specificity	18 (54.54)	0.89 [0.87–0.94], 0.87 to 0.99
Precision	10 (30.3)	0.83 [0.83–0.93], 0.77 to 0.99
Recall	7 (21.21)	0 [0.85–0.85], 0.82 to 0.92

Abbreviations: AUC/ROC, Area Under the Receiver Operating Characteristic Curve, AUPRC, Area Under the Precision-Recall Curve, Min (Re, Pr), Minimum value between Recall and Precision.

[a] Counts are absolute numbers with column percentages in parentheses. The percentages sometimes do not add up to 100 % because studies reported more than one option.

individuals who may benefit from early interventions or personalized treatment plans.

Despite the promising results of some studies, we found a lack of consistency in reporting model development and validation procedures across the selected articles. For instance, some studies did not provide detailed information about data sources or model construction methods. This lack of transparency can hinder reproducibility and limit the generalizability of the models to other populations or settings.

Concerning data sources, there are several biases in datasets used for machine learning models. Bias is a statistical term that can occur when a model fails to provide an accurate representation of the population. Some biases present in the datasets include the following:

- **Selection Bias**: This bias arises when data from a specific part of the population is used, not representing the entire target population of the study. To mitigate this bias, it is essential to audit the dataset, ensuring the sample accurately represents the study's target population.

- **Overgeneralization Bias**: Researchers encounter this bias when assuming that observations in their dataset mirror those in any dataset aimed at assessing the same problem. To address this issue, external validation is crucial for evaluating model performance.
- **Automation Bias**: This bias occurs when researchers heavily rely on automation tools for data processing before model training. Complete trust in these tools is discouraged; it's vital to verify correct data transformation outcomes.
- **Sampling Bias**: This bias occurs when sampling techniques are not used to balance classes within the dataset. This may lead to models with high accuracy in classifying the most represented class in the dataset.

Still regarding the biases in the data, the most common inconsistency observed in the identified articles pertains to the failure to share modified training data. Researchers should elucidate the state of data post-modifications made for training, including the removal of erroneous features, handling of features with substantial missing data, categorical variable encoding, data sampling, and other relevant procedures. Mere mention of using data from a specific website is insufficient. Without this crucial information, the assessment of the actual data employed in the studies becomes challenging. This can lead to indications of biases in the models and render them less interpretable. Studies that do not disclose their data and source code make the research less transparent.

To address these issues, future studies should follow established guidelines for developing and reporting predictive models (e.g., TRIPOD statement) [18]. Additionally, researchers should consider external validation of their models to assess their performance in independent datasets [19].

Another consideration is the ethical implications of using machine learning models in clinical practice. For instance, there is a risk of perpetuating bias or discrimination if the models are trained on biased data or are not validated across diverse populations [20]. Therefore, ensuring that these models are developed and used ethically and responsibly is crucial.

Another finding from our review is that most of the selected studies used administrative databases as their primary data source. This suggests that machine learning techniques may be particularly useful for analyzing large-scale administrative datasets to identify patterns and predict outcomes.

Machine learning models have a potential impact on clinical decision-making. These models have shown promise for improving patient outcomes by identifying high-risk individuals or predicting disease progression. However, the models should not be viewed as a replacement for clinical judgment or human expertise [21]. Instead, they should be used as a tool to support clinical decision-making and improve patient care.

There is a deficiency how the selected studies reported data in the models. The limitations include inadequate reporting of sample sizes, missing information about hyperparameter tuning, lack of implementation details, and performance measures of the models. These issues are essential for reproducibility purposes [22].

Few studies employed cross-validation techniques in model development. Cross-validation helps to prevent the phenomenon of overfitting [23], where the model achieves 100 % accuracy on the test data, which represents the model's development data that has not been seen by the model before. However, if the test data happens to be identical to the training data, it is necessary to train and test the model using different folds of the data. Cross-validation divides the model development data into multiple folds, using each fold as both training and testing data. The lack of cross-validation can lead to inaccurate information regarding the performance of the models.

The most commonly used method for predictor selection in the selected studies was the Spearman correlation. Few studies discussed techniques for dimensionality reduction of predictors, although most studies had a low number of features for model development. The selected studies did not provide clear information about missing data and how they handled it. Many methodological details in the majority of studies were unclear. Several studies did not make their code or data available in separate repositories for other researchers to read and reproduce the analysis. Many studies did not report information regarding the calibration and discrimination of the models. It is important to report data about the calibration and discrimination of a machine learning model because these metrics provide insights into the model's performance and reliability. Calibration measures the agreement between the predicted probabilities and the observed outcomes, indicating whether the model's predictions are well-calibrated and accurate. Conversely, discrimination assesses the model's ability to distinguish between different outcomes or classes, indicating its predictive power. Reporting these metrics allows researchers and practitioners to evaluate the model's effectiveness, identify potential biases or limitations, and compare its performance against other models or benchmarks. Ultimately, it promotes transparency, reproducibility, and informed decision-making in utilizing machine learning models.

The studies did not provide a solid contribution to the medical community as they did not create any website or other means for physicians and other interested parties to test the model. There is a need for closer collaboration between this emerging field of evidence-based medicine and practicing clinicians. The availability of models is crucial for other physicians to provide feedback on the performance of the models developed for data specific to their regions.

No selected study provided information on the prediction horizon of the models. This type of information can be necessary for the clinical field to understand the predictions' validity.

The lack of external validation to effectively test the selected models with unseen data is worth noting. However, obtaining external validation data can be challenging, and testing models with multiple sources requires time and effort to acquire and organize large databases for evaluation by machine learning models.

