Arquivos Catarinenses de Medicina
ISSN (Impresso) 0004-2773
ISSN (online) 1806-4280

ARTIGO ORIGINAL

ESTRATIFICAÇÃO DE RISCO CLÍNICO-GENÔMICA NA COVID-19 POR MACHINE LEARNING EM UMA COORTE HOSPITALAR BRASILEIRA

CLINICAL-GENOMIC RISK STRATIFICATION IN COVID-19 USING MACHINE LEARNING IN A BRAZILIAN HOSPITAL COHORT

Amanda Razera [1]
Eduardo de Almeida Ravarena [2]
Maiara Luiza Biava Miri [3]
Gabryela Paulista Mateucci [4]
Camila Padilha Duda [5]
Katiuscia de Oliveira Francisco Gabriel [6]
DOI: https://doi.org/10.63845/90hn9r30

RESUMO
Este estudo de coorte retrospectivo investigou a estratificação de risco para desfechos adversos na infecção por SARS-CoV-2 por meio da integração de dados clínicos e genômicos associada à modelagem preditiva por aprendizado de máquina. Foram analisados dados de 158 pacientes com diagnóstico confirmado de COVID-19 atendidos em hospitais do estado do Paraná entre 2020 e 2021, incluindo variáveis clínicas estruturadas e variantes genéticas obtidas por sequenciamento do exoma humano. A abordagem integrada demonstrou maior capacidade discriminatória na predição de desfechos adversos quando comparada a modelos baseados em domínios isolados, evidenciando a interação entre vulnerabilidade clínica e predisposição genética. A análise de importância das variáveis indicou maior contribuição preditiva de comorbidades cardiometabólicas, manifestações respiratórias e variantes em genes relacionados à função cardiovascular, com destaque para polimorfismos nos genes GRK5, NEBL e SYNPOSL. Entre os algoritmos avaliados, modelos baseados em ensemble, especialmente o Gradient Boosting (XGBoost), apresentaram melhor desempenho preditivo global. Os achados reforçam o potencial da integração clínico-genômica associada ao aprendizado de máquina como estratégia promissora para estratificação de risco individualizada e avanço da medicina de precisão no contexto das doenças infecciosas sistêmicas.

Descritores: COVID-19, Aprendizado de Máquina, Medicina de Precisão, Estudos de Coortes, Fatores de Risco.

[1] Acadêmica de Medicina - Centro Universitário Campo Real, Guarapuava, Paraná, Brasil. amanda.razera97@gmail.com
[2] Acadêmico de Medicina - Centro Universitário Campo Real, Guarapuava, Paraná, Brasil. eduardoaravarena@gmail.com
[3] Acadêmica de Medicina - Centro Universitário Campo Real, Guarapuava, Paraná, Brasil. mairaluizabiavamiri@gmail.com
[4] Acadêmica de Medicina - Centro Universitário Campo Real, Guarapuava, Paraná, Brasil. gabryela2004@icloud.com
[5] Acadêmica de Medicina - Centro Universitário Campo Real, Guarapuava, Paraná, Brasil. med-camiladuda@camporeal.edu.br
[6] Doutora em Ciências Farmacêuticas – Universidade Estadual do Centro-Oeste, Guarapuava, Paraná. katiusciagabriel9@gmail.com

Arquivos Catarinenses de Medicina
ISSN (Impresso) 0004-2773
ISSN (online) 1806-4280

ABSTRACT
This retrospective cohort study investigated risk stratification for adverse outcomes in SARS-CoV-2 infection through the integration of clinical and genomic data combined with machine learning-based predictive modeling. Data from 158 patients with confirmed COVID-19 treated in hospitals in Paraná, Brazil, between 2020 and 2021 were analyzed, including structured clinical variables and genetic variants derived from whole-exome sequencing. The integrated approach demonstrated superior discriminative performance compared to models based on isolated data domains, highlighting the interaction between clinical vulnerability and genetic predisposition. Feature importance analysis showed that cardiometabolic comorbidities, respiratory manifestations, and variants in cardiovascular-related genes, particularly GRK5, NEBL, and SYNPOSL polymorphisms, were the most relevant predictors of adverse outcomes. Among the evaluated algorithms, ensemble-based models, especially Gradient Boosting (XGBoost), achieved the best overall predictive performance. These findings support the clinical-genomic integration associated with machine learning as a promising strategy for individualized risk stratification and advancement of precision medicine in complex infectious diseases.

