Journal of Health Informatics

J Health Inform. 2026;18:e1848
DOI: https://doi.org/10.59681/2175-4411.v18.iEspecial.2026.1848

ARTIGO COMPLETO
CBIS 2026

GoslonAI: sistema inteligente de apoio ao diagnóstico ortodôntico em anomalias craniofaciais

GoslonAI: intelligent system to support orthodontic diagnosis in craniofacial anomalies

Ismael Alves Lima[1], Alice Castro Guedes Mendonça[2], Livian Isabel de Medeiros Carvalho[1], Adriel dos Santos Araújo Cabral[1], José Alisson Rocha da Silva[1], Eduarda Gomes Onofre de Araújo[1], Danilo Rangel Arruda Leite[1], Terumi Okada Ozawa[3], Thaís Gaudencio do Rêgo[1], Rosa Helena Wanderley Lacerda[4] ✉

ORCID: 0000-0001-7711-7256 · 0000-0003-1572-7547 · 0000-0001-7605-1523 · 0009-0002-2520-9831 · 0009-0007-4089-6293 · 0000-0001-7107-6107 · 0000-0001-5323-2508 · 0000-0002-1051-8140 · 0000-0002-6608-4900 · 0000-0003-0011-6487

[1] Universidade Federal da Paraíba, João Pessoa (PB), Brasil.
[2] Universidade Federal do Rio Grande do Norte, Natal (RN), Brasil.
[3] Universidade de São Paulo, São Paulo (SP), Brasil.
[4] Hospital Universitário Lauro Wanderley, Universidade Federal da Paraíba, João Pessoa (PB), Brasil.

FICHA DO ARTIGO

TIPO
Artigo completo · CBIS 2026

IDENTIFICADOR
e1848

DOI
10.59681/2175-4411.v18.iEspecial.2026.1848

RECEBIDO
17/05/2026

APROVADO
20/07/2026

PROCESSO DE AVALIAÇÃO
Double Blind Review

EDITOR RESPONSÁVEL
Juliano de Souza Gaspar

ACESSO
Aberto, CC BY 4.0

AUTOR CORRESPONDENTE
Rosa Helena Wanderley Lacerda
rhelenawanderley@gmail.com

LICENÇA

Este é um artigo publicado em acesso aberto sob a licença Creative Commons CC BY 4.0, que permite uso, distribuição e reprodução em qualquer meio, desde que o trabalho original seja corretamente citado.

CC BY 4.0

RESUMO

Objetivo: Desenvolver e avaliar a plataforma GoslonAI, um sistema de apoio à decisão clínica baseado em redes neurais convolucionais (CNNs) para classificação prognóstica automatizada de pacientes com fissura labiopalatina (FLP) utilizando o índice de GOSLON. **Método:** Estudo metodológico e retrospectivo realizado a partir de escaneamentos intraorais de pacientes com FLP provenientes de centros de referência. Foram avaliadas arquiteturas de CNNs pré-treinadas associadas a técnicas de fine-tuning e Grad-CAM. **Resultados:** A arquitetura VGG19 apresentou melhor desempenho, com acurácia de 83,5%, F1-score de 0,84 e AUC de 0,93. Os mapas Grad-CAM demonstraram coerência clínica entre as regiões analisadas pelo modelo e os critérios ortodônticos do índice de GOSLON. Além disso, foi desenvolvida uma plataforma web responsiva para classificação automatizada e geração de relatórios. **Conclusão:** A GoslonAI demonstrou potencial como ferramenta de apoio ao diagnóstico ortodôntico em pacientes com FLP.

Descritores: Inteligência artificial; Aprendizado profundo; Fenda labial; Fissura palatina.

ABSTRACT

Objective: To develop and evaluate the GoslonAI platform, a clinical decision support system based on convolutional neural networks (CNNs) for automated prognostic classification of patients with cleft lip and palate (CLP) using the GOSLON index. **Methods:** A methodological and retrospective study was conducted using intraoral scans from patients with CLP obtained from reference centers. Pre-trained CNN architectures associated with fine-tuning and Grad-CAM techniques were evaluated. **Results:** The VGG19 architecture achieved the best performance, with 83.5% accuracy, 0.84 F1-score, and 0.93 AUC. Grad-CAM maps demonstrated clinical coherence between the regions analyzed by the model and the orthodontic criteria of the GOSLON index. In addition, a responsive web platform was developed for automated classification and report generation. **Conclusion:** GoslonAI demonstrated potential as a support tool for orthodontic diagnosis in patients with CLP.

Keywords: Artificial intelligence; Deep learning; Cleft lip; Cleft palate.

CBIS 2026 XXI CONGRESSO BRASILEIRO DE INFORMÁTICA EM SAÚDE

Artigo apresentado no XXI Congresso Brasileiro de Informática em Saúde (CBIS 2026), realizado de 23 a 25 de setembro de 2026, em Brasília (DF), e publicado na seção CBIS 2026 do Journal of Health Informatics.

COMO CITAR ESTE ARTIGO: Lima IA, Mendonça ACG, Carvalho LIM, Cabral ASA, da Silva JAR, de Araújo EGO, et al. GoslonAI: sistema inteligente de apoio ao diagnóstico ortodôntico em anomalias craniofaciais. J Health Inform 2026;18:e1848. doi: https://doi.org/10.59681/2175-4411.v18.iEspecial.2026.1848

1 INTRODUÇÃO

As fissuras labiopalatinas (FLP) figuram entre as anomalias craniofaciais congênitas mais prevalentes no mundo, apresentando frequência média estimada entre 1 a cada 1.000 e 1.500 nascimentos, segundo a Organização Mundial da Saúde (OMS) [1]. As FLP representam um relevante problema de saúde pública devido às repercussões funcionais