The nature of the data was not widely discussed in the majority of articles. As important as the model itself, the quality and preparation of the data used for training greatly influence the model's performance. If the data is not properly prepared before the training, biases may be introduced, affecting the model's true performance. Few studies mentioned how the data were treated in terms of their nature (continuous, discrete, and others) and how the data were encoded for evaluation by the developed models.

4.2. How models were externally validated

In the three studies that use external validation, the procedure has been conducted to assess the model's real-world applicability [24–26]. The studies conducted external validation, adapting to their specific dataset characteristics. For models with small sample sizes, the researchers in the first study employed data splitting, allocating a portion of the dataset for training and another for validation. Additionally, they acquired external data from independent sources to further validate the model's performance. Key performance metrics, such as accuracy and precision, were calculated and compared between the internal and external datasets, ensuring a comprehensive assessment of results generalization. In the second and third study, addressing models with large sample sizes, the authors adopted a similar approach, splitting their dataset into training and validation subsets. They emphasized the importance of external validation, even with large data, by obtaining an independent and unseen dataset. Performance metrics were evaluated in both internal and external validation datasets. Data splitting was complemented by techniques such as k-fold cross-validation to maximize data utilization. Since these three studies reported good metric values in the tests with external validation datasets, this can exemplify the importance of external validation in machine learning research, contributing to the transparency and real-world applicability of their findings.

4.3. *Traditional statistical models versus machine learning models*

Traditional statistics has greater transparency and interpretability of relationships between different variables in the data, clearly showing insights between dependent and independent variables. On the other hand, machine learning models can learn different relationships between data that were not detected by traditional statistical models, but this is not the focus. Until recently, developers paid little attention to the explainability of machine learning models. The models were seen as black boxes. This scenario has changed, and today's models are more explicit about their results. However, the aim of machine learning models is different from traditional statistical models. The aim of these models is that, by using a set of data that the model has never seen, it is able to classify the data correctly or predict something accurately. The machine learning models are oriented towards the result and the final performance of the prediction.

For example, when a machine learning model is used to diagnose a disease, the doctor must enter the patient's data into the system and it will tell whether the patient is likely to have the disease, showing which variables contributed most to that outcome. These models often see different relationships between the data compared to traditional statistics, as the focus is on providing an answer with a higher degree of accuracy for the task proposed to the model. For the same set of data, machine learning models often find different relationships between the data than statistical models. This is because the variables that allow the model to give the correct answers are different.

The comparison between machine learning models and traditional statistical models regarding performance and utility has the following strengths and limitations [27–29].

4.4. *Strengths and limitations of traditional statistical models*

- Statistical models are designed to infer relationships between variables. They are used to identify the underlying patterns and relationships in the data and establish both the scale and significance of the relationship.
- Statistical models explicitly specify a probabilistic model for the data and identify variables that are usually interpretable and of special interest, such as effects of predictor variables.
- Statistical models are best suited for small to medium-sized datasets.
- Statistical models require many assumptions to identify the underlying relationships between variables.
- Statistical models presuppose that the input variables are not highly associated with one another and do not exhibit multicollinearity.
- Certain statistical models rely on a sufficiently large sample size to guarantee precise parameter estimates.

4.5. *Strengths and limitations of machine learning models*

- Machine learning models are designed to make the most accurate predictions possible. They are built to provide accurate predictions without explicit programming.
- Machine learning models can provide better predictions than statistical models.
- Machine learning models are more empirical and do not impose relationships between predictors and outcomes, or isolate the effects of any single variable.
- Machine learning models are best suited for large datasets.
  Machine learning models are more difficult to understand and explain than statistical models.
- Machine learning models do not provide the level of interpretability that is possible using statistical models.

The choice between machine learning models and traditional statistical models depends on the purpose of the analysis. If the goal is to determine and explain the relationships between variables, statistical models are the best approach. If the goal is to make accurate predictions, machine learning models are the most adequate option.

5. Comparison to previous studies

To the best of our knowledge, at the present moment of writing the results of this study, we did not find another study that has conducted a systematic review to identify the methodological conduct and study design of research utilizing prediction models for outcomes in children and adolescents using machine learning algorithms. However, studies evaluating machine-learning models for adult patients have identified similar methodological conduct and reporting issues in various reviews exploring different machine-learning techniques [30–32]. Neglected aspects such as missing data, sample size, calibration, and model availability have been consistently observed [31,32–34]. In a review examining the trends of prediction models utilizing electronic health records (EHR), it was noted that the utilization of ensemble models increased from 6% to 19% [35]. Another comprehensive review focusing on prediction models for hospital readmission revealed a substantial growth in the application of algorithms, including Support Vector Machine (SVM), Random Forest (RF), and Neural Networks (NN), with an increase from none to 38% over the past 5 years [36]. Additionally, the adoption of methods to address a class imbalance in EHR datasets increased from 7% to 13% [35].

6. Limitations of this study

The information extracted in our study was solely based on the content reported in the articles. Regrettably, only a few articles provided the essential information required by reporting guidelines, making the data extraction process challenging [37]. Additionally, there was inconsistency in the terminology used across papers. For instance, the term "validation" was frequently used to describe both tuning and testing (i.e., internal validation), a concern previously identified in a review of studies on deep learning models [38]. This fact highlights the necessity of uniform terminology for critically evaluating machine learning models [39].

In our study, we encountered limitations that prevented us from conducting a meta-analysis. The scarcity of studies refers to the limited number of relevant studies available, which may arise due to the novelty of the research area, ethical considerations, or limited research resources. Additionally, the heterogeneity among studies, in terms of study design, population characteristics, interventions, or outcome measures, and the variation in methodologies and findings across studies may introduce substantial clinical and methodological heterogeneity, making it inappropriate to combine the results quantitatively.