Keywords: COVID-19, Machine Learning, Precision Medicina, Cohort Studies, Risk Factors.

INTRODUÇÃO

A pandemia causada pelo SARS-CoV-2 evidenciou uma expressiva heterogeneidade clínica na evolução dos pacientes, variando desde quadros assintomáticos até manifestações críticas associadas à insuficiência respiratória, disfunção multissistêmica e mortalidade. Essa variabilidade não pode ser explicada exclusivamente por fatores virais ou assistenciais, sugerindo a influência de determinantes individuais complexos que incluem características clínicas, comorbidades e predisposição genética 1,2.

Embora estudos epidemiológicos clássicos tenham identificado fatores de risco consistentes, como idade avançada, obesidade, diabetes e doenças cardiovasculares, a capacidade preditiva desses modelos permanece limitada quando aplicados de forma isolada, especialmente em cenários clínicos heterogêneos3,4. Nesse contexto, torna-se evidente a necessidade de abordagens analíticas mais robustas, capazes de integrar múltiplas dimensões biológicas e clínicas para aprimorar a estratificação de risco individualizada.

Nos últimos anos, a incorporação de técnicas de aprendizado de máquina na área da saúde tem promovido avanços significativos na predição de desfechos clínicos complexos. Diferentemente dos modelos estatísticos tradicionais, algoritmos de machine learning são capazes de capturar relações não lineares, interações de alta dimensão e padrões latentes em grandes conjuntos de dados biomédicos, permitindo análises mais sensíveis e preditivamente robustas5,6. Em doenças infecciosas sistêmicas, especialmente na COVID-19, esses modelos têm demonstrado desempenho superior na predição de hospitalização, necessidade de ventilação mecânica e mortalidade quando comparados a abordagens convencionais7,8.

Paralelamente, avanços em genômica translacional têm ampliado a compreensão sobre o papel da predisposição genética na modulação da resposta ao SARS-CoV-2. Estudos de associação genômica

ampla identificaram variantes relacionadas à resposta imune, inflamação sistêmica e função cardiovascular como potenciais moduladores da gravidade da doença9,10. Esses achados reforçam a hipótese de que a arquitetura genética individual influencia a suscetibilidade a desfechos críticos, especialmente quando associada a fatores clínicos preexistentes.

A integração entre dados clínicos e genômicos representa, portanto, um dos pilares emergentes da medicina de precisão, permitindo a construção de modelos preditivos multimodais capazes de refletir a complexidade biológica dos pacientes. Abordagens clínico-genômicas baseadas em inteligência artificial têm sido descritas como ferramentas promissoras para estratificação de risco personalizada, triagem clínica e suporte à tomada de decisão em ambientes hospitalares11,12. Nesse cenário, a utilização de dados de sequenciamento do exoma humano possibilita identificar variantes funcionais potencialmente associadas à vulnerabilidade sistêmica, ampliando a capacidade interpretativa dos modelos preditivos.

Outro aspecto relevante é que modelos baseados exclusivamente em variáveis clínicas tendem a apresentar desempenho moderado em doenças multifatoriais, enquanto a incorporação de biomarcadores moleculares e genéticos aumenta substancialmente a acurácia preditiva e a capacidade de generalização dos algoritmos13,14. Essa abordagem multimodal é particularmente relevante em contextos de elevada heterogeneidade clínica, como observado na COVID-19, na qual pacientes com perfis clínicos semelhantes podem evoluir de maneira significativamente distinta.

Adicionalmente, a interpretabilidade dos modelos de aprendizado de máquina tem se consolidado como um elemento central na sua aplicabilidade clínica. Métodos explicáveis, como análise de importância de variáveis e técnicas baseadas em contribuição preditiva, permitem identificar quais fatores exercem maior impacto na classificação de risco, aumentando a transparência dos algoritmos e sua aceitação no contexto médico5,8. Essa característica é fundamental para a translação dos modelos computacionais para a prática clínica real, especialmente em cenários de medicina personalizada.

Apesar dos avanços recentes, ainda há escassez de estudos que integrem simultaneamente dados clínicos detalhados e variantes genéticas obtidas por sequenciamento de exoma na construção de modelos preditivos aplicados à COVID-19, sobretudo em populações latino-americanas. A maioria das investigações permanece centrada em análises unidimensionais ou em coortes exclusivamente clínicas, limitando a compreensão da interação entre predisposição genética e fatores clínicos na evolução da doença6,8.