Our data extraction form was primarily drawn based on the items and signaling questions from the TRIPOD and PROBAST tools. Although these tools were initially developed for regression-based prediction models, most items and signaling questions were still applicable to studies on machine learning-based models.

7. Implications for future research

The extent to which the selected studies aimed to improve clinical care with the developed models or primarily sought to showcase promising results with the proposed models is questionable. There was a limited emphasis on aspects including the study's objective, clinical workflow, outcome format, prediction horizon, and clinically relevant performance metrics. Guidelines and meta-epidemiological studies have strongly emphasized the importance of applying optimal methodology and transparent reporting in prediction model studies [40,43]. The TRIPOD and PROBAST provide best practice recommendations for designing, conducting, and reporting prediction models, regardless of the modeling technique employed [13,18,40,41]. However, extending these recommendations to include areas such as data preparation, tunability, fairness, and data leakage is crucial.

Extensions of PROBAST and TRIPOD specifically designed for artificial intelligence (AI) or machine learning-based prediction models, namely PROBAST-AI and TRIPOD-AI, are currently being developed [39,42]. As machine learning continues to gain importance in healthcare, it is highly recommended for future studies to reinforce the adoption of a minimum standard in methodological conduct and reporting to increase the generalizability and applicability of these models [13,18,40,41].

Another notable aspect in the selected studies of this review concerns the explainability of the developed models. Only three of the nine studies addressed the explainability of the models using the Shapley Additive exPlanations (SHAP) method. The explainability of machine learning models in the healthcare domain is crucial as this research field deals with various interested parties who demand fair, unbiased, reliable, and interpretable learning models rather than black-box machine learning models. The findings of our study align with recent research highlighting that most machine learning models developed in healthcare did not employ explainable artificial intelligence (XAI) methods to elucidate the predictions made by the models [44,45]. This is an issue that requires the attention of researchers.

Furthermore, the limited accessibility of the developed models poses a barrier to conducting independent validation, a crucial step before their integration into clinical practice. Openly sharing the source code and, ultimately, the clinical prediction model itself is a fundamental measure to establish trust and credibility in applying AI and machine learning in the clinical setting [46].

8. Conclusion

Our study highlights important considerations when developing and using machine learning models in healthcare settings. Future research should address limitations, including small sample sizes, inconsistent reporting practices, biases in data sources, and ethical implications, to ensure that these models are developed and used responsibly to improve patient care.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.artmed.2024.102824.

Declaration of competing interest

The funding organizations had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript.

Acknowledgments

The authors would like to thank State Research Foundation- FAPEMIG, Minas Gerais, Brazil, the National Council for Scientific and Technological Development - CNPq, Brazil, and the Coordination for the Improvement of Higher Education Personnel, CAPES, Brazil.

References

[1] Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019; 380(14):1347–58. https://doi.org/10.1056/NEJMra1814259.

[2] Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. N Engl J Med 2016;375(13):1216–9. https://doi.org/10.1056/NEJMp1606181.

[3] Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018; 319(13):1317–8. https://doi.org/10.1001/jama.2017.18391.

[4] Wyatt JC. Clinical data systems: overcoming the barriers to their development. JAMIA 1996;3(6):408–12. https://doi.org/10.1136/jamia.1996.97046762.

[5] Moons KGM, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart 2012;98(9):691–8. https://doi.org/10.1136/heartjnl-2011-301247.

[6] Van Smeden M, Reitsma JB, Riley RD, Collins GS, Moons KG. Clinical prediction models: diagnosis versus prognosis. J Clin Epidemiol 2021;132:142–5. https://doi.org/10.1016/j.jclinepi.2021.01.009.

[7] Oliveira Eduardo A, Oliveira Maria Christina L, Silva Ana Cristina Simões e, Colosimo Enrico A, Mak Robert H, Vasconcelos Mariana A, et al. Clinical outcomes of omicron variant (B.1.1.529) infection in children and adolescents hospitalized with COVID-19 in Brazil with observational data on the efficacy of the vaccines in adolescents. Pediatr Infect Dis J March 2023;42(3):218–25. https://doi.org/10.1097/INF.0000000000003783.

[8] Comparison of the first and second waves of the Coronavirus disease 2019 pandemic in children and adolescents in a middle-income country: clinical impact associated with severe acute respiratory syndrome coronavirus 2 gamma lineage Oliveira, Eduardo A. et al J Pediatr, Volume 244, 178 - 185.(e3).

[9] Vasconcelos MA, Mendonça ACQ, Colosimo EA, et al. Outcomes and risk factors for death among hospitalized children and adolescents with kidney diseases and COVID-19: an analysis of a nationwide database. Pediatr Nephrol 2023;38:181–91. https://doi.org/10.1007/s00467-022-05588-0.

[10] Andaur Navarro CL, Damen JA, van Smeden M, Takada T, Nijman SW, Dhiman P, et al. Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models. J Clin Epidemiol 2023;154: 8–22.

[11] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021:372.

[12] Tawfik GM, Dila KAS, Mohamed MYF, et al. A step-by-step guide for conducting a systematic review and meta-analysis with simulation data. Trop Med Health 2019; 47:46. https://doi.org/10.1186/s41182-019-0165-6.

[13] Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMC Med 2015;13:1. https://doi.org/10.1186/s12916-014-0241-z.

[14] Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019;170(1):51–8. https://doi.org/10.7326/M18-1376.