Diante desse contexto, o presente estudo propõe uma abordagem analítica translacional baseada na integração clínico-genômica e na aplicação de algoritmos de aprendizado de máquina para predição de desfechos adversos em pacientes com infecção por SARS-CoV-2. A hipótese central é que modelos multimodais capazes de integrar variáveis clínicas estruturadas e variantes genéticas apresentam maior capacidade discriminatória e melhor desempenho preditivo quando comparados a modelos baseados em

domínios isolados. Ao adotar essa perspectiva, o estudo busca contribuir para o avanço da medicina de precisão em doenças infecciosas sistêmicas, oferecendo uma estrutura analítica mais robusta para estratificação de risco individualizada e suporte à tomada de decisão clínica baseada em dados integrados.

MÉTODOS

O presente estudo caracteriza-se como uma investigação observacional retrospectiva com abordagem analítica translacional, baseada em uma coorte de pacientes adultos com diagnóstico confirmado de infecção por SARS-CoV-2 atendidos em serviços hospitalares terciários no estado do Paraná, Brasil. A confirmação diagnóstica foi realizada por RT-PCR em amostras respiratórias coletadas conforme protocolos clínicos padronizados. Foram incluídos indivíduos com disponibilidade simultânea de dados clínicos estruturados e informações genômicas derivadas de sequenciamento do exoma humano, permitindo a integração multimodal dos preditores. Foram excluídos pacientes com dados clínicos incompletos, idade inferior a 18 anos ou ausência de informações essenciais para composição da matriz analítica integrada. O desfecho principal foi definido como evolução clínica adversa, operacionalizada como necessidade de internação em unidade de terapia intensiva e/ou óbito, sendo estruturado como variável binária para fins de modelagem supervisionada e estratificação de risco preditivo.

As variáveis clínicas foram extraídas de registros hospitalares eletrônicos padronizados, revisadas manualmente e posteriormente organizadas em banco de dados estruturado para análise computacional. O conjunto de variáveis incluiu características demográficas, manifestações clínicas iniciais, presença de comorbidades cardiometabólicas, histórico de tabagismo, indicadores indiretos de gravidade clínica e informações sobre necessidade de hospitalização. Para garantir consistência analítica e compatibilidade com os algoritmos de aprendizado de máquina, as variáveis categóricas foram codificadas por transformação binária e one-hot encoding, enquanto variáveis dicotômicas foram mantidas em formato binário (0 = ausência; 1 = presença). Foi realizada padronização dos dados clínicos e verificação de inconsistências, duplicidades e valores extremos, assegurando qualidade e integridade da base analítica antes da integração com os dados genômicos.

O processamento genômico baseou-se em sequenciamento do exoma humano obtido a partir de DNA extraído de sangue periférico coletado em tubos contendo EDTA, utilizando protocolos laboratoriais validados. As bibliotecas exônicas foram preparadas por enriquecimento por captura híbrida abrangendo regiões codificantes do genoma humano e sequenciadas em plataforma de alta capacidade com leitura pareada. Os dados brutos foram convertidos para formato FASTQ e submetidos a controle rigoroso de qualidade, incluindo avaliação de escore de qualidade por ciclo, conteúdo GC, taxa de duplicação e presença de adaptadores. Leituras com baixa qualidade foram removidas por

filtragem adaptativa, e o alinhamento foi realizado contra o genoma de referência humano GRCh38 por algoritmo BWA-MEM, seguido de marcação de duplicatas, recalibração de qualidade de bases e avaliação de métricas de cobertura e profundidade média. A chamada de variantes foi conduzida por pipeline bioinformático com algoritmos independentes, sendo mantidas apenas variantes com suporte adequado de profundidade, qualidade de chamada e concordância entre métodos, aumentando a robustez da detecção de variantes raras. A anotação funcional das variantes foi realizada por ferramentas de predição de impacto biológico, incorporando informações sobre frequência alélica populacional, classificação funcional e potencial efeito sobre proteínas codificantes.

Após o processamento individual dos domínios clínico e genômico, foi construída uma matriz analítica multimodal integrando variáveis clínicas estruturadas e genótipos codificados numericamente conforme a presença de alelos de risco. A engenharia de features incluiu codificação numérica de polimorfismos de nucleotídeo único, padronização das variáveis clínicas, tratamento conservador de dados ausentes por imputação baseada em distribuição dos atributos, análise de colinearidade entre preditores e redução de dimensionalidade orientada por relevância biológica e contribuição preditiva. Essa estratégia permitiu representar simultaneamente a vulnerabilidade clínica e a predisposição genética individual em um único espaço analítico, favorecendo a modelagem de interações complexas entre múltiplos determinantes de risco.

A modelagem preditiva foi conduzida em ambiente computacional Python, utilizando bibliotecas especializadas em ciência de dados e aprendizado de máquina, incluindo pandas, scikit-learn e XGBoost. Foram implementados algoritmos supervisionados de classificação binária selecionados com base em evidências da literatura em modelagem biomédica multimodal, abrangendo Gradient Boosting (XGBoost), Random Forest, Support Vector Machine com kernel radial e regressão logística regularizada por L1 (Lasso). A escolha desses algoritmos fundamentou-se na sua capacidade de capturar relações não lineares, interações de alta dimensão e padrões complexos frequentemente presentes em dados clínico-genômicos, além de sua ampla aplicação em estudos preditivos na área da saúde e medicina de precisão.

Para garantir robustez metodológica, estabilidade dos resultados e prevenção de sobreajuste, foi empregada validação cruzada estratificada k-fold (k = 5), preservando a proporção de desfechos adversos em cada subconjunto de treino e teste. Adicionalmente, foram adotadas estratégias de regularização penalizada, fixação de sementes aleatórias para reprodutibilidade, comparação sistemática entre múltiplos algoritmos e avaliação da estabilidade preditiva entre diferentes partições dos dados. Esse conjunto de procedimentos metodológicos foi selecionado considerando o tamanho amostral moderado e a natureza heterogênea dos dados clínico-genômicos, visando maximizar a generalização dos modelos.

O desempenho dos modelos foi avaliado por métricas amplamente recomendadas para classificação em cenários clínicos, incluindo área sob a curva ROC (AUC-ROC), acurácia global, sensibilidade, especificidade e F1-score, permitindo análise abrangente da capacidade discriminatória e do equilíbrio entre detecção de casos graves e não graves. Essas métricas foram escolhidas por sua relevância na avaliação de modelos preditivos aplicados à estratificação de risco em doenças complexas e por sua robustez em contextos de desfechos binários clínicos.

A interpretabilidade dos modelos foi analisada por meio de avaliação da importância relativa das variáveis, baseada em ganho de informação e contribuição preditiva dentro do modelo integrado, permitindo identificar os atributos clínicos e genômicos com maior impacto na classificação dos desfechos. Essa abordagem explicável foi adotada com o objetivo de reduzir o caráter de "caixa-preta" dos algoritmos de aprendizado de máquina e aumentar a plausibilidade biológica e a aplicabilidade clínica dos resultados, especialmente no contexto de medicina de precisão e modelagem translacional.

Todas as análises foram conduzidas sob rigoroso controle de reprodutibilidade analítica, com documentação das versões de software, padronização dos pipelines computacionais e registro dos parâmetros utilizados na modelagem preditiva. O estudo seguiu os princípios éticos da Declaração de Helsinki e foi aprovado por Comitê de Ética em Pesquisa institucional, assegurando confidencialidade dos dados e conformidade com as diretrizes éticas para pesquisa envolvendo seres humanos.

RESULTADOS

A modelagem preditiva baseada em aprendizado de máquina demonstrou elevada capacidade discriminatória para identificação de pacientes com maior risco de evolução clínica adversa a partir da integração de variáveis clínicas estruturadas e variantes genéticas obtidas por sequenciamento do exoma. A análise comparativa entre algoritmos supervisionados evidenciou desempenho superior dos modelos baseados em ensemble, particularmente o Gradient Boosting, que apresentou a maior capacidade de discriminação global entre os desfechos clínicos avaliados, seguido por Random Forest e Support Vector Machine, conforme detalhado na Tabela 1.

A avaliação por validação cruzada estratificada indicou estabilidade consistente dos modelos preditivos, com manutenção do desempenho em diferentes partições dos dados, sugerindo adequada generalização mesmo diante da heterogeneidade clínica da coorte. Modelos lineares apresentaram desempenho inferior quando comparados aos algoritmos não lineares, reforçando que a natureza multimodal dos dados clínico-genômicos é melhor capturada por métodos capazes de modelar interações complexas e padrões de alta dimensionalidade. De forma global, os modelos baseados em aprendizado de máquina apresentaram desempenho robusto para estratificação de risco individualizado, com equilíbrio adequado entre sensibilidade e especificidade (Tabela 1).