[15] Darapaneni N, Srinivas P, Reddy KM, Paduri AR, Kanugovi L, J P, et al. Tree Based Models: A Comparative and Explainable Study for Credit Default Classification. 2022 IEEE 9th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON). 2022. p. 1–8.

[16] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016.

[17] Car Z, Segota SB, Andelic N, Lorencin I, Mrzljak V. Modeling the spread of COVID-19 infection using a multilayer perceptron. Comput Math Methods Med 2020;2020.

[18] Moons KGM, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162:W1e73.

[19] Campagner A, Carobene A, Cabitza F. External validation of machine learning models for COVID-19 detection based on complete blood count. Health Inf Sci Syst 2021;9(1):37. https://doi.org/10.1007/s13755-021-00167-3.

[20] Michelson KN, Klugman CM, Kho AN, Gerke S. Ethical considerations related to using machine learning-based prediction of mortality in the pediatric intensive care unit. J Pediatr 2022;247:125–8. https://doi.org/10.1016/j.jpeds.2021.12.069.

[21] Xu Y, Liu X, Cao X, Huang C, Liu E, Qian S, et al. Artificial intelligence: a powerful paradigm for scientific research. Innovation (Cambridge (Mass.)) 2021;2(4): 100179. https://doi.org/10.1016/j.xinn.2021.100179.

[22] Probst Philipp, Boulesteix Anne-Laure, Bischl Bernd. Tunability: importance of hyperparameters of machine learning algorithms. J Mach Learn Res 2019;20(1): 1934–65 (January 2019).

[23] Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, MA: MIT Press; 2016.

[24] Byeon H. Predicting South Korean adolescents vulnerable to obesity after the COVID-19 pandemic using categorical boosting and shapley additive explanation values: a population-based cross-sectional survey. Front Pediatr 2022;10:955339. https://doi.org/10.3389/fped.2022.955339.

[25] Gao J, Yang C, Heintz J, Barrows S, Albers E, Stapel M, et al. MedML: fusing medical knowledge and machine learning models for early pediatric COVID-19 hospitalization and severity prediction. iScience 2022;25(9):104970. https://doi.org/10.1016/j.isci.2022.104970.

[26] Zhang Z, Xiao Q, Luo J. Infant death prediction using machine learning: a population-based retrospective study. Comput Biol Med 2023;165:107423. https://doi.org/10.1016/j.compbiomed.2023.107423.

[27] Rajula HSR, Verlato G, Manchia M, Antonucci N, Fanos V. Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. Medicina (Kaunas) 2020;56(9):455. https://doi.org/10.3390/medicina56090455.

[28] Bennett M, Kleczyk EJ, Hayes K, Mehta R. Evaluating similarities and differences between machine learning and traditional statistical modeling in healthcare analytics. IntechOpen 2022. https://doi.org/10.5772/intechopen.105116.

[29] Ley C, Martin RK, Pareek A, et al. Machine learning and conventional statistics: making sense of the differences. Knee Surg Sports Traumatol Arthrosc 2022;30: 753–7. https://doi.org/10.1007/s00167-022-06896-6.

[30] Dhiman P, Ma J, Andaur Navarro CL, Speich B, Bullock G, Damen JAA, et al. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. BMC Med Res Methodol 2022;22:1e16.

[31] Dhiman P, Ma J, Andaur Navarro C, Speich B, Bullock G, Damen JA, et al. Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. J Clin Epidemiol 2021;138:60e72.

[32] Collins GS, De Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol 2014;14:40.

[33] Damen JAAG, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ 2016;353:i2416.

[34] Bouwmeester W, Zuithoff NPA, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. PLoS Med 2012;9(5):1e12.

[35] Yang C, Kors JA, Ioannou S, John LH, Markus AF, Rekkas A, et al. Trends in the conduct and reporting of clinical prediction model development and validation: a systematic review. J Am Med Inform Assoc 2022;29(5):983e9.

[36] Artetxe A, Beristain A, Graña M. Predictive models for hospital readmission risk: a systematic review of methods. Comput Methods Prog Biomed 2018;164:49e64.

[37] Andaur Navarro CL, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review. BMC Med Res Methodol 2022;22:12.

[38] Kim DW, Jang HY, Ko Y, Ko Y, Son JH, Kim PH, et al. Inconsistency in the use of the term "validation" in studies reporting the performance of deep learning algorithms in providing diagnosis from medical imaging. PLoS One 2020;15:1e10.

[39] Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open 2021;11(7):e048008.

[40] Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019;170:51e8.

[41] Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019;170:W1e33.

[42] Collins GS, Moons M, KG.. Reporting of artificial intelligence prediction models. Lancet 2019;393:1577e9.

[43] Damen JAAG, Debray TPA, Pajouheshnia R, Reitsma JB, Scholten RJPM, Moons KGM, et al. Empirical evidence of the impact of study characteristics on the performance of prediction models: a meta-epidemiological study. BMJ Open 2019; 9(4):1e12.

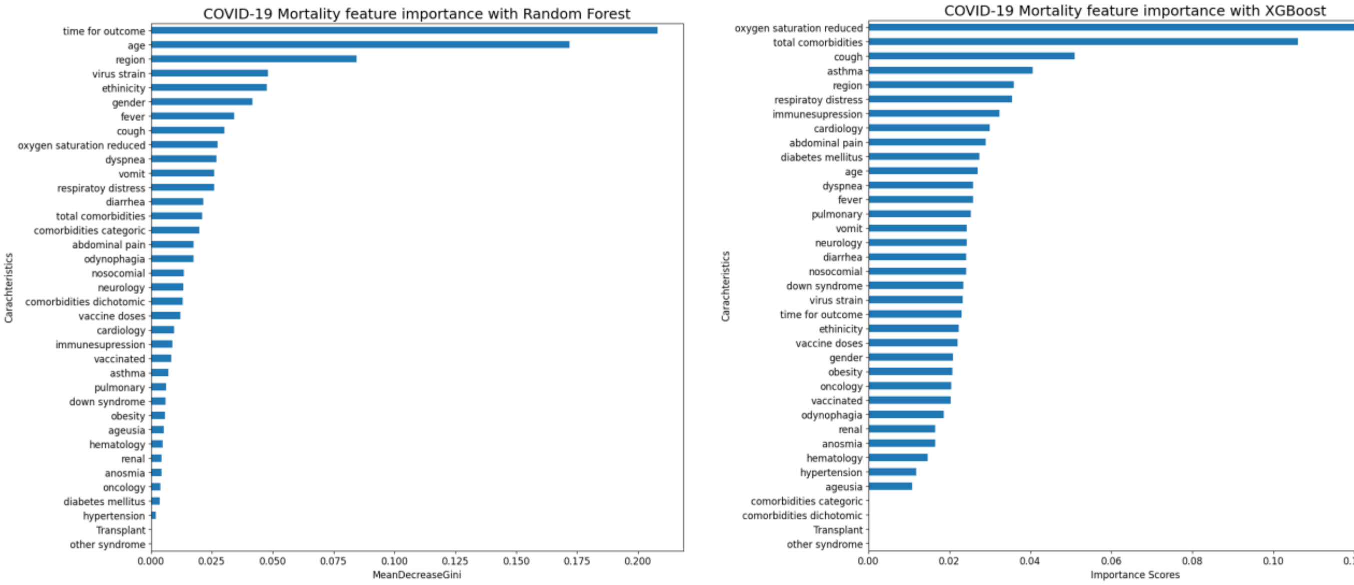
[44] Allgaier Johannes, Mulansky Lena, Draelos Rachel Lea, Pryss Rüdiger. How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare. Artif Intell Med 2023;143:102616. ISSN 0933-3657, https://doi.org/10.1016/j.artmed.2023.102616.

[45] Parimbelli Enea, Buonocore Tommaso Mario, Nicora Giovanna, Michalowski Wojtek, Wilk Szymon, Bellazzi Riccardo. Why did AI get this one wrong? — Tree-based explanations of machine learning model predictions. Artif Intell Med 2023;135:102471 (ISSN 0933-3657), https://doi.org/10.1016/j.artmed.2022.102471.

[46] Van Calster B, Wynants L, Timmerman D, Steyerberg EW, Collins GS. Predictive analytics in health care: how can we know it works? J Am Med Inform Assoc 2019; 26(12):1651e4.

Supplementary material from the original article

Supplementary Figure 1 - The importance scores of the predictors were calculated using random forest (a) and XGBoost (b) tests.



Supplementary Table 1 – Primary features documented in the SIVEP-Kids database.

No.	Feature name	Variable Type	No.	Feature name	Variable type
1	Gender	Nominal	21	Hypertension	Nominal
2	Age	Numeric	22	Immunosuppression	Nominal
3	Ethnicity	Nominal	23	Renal disease	Nominal
4	Region	Nominal	24	Asthma	Nominal
5	Virus strain	Nominal	25	Hematology disease	Nominal
6	Dyspnea	Nominal	26	Neurology	Nominal
7	Fever	Nominal	27	Oncology	Nominal
8	Cough	Nominal	28	Transplanted	Nominal
9	Odynophagia	Nominal	29	Down Syndrome	Nominal
10	Diarrhea	Nominal	30	Other Syndrome	Nominal
11	Vomit	Nominal	31	Nosocomial	Nominal
12	Abdominal pain	Nominal	32	Comorbidities dichotomic	Nominal
13	Ageusia	Nominal	33	Total Comorbidities	Numeric
14	Anosmia	Nominal	34	Number of vaccine doses	Numeric
15	Respiratory distress	Nominal	35	Comorbidities categoric	Nominal
16	Oxygen saturation reduced	Nominal	36	Time for outcome	Numeric
17	Diabetes	Nominal	37	Vaccinated	Nominal
18	Obesity	Nominal	38	Outcome (Target Variable)	Nominal
19	Cardiology	Nominal			
20	Pulmonary	Nominal			

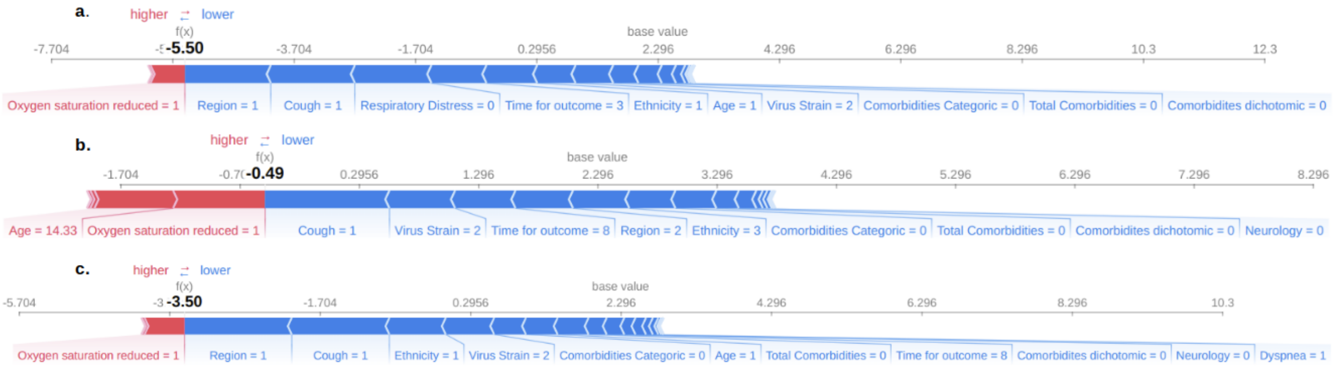
Supplementary Table 2 - The Hyperparameters of the selected ML algorithms for COVID-19 mortality prediction in children and adolescents.