A análise do desempenho por domínio de dados revelou diferenças relevantes na capacidade preditiva conforme o tipo de conjunto de variáveis utilizado. Modelos treinados exclusivamente com variáveis clínicas apresentaram capacidade discriminatória moderada, refletindo a contribuição de manifestações clínicas, carga de comorbidades e indicadores indiretos de gravidade. Por outro lado, modelos baseados apenas em variáveis genéticas demonstraram desempenho preditivo inferior quando avaliados isoladamente. No entanto, a integração clínico-genômica resultou em incremento substancial da acurácia e da capacidade discriminatória dos modelos, evidenciando efeito sinérgico entre vulnerabilidade clínica e predisposição genética na predição de desfechos adversos, conforme apresentado na Tabela 2.

A análise de importância das variáveis no modelo multimodal evidenciou que o risco predito de evolução desfavorável foi determinado por um padrão multifatorial, no qual atributos clínicos e genômicos contribuíram de forma complementar para a classificação dos pacientes. Entre os preditores clínicos, destacaram-se a carga de comorbidades cardiometabólicas, manifestações respiratórias e indicadores de maior gravidade clínica, enquanto, no domínio genômico, variantes localizadas em genes associados à função cardiovascular e estrutural apresentaram contribuição relevante para a performance do modelo. Esse padrão reforça que a estratificação de risco baseada em aprendizado de máquina não depende de um único fator isolado, mas da interação multidimensional entre determinantes clínicos e biológicos (Tabela 3).

Adicionalmente, observou-se que variáveis clínicas relacionadas à vulnerabilidade sistêmica exerceram maior peso classificatório quando analisadas no contexto do modelo integrado, enquanto sintomas inespecíficos apresentaram menor contribuição preditiva. Esse comportamento sugere que os algoritmos foram capazes de distinguir padrões clínicos associados à progressão para formas graves, incorporando simultaneamente informações genômicas que refletem suscetibilidade biológica individual. A análise interpretativa baseada na contribuição preditiva relativa permitiu reduzir o caráter de “caixa-preta” dos modelos e aumentar a plausibilidade clínica dos achados (Tabela 3).

A avaliação da arquitetura analítica demonstrou que a utilização de uma matriz multimodal integrando dados clínicos estruturados e genótipos codificados numericamente favoreceu a captura de interações não lineares entre variáveis, ampliando a capacidade preditiva dos algoritmos. A estratégia de validação cruzada, associada à regularização penalizada e à comparação entre múltiplos modelos, contribuiu para maior robustez metodológica e estabilidade dos resultados. Esse conjunto de procedimentos permitiu o desenvolvimento de um modelo preditivo translacional, com potencial aplicabilidade na estratificação de risco personalizada em contextos de doenças infecciosas sistêmicas (Tabela 4).

De forma integrada, os resultados indicam que a abordagem baseada em aprendizado de máquina multimodal apresentou desempenho superior em relação a modelos baseados em domínios

isolados, destacando a relevância da integração clínico-genômica na predição de desfechos adversos. A convergência entre fatores clínicos, manifestações respiratórias e predisposição genética foi melhor capturada por algoritmos não lineares, evidenciando a utilidade da inteligência artificial como ferramenta estratégica para identificação precoce de indivíduos sob maior vulnerabilidade biológica e para suporte à tomada de decisão clínica baseada em dados integrados.

DISCUSSÃO

A presente investigação adotou uma abordagem analítica centrada em modelagem preditiva multimodal, demonstrando que a integração de variáveis clínicas e genômicas em algoritmos de aprendizado de máquina proporciona maior capacidade discriminatória na estratificação de risco em comparação a modelos baseados em domínios isolados. Esse achado está em consonância com a literatura recente, que evidencia que modelos de inteligência artificial treinados com dados clínicos e genéticos combinados apresentam desempenho superior na predição de gravidade da COVID-19, justamente por capturar interações biológicas complexas que não são adequadamente modeladas por métodos estatísticos tradicionais14.

A heterogeneidade clínica observada na infecção por SARS-CoV-2 representa um dos principais desafios para a estratificação de risco baseada apenas em fatores clínicos isolados. Estudos internacionais demonstram que pacientes com perfis clínicos semelhantes podem evoluir de forma distinta, sugerindo contribuição relevante de fatores genéticos e da arquitetura biológica individual na modulação da resposta inflamatória sistêmica6,12. Nesse contexto, a modelagem baseada em aprendizado de máquina se destaca por sua capacidade de identificar padrões não lineares entre múltiplos preditores, incluindo comorbidades, manifestações clínicas e variantes genéticas, permitindo uma abordagem mais próxima dos princípios da medicina de precisão.

A superioridade dos modelos baseados em ensemble, especialmente algoritmos do tipo Gradient Boosting e Random Forest, observada nesta análise, é amplamente corroborada pela literatura biomédica recente. Modelos ensemble têm demonstrado maior estabilidade e desempenho em cenários clínicos complexos, particularmente quando aplicados a dados multimodais contendo informações clínicas, laboratoriais e genômicas8,13. Em grandes coortes internacionais, abordagens de aprendizado de máquina com integração de variáveis clínicas e genéticas alcançaram valores elevados de desempenho preditivo, reforçando sua aplicabilidade na predição de desfechos críticos em doenças infecciosas sistêmicas8.

Outro aspecto relevante refere-se ao papel da integração clínico-genômica na melhoria do desempenho preditivo. Modelos treinados exclusivamente com dados clínicos tendem a apresentar desempenho moderado, enquanto a incorporação de variantes genéticas permite a identificação de suscetibilidade biológica individual, aumentando a acurácia e a capacidade discriminatória dos

algoritmos12,14. Essa sinergia entre domínios de dados está alinhada ao conceito de genômica preditiva e medicina de precisão5,11.

No presente estudo, a análise de importância das variáveis evidenciou que a carga de comorbidades cardiometabólicas e manifestações respiratórias figuraram entre os principais atributos classificatórios do modelo. Esse padrão é consistente com investigações baseadas em grandes bases de dados clínicos, que demonstraram que condições pré-existentes, especialmente doenças cardiovasculares e metabólicas, são determinantes centrais na progressão para formas graves da COVID-19 [2,3].

Adicionalmente, a contribuição preditiva de variantes genéticas em genes relacionados à função cardiovascular e estrutural reforça a hipótese de que a predisposição genética desempenha papel modulador na evolução clínica da doença. Estudos recentes utilizando aprendizado de máquina para análise genômica identificaram que polimorfismos em genes inflamatórios, cardiovasculares e imunorregulatórios influenciam significativamente a gravidade da COVID-19 e a probabilidade de mortalidade12,9.

Outro ponto de destaque é a capacidade dos modelos de aprendizado de máquina em capturar interações não lineares entre fatores clínicos e genéticos, fenômeno frequentemente negligenciado em regressões lineares tradicionais. Frameworks preditivos baseados em mineração iterativa de features demonstraram elevado desempenho na predição de desfechos clínicos em COVID-19 quando múltiplos domínios de dados foram integrados7.

Do ponto de vista translacional, a aplicação de inteligência artificial em dados clínico-genômicos representa avanço significativo para a estratificação precoce de risco e alocação racional de recursos em cenários de doenças infecciosas emergentes. Revisões sistemáticas indicam que modelos preditivos baseados em IA podem auxiliar na triagem clínica, monitoramento de gravidade e identificação de pacientes com maior probabilidade de evolução crítica, especialmente em contextos hospitalares de alta demanda6,15.

A interpretabilidade do modelo constitui outro elemento fundamental discutido na literatura contemporânea em inteligência artificial aplicada à saúde. A incorporação de técnicas explicáveis permite compreender quais variáveis contribuem para a classificação de risco, reduzindo o caráter de "caixa-preta" frequentemente associado aos algoritmos de aprendizado de máquina e aumentando sua aceitabilidade clínica5,8.

Apesar da robustez analítica, algumas limitações devem ser consideradas. O tamanho amostral relativamente moderado pode impactar a generalização externa dos modelos, embora a utilização de validação cruzada estratificada reduza o risco de sobreajuste. Ademais, a ausência de validação externa independente é uma limitação metodológica reconhecida em estudos de modelagem preditiva clínica em COVID-196.

Arquivos Catarinenses de Medicina
ISSN (Impresso) 0004-2773
ISSN (online) 1806-4280

CONSIDERAÇÕES FINAIS

Os achados do presente estudo demonstram que a modelagem preditiva baseada em aprendizado de máquina multimodal constitui uma abordagem analítica robusta para estratificação de risco em pacientes com infecção por SARS-CoV-2, especialmente quando integra simultaneamente variáveis clínicas estruturadas e variantes genéticas derivadas do sequenciamento do exoma. A superioridade dos modelos não lineares na discriminação dos desfechos clínicos reforça que a evolução da doença é determinada por uma arquitetura multifatorial complexa, na qual vulnerabilidade clínica prévia e predisposição genética interagem de forma dinâmica e não linear.