ML Algorithms	Hyperparameters used to create the models
GBC	criterion='friedman_mse', learning_rate=0.0005, max_depth=9, max_features='log2', min_impurity_decrease=0.001, min_samples_leaf=1, min_samples_split=9, n_estimators=120, subsample=0.9, tol=0.0001, validation_fraction=0.1.
ADA	algorithm='SAMME', learning_rate=0.005, n_estimators=260.
CATBOOST	Iterations=1000, learning_rate=0.1, depth=6, l2_leaf_reg=3.0, subsample=0.8, colsample_bylevel=0.8, border_count=128, loss='log_loss'.
RF	criterion='gini', max_depth=4, max_features=1.0, max_leaf_nodes=None, min_impurity_decrease=0.3, min_samples_leaf=2, min_samples_split=7, n_estimators=90.
XGBOOST	booster='gbtree', colsample_bytree=1, learning_rate=0.4, max_depth=1, min_child_weight=2, n_estimators=120, objective='binary:logistic'
ET	criterion='gini', max_depth=4, max_features=1.0, min_impurity_decrease=0.3, min_samples_leaf=2, min_samples_split=7, n_estimators=90.
LR	C=0.662, fit_intercept=True, intercept_scaling=1, l1_ratio=None, max_iter=1000, penalty='l2', solver='lbfgs', tol=0.0001.
LDA	shrinkage=0.4, solver='lsqr', tol=0.0001.
DT	criterion='entropy', max_depth=4, max_features=1.0, min_impurity_decrease=0.5, min_samples_leaf=3, min_samples_split=2, splitter='best'.
NB	var_smoothing=1
KNN	leaf_size=30, metric='manhattan', n_neighbors=50, p=2, weights='distance'.
QDA	reg_param=0.29, tol=0.0001.

Supplementary Table 3 - Descriptive statistics of the most important variables selected in the feature selection phase for mortality in COVID-19 children and adolescents' patients.

Nº	Feature name	Variable Type	Frequency or mean ± SD
1	Age	Numeric	5.04 ± 5.25
2	Region	Nominal	Southeast (10,819) South (4,379) Northeast (4,033) North (2,609) Central West (2,257)
3	Ethnicity	Nominal	Asian (178) Black (778) Brown (11,467) Indigenous (221) White (11,453)
4	Virus Strain	Nominal	Omicron (13,432) Gamma (8,251) Delta (2,414)
5	Dyspnea	Nominal	Haven't (11,126) Have (12,971)
6	Cough	Nominal	Haven't (7,198) Have (16,899)
7	Respiratory distress	Nominal	Haven't (11245) Have (12852)
8	Oxygen saturation reduced at admission	Nominal	Haven't (12018) Have (12079)
9	Obesity	Nominal	Haven't (23675) Have (422)
10	Cardiology	Nominal	Haven't (23314) Have (783)
11	Pulmonary	Nominal	Haven't (23608) Have (489)
12	Hypertension	Nominal	Haven't (24029) Have (68)
13	Immunosuppression	Nominal	Haven't (23580) Have (517)
14	Renal	Nominal	Haven't (23876) Have (221)
15	Asthma	Nominal	Haven't (22711) Have (1386)

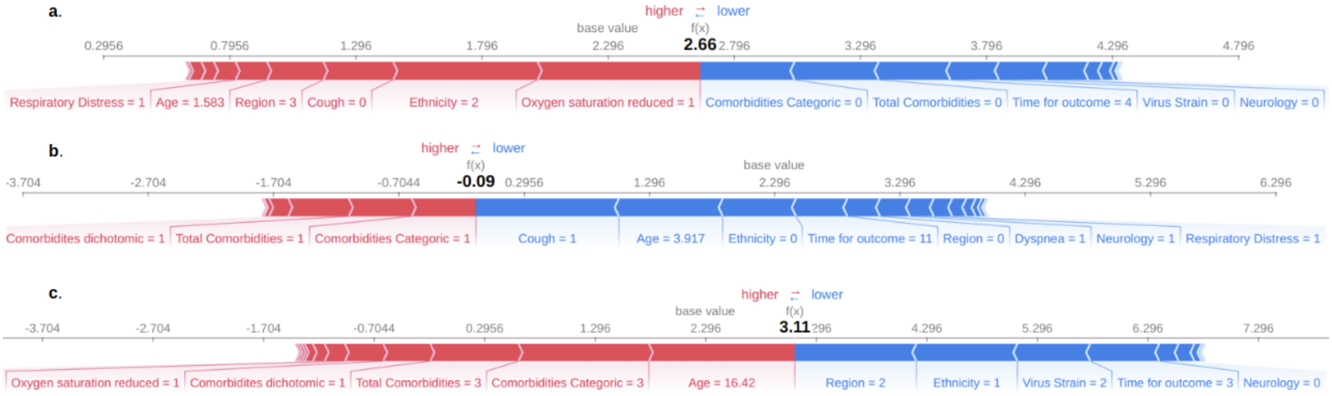
16	Total Comorbidities	Numeric	0.22 ± 0.54 (0, 19670) (1, 3512) (2, 757) (3, 128) (4, 24) (5, 2) (6, 2) (7, 1) (10, 1)
17	Comorbidities dichotomic	Nominal	Haven't (19670) Have (4427)
18	Comorbidities categoric	Nominal	Haven't (19670) One (3512) Two (757) Three or more (158)
19	Time for Outcome	Numeric	7.63 ± 6.83
20	Hematology	Nominal	Haven't (23649) Have (448)
21	Neurology	Nominal	Haven't (22397) Have (1700)
22	Oncology	Nominal	Haven't (24048) Have (49)
23	Down Syndrome	Nominal	Haven't (23681) Have (416)
24	Nosocomial	Nominal	Haven’t (23476) Have (621)



Supplementary Figure 2 – Force plot of feature contributions to the decision-making process of the model for discharge outcome.

In a force plot for SHAP values, as can be seen in Supplementary Figure 2 (showing specific cases where the model made medical discharge predictions) and Figure 5 (showing specific cases where the model made death predictions), the goal is to illustrate the contributions of individual features to a specific model prediction. Each feature is represented by a horizontal bar, and the length of the bar corresponds to the magnitude of the SHAP value for that feature. The features are arranged horizontally based on their importance to the prediction. We show in each figure three examples of force plots for each class of model prediction (death or discharge).

The plots in Supplementary Figure 2a illustrate the forces for three patients who were discharged from the dataset. Each plot corresponds to an individual patient, and it is evident that reduced oxygen saturation is the most influential feature in the model's decision-making process. However, the force associated with this feature is lower compared to other values, resulting in the classification of the patient into class 0 (medical discharge). It is noteworthy that the region representing the southern region of Brazil had the highest force value, influencing the model's decision to classify the patient as discharged. The absence of respiratory distress, short hospitalization duration, white ethnicity, and the absence of comorbidities were also important factors in the model's decision to classify these patients as discharged. A comparison of Supplementary Figure 2a, 2b, and 2c reveals that lower age values tend to lead the model to classify the patient as discharged, while higher age values, as seen in Supplementary Figure 2b, contribute to the model classifying the patient as deceased. Additionally, variables associated with comorbidities with lower values contribute to the model classifying the patient as discharged.



Suplementary Figure 3 – Force plot of feature contributions to the decision-making process of the model for discharge outcome.

In Supplementary Figure 3a, we observe that a patient presenting reduced oxygen saturation was the primary feature contributing to the model classifying the patient into class one (deceased). Additionally, the model considered an ethnicity value of 2, indicating indigenous, as a variable contributing to an increased prediction value for the higher class, which is deceased. Furthermore, factors such as the northern region of Brazil and reports of respiratory distress contribute to the model classifying the patient as deceased. In Supplementary Figure 3b, the presence of comorbidities emerges as an important factor for the model to classify as deceased. However, solely relying on the presence of comorbidities in patients does not serve as a strong predictor for the model, as the final value of the model's decision function was negative, close to zero. Nevertheless, the model correctly classified this instance as deceased. In Supplementary Figure 3c, besides variables crucial for decision-making regarding mortality, we observe that higher age leads the model to predict the patient as deceased, in contrast to what was shown in Supplementary Figure 2.

Other studies in the literature that evaluated ML algorithms for predicting deaths in children and adolescents.

Other studies evaluated ML models for predicting the deaths of children and adolescents from COVID-19. The study conducted by Zhang et al [1]. utilized ML techniques to predict infant mortality rates in the United States based on factors related to birth facility, prenatal care, labor and delivery, and newborn characteristics. The analysis was performed on data from 2016 to 2021, including 116,309 infant deaths among 22,669,736 live births. Among the five ML models compared, XGBoost demonstrated the best predictive performance, achieving an AUC of 93% and an Average Precision (AP) score of 0.55. The study highlighted the significance of utilizing the original imbalanced dataset over balanced datasets created through oversampling techniques, as the former yielded superior predictive outcomes. The validation of the predictive model on data from 2020 to 2021 maintained the performance level, with an AUC of 93% and an AP value of 0.52. The performance of the model during both pre-pandemic (2016–2019) and pandemic periods (2020–2021) shows potential utility in informing strategies to mitigate infant mortality rates.

In the study conducted by Byeon et al. [2], a population-based cross-sectional survey was employed to investigate the impact of the COVID-19 pandemic on the prevalence of obesity among South Korean adolescents. The research utilized categorical boosting, specifically the CatBoost algorithm, to develop a predictive model for adolescent obesity. The model's performance was evaluated using various metrics, and the results indicated that the model achieved an AUC of 68%, with a general accuracy of 82%. The data used in the study encompassed a range of factors including exercise, academic performance, and lifestyle habits, which were analyzed to identify potential risk factors for adolescent obesity. The utilization of the CatBoost algorithm, in conjunction with the evaluation of various performance metrics, underscores the rigorous approach taken to predict vulnerability to obesity in South Korean adolescents post-pandemic.

Gao et al. [3] presents a hybrid approach that combines domain knowledge-based features with data-driven methods to predict pediatric COVID-19 hospitalization and severity. The authors split two cohorts into training, validation, and testing sets by 6:1:3 and used the training set to fit the models, the validation set to determine the hyper-parameters, and the testing set to evaluate the models. The evaluation metrics were AUROC, AUPRC, and Min (Re, Pr). The best model, MedML, achieved a 3% higher AUROC and 4% higher AUPRC on the hospitalization prediction task and a 7% higher AUROC and 14% higher AUPRC on the severity prediction task compared to the best baseline model. The authors used the N3C Data Enclave with Code Workbook and the mini-batch gradient descent to train the models and the batch size was set to 128. The results showed that MedML is generalizable in all nine national geographical regions of the United States and temporally across all consecutive pandemic stages. The authors state that MedML serves as a bridge between clinicians, data engineers, and computer scientists to augment the clinical decision-making process through intuitive knowledge representation, explainable construction, and powerful computation.