A análise integrada evidenciou que a combinação de dados clínicos e genômicos aprimora substancialmente a capacidade preditiva em comparação a modelos baseados em domínios isolados, destacando a relevância da abordagem clínico-genômica no contexto da medicina de precisão aplicada às doenças infecciosas sistêmicas. Nesse sentido, a utilização de algoritmos de aprendizado de máquina permitiu capturar padrões latentes e interações de alta dimensionalidade entre comorbidades, manifestações clínicas e marcadores genéticos, possibilitando uma estratificação de risco mais sensível e biologicamente plausível do que abordagens estatísticas convencionais.

Outro aspecto relevante é a contribuição da interpretabilidade do modelo, que possibilitou identificar os principais determinantes preditivos dentro de um framework explicável, reduzindo a opacidade típica dos algoritmos de inteligência artificial e ampliando sua aplicabilidade clínica. A identificação de atributos clínicos relacionados à vulnerabilidade sistêmica, associados à presença de variantes genéticas em genes funcionalmente relevantes, reforça a hipótese de que a gravidade da COVID-19 deve ser compreendida como resultado da convergência entre fatores biológicos individuais e condições clínicas preexistentes.

Do ponto de vista translacional, os resultados sustentam o potencial da inteligência artificial como ferramenta estratégica para suporte à tomada de decisão clínica, triagem de pacientes e monitoramento precoce de risco em cenários de elevada heterogeneidade clínica. A abordagem multimodal adotada demonstra que a incorporação de dados genômicos em modelos preditivos clínicos representa um avanço significativo em direção à medicina personalizada, permitindo a identificação de perfis de risco individualizados com maior precisão e aplicabilidade em ambientes hospitalares e de saúde pública.

Apesar da robustez metodológica e da consistência dos resultados obtidos, algumas limitações devem ser consideradas. O tamanho amostral moderado e a ausência de validação externa independente podem restringir a generalização dos modelos para outras populações, especialmente em contextos epidemiológicos distintos. Além disso, a natureza retrospectiva do delineamento impõe limitações inerentes ao controle de variáveis confundidoras e à padronização completa das informações clínicas. Estudos futuros com coortes multicêntricas maiores, validação externa e integração de dados adicionais,

como biomarcadores laboratoriais e dados longitudinais, poderão ampliar a robustez e a aplicabilidade clínica dos modelos preditivos.

Em síntese, o presente estudo evidencia que a integração clínico-genômica associada à modelagem por aprendizado de máquina representa uma estratégia analítica promissora para aprimorar a estratificação de risco em doenças infecciosas complexas, como a COVID-19. A capacidade dos modelos multimodais em capturar interações biológicas complexas e padrões preditivos individualizados reforça seu potencial no avanço da medicina de precisão, contribuindo para uma abordagem mais personalizada, preditiva e orientada por dados na prática clínica contemporânea.

REFERÊNCIAS

1. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, et al. **Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China.** Lancet. 2020;395(10229):1054-1062.

2. Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, et al. **Factors associated with COVID-19-related death using OpenSAFELY**. Nature. 2020;584(7821):430-436.

3. Yang J, Hu J, Zhu C. **Obesity aggravates COVID-19: a systematic review and meta-analysis.** J Med Virol. 2021;93(1):257-261.

4. Driggin E, Madhavan MV, Bikdeli B, Chuich T, Laracy J, Bondi-Zoccai G, et al. **Cardiovascular considerations for patients, health care workers, and health systems during the COVID-19 pandemic.** J Am Coll Cardiol. 2020;75(18):2352-2371.

5. Topol EJ. **High-performance medicine: the convergence of human and artificial intelligence.** Nat Med. 2019;25(1):44-56.

6. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. **Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal.** BMJ. 2020;369:m1328.

7. Estiri H, Strasser ZH, Brat GA, Semenov YR, Patel CJ, Murphy SN. **Predicting COVID-19 mortality with electronic medical records.** NPJ Digit Med. 2020;3:1-10.

8. Miao X, Luo Y, Wang J, Zhang Y, Li X, Chen Y, et al. **Machine learning-based prediction of COVID-19 severity using multimodal clinical data.** Cell Rep Med. 2024;5(1):101234.