Pavliuk et al. [4] developed a ML model for analyzing and predicting the hospitalization numbers of children in the Lviv region during the fourth wave of the COVID-19 pandemic, characterized by the Omicron strain's dominance. The surge in hospitalizations, especially among children, is attributed to their high sociability and low vaccination rates in Ukraine. Utilizing publicly available data, the ML model comprises analysis and prediction components. Pearson correlation coefficient was employed for analyzing hospitalized children's numbers, while short and medium-term predictions utilized neural networks.

The study of Mamlook et al. [5] focuses on evaluating and comparing five well-known ML approaches, including artificial neural network (ANN), random forest (RF), support vector machines (SVM), decision trees (DT), and gradient boosted trees (GBM), to detect COVID-19 in children. The classification performance of each model was assessed using a standard 10-fold cross-validation procedure. The findings reveal that the classification model based on decision trees (CART) outperforms others, achieving 92.5% accuracy for binary classes (positive vs. negative) based on laboratory findings. Important predictors such as Leukocytes, Monocytes, Potassium, and Eosinophils were identified, suggesting their crucial role in COVID-19 detection. The proposed model offers a tool for medical experts to predict COVID-19 in children and validate primary laboratory findings, showcasing the potential of ML methods in facilitating accurate predictions for COVID-19 laboratory outcomes in pediatric cases.

Ma et al. [6] investigate whether clinical symptoms and laboratory results can serve as predictors for the necessity of CT (Computed Tomography) scans in pediatric patients with positive RT-PCR results. Data from 244 pediatric patients were collected, and advanced decision tree-based ML models were employed. The study revealed that age, lymphocyte count, neutrophils, ferritin, and C-reactive protein are crucial indicators for predicting CT outcomes. The developed decision support system demonstrated promising performance, achieving an AUC of 84% with accuracy of 82% and sensitivity of 84%. These findings suggest a reconsideration of CT use in pediatric patients, highlighting the potential non-indispensability of this imaging modality.

Nugawela et al. [7] developed a predictive model for identifying children and young people at a higher risk of experiencing long COVID, defined as having at least one impairing symptom three months after SARS-CoV-2 positive RT-PCR testing. The research utilized data from a nationally matched cohort of SARS-CoV-2 test-positive and test-negative patients aged 11 to 17 years. Predictors considered included SARS-CoV-2 status, demographic factors, quality of life/functioning, physical and mental health, loneliness, and the number of symptoms at testing. The logistic regression model demonstrated an accuracy of 83%, achieving good calibration and discrimination measures.

References

1 - Zhang Z, Xiao Q, Luo J. Infant death prediction using machine learning: A population-based retrospective study. Comput Biol Med. 2023 Oct; 165:107423. doi: 10.1016/j.compbiomed.2023.107423. Epub 2023 Sep 1. PMID: 37672926.

2 - Byeon H. Predicting South Korean adolescents vulnerable to obesity after the COVID-19 pandemic using categorical boosting and shapley additive explanation values: A population-based cross-sectional survey. Front Pediatr. 2022 Sep 21; 10:955339. doi: 10.3389/fped.2022.955339. PMID: 36210956; PMCID: PMC9532523.

3 - Gao J, Yang C, Heintz J, Barrows S, Albers E, Stapel M, Warfield S, Cross A, Sun J; N3C consortium. MedML: Fusing medical knowledge and machine learning models for early pediatric COVID-19 hospitalization and severity prediction. iScience. 2022 Sep 16;25(9):104970. doi: 10.1016/j.isci.2022.104970. Epub 2022 Aug 17. PMID: 35992304; PMCID: PMC9384332.

4 - Pavliuk O, Kolesnyk H. Machine-learning method for analyzing and predicting the number of hospitalizations of children during the fourth wave of the COVID-19 pandemic in the Lviv region. J Reliab Intell Environ. 2023;9(1):17-26. doi: 10.1007/s40860-022-00188-z. Epub 2022 Sep 1. PMID: 36065343; PMCID: PMC9434091.

5 - Mamlook, R. A., Al-Mawee, W., Alden, A. Y. Q., Alsheakh, H., & Bzizi, H. (2021). Evaluation of Machine Learning Models to Forecast COVID-19 Relying on Laboratory Outcomes Characteristics in Children. IOP Conference Series: Materials Science and Engineering, 1094(1), 012072. https://doi.org/10.1088/1757-899X/1094/1/012072

6 - Ma H, Ye Q, Ding W, Jiang Y, Wang M, Niu Z, Zhou X, Gao Y, Wang C, Menpes-Smith W, Fang EF, Shao J, Xia J, Yang G. Can Clinical Symptoms and Laboratory Results Predict CT Abnormality? Initial Findings Using Novel Machine Learning Techniques in Children With COVID-19 Infections. Front Med (Lausanne). 2021 Jun 14; 8:699984. doi: 10.3389/fmed.2021.699984. PMID: 34195215; PMCID: PMC8236538.

7 - Nugawela, M.D., Stephenson, T., Shafran, R. et al. Predictive model for long COVID in children 3 months after a SARS-CoV-2 PCR test. BMC Med 20, 465 (2022). https://doi.org/10.1186/s12916-022-02664-y