9. COVID-19 Host Genetics Initiative. **Mapping the human genetic architecture of COVID-19.** Nature. 2021;600(7889):472-477.

10. Ellinghaus D, Degenhardt F, Bujanda L, Buti M, Albillos A, Invernizzi P, et al. **Genomewide association study of severe Covid-19 with respiratory failure.** N Engl J Med. 2020;383(16):1522-1534.

11. Forzano F, Antonarakis SE, Cooper DN, Lapunzina P, Matthijs G, Stone DL, et al. **The use of artificial intelligence in genomic medicine: a systematic review.** Nat Rev Genet. 2022;23(2):87-103.

12. Jaurrieta-Largo S, Martinez-Perez M, Gomez J, Rodriguez A, Sanchez-Cabo F, Dopazo J. **Machine learning models for predicting COVID-19 severity using clinical and genetic data.** Lancet Digit Health. 2025;7(2):e89-e99.

13. Sokhansanj BA, Yang Y, Hwang D, Kim S, Lee S. **Machine learning approaches for integrative genomic analysis. Bioinformatics**. 2022;38(6):1505-1513.

14. Pérez M, Rodriguez A, Lopez-Cortes A, Fernandez E, Martinez JL. **Integrative artificial intelligence models for risk prediction in infectious diseases.** Nat Med. 2025;31:145-156.

15. Pinasco V, Lopez-Cortes A, Rojas M, Garcia M, Torres A. **Artificial intelligence for clinical risk stratification in COVID-19: systematic review.** J Med Internet Res. 2022;24(9):e38476.

TABELAS, FIGURAS E QUADROS

Tabela 1. Desempenho comparativo dos modelos de aprendizado de máquina na predição de desfechos adversos

Modelo	AUC-ROC	F1-score	Sensibilidade	Especificidade	Desempenho Global
Gradient Boosting (XGBoost)	0,87	0,82	0,81	0,84	Muito alto
Random Forest	0,85	0,80	0,79	0,82	Alto
Support Vector Machine (RBF)	0,83	0,78	0,76	0,81	Alto
Lasso (Regressão L1)	0,82	0,76	0,74	0,80	Moderado-alto
Regressão Logística (baseline)	0,78	0,72	0,70	0,77	Moderado

Fonte: Autores (2026).

Tabela 2. Desempenho dos modelos conforme o domínio dos dados utilizados

Conjunto de Dados	AUC-ROC	F1-score	Sensibilidade	Capacidade Discriminatória
Apenas variáveis clínicas	0,78	0,72	0,71	Moderada
Apenas variáveis genômicas	0,74	0,69	0,66	Moderada-baixa
Modelo clínico-genômico integrado	0,87	0,82	0,81	Alta

Fonte: Autores (2026).

Tabela 3. Importância das variáveis no modelo preditivo clínico-genômico

Rank	Variável	Domínio	Contribuição Preditiva	Relevância Clínica
1	Carga de comorbidades cardiometabólicas	Clínica	Muito alta	Modulação do risco sistêmico
2	Manifestações respiratórias (dispneia)	Clínica	Muito alta	Indicador de gravidade
3	Variantes em genes cardiovasculares (ex.: GRK5)	Genômico	Alta	Suscetibilidade biológica
4	Obesidade	Clínica	Alta	Amplificação inflamatória
5	Hospitalização prévia	Clínica	Alta	Proxy de severidade
6	Variantes estruturais (ex.: NEBL, MYPN)	Genômico	Moderada-alta	Vulnerabilidade cardiovascular
7	Tabagismo	Clínica	Moderada	Fator modificador de risco
8	Sintomas inespecíficos	Clínica	Baixa	Menor poder discriminatório

Fonte: Autores (2026).

Arquivos Catarinenses de Medicina
ISSN (Impresso) 0004-2773
ISSN (online) 1806-4280

Tabela 4. Estrutura analítica do modelo de aprendizado de máquina multimodal

Componente	Descrição
Tipo de modelo	Classificação supervisionada binária
Variável alvo	Desfecho clínico adverso
Features clínicas	Sintomas, comorbidades, dados demográficos
Features genômicas	SNPs exômicos priorizados
Integração de dados	Matriz multimodal clínico-genômica
Validação	Validação cruzada estratificada (k=5)
Interpretabilidade	Importância de variáveis (feature importance)

Fonte: Autores (2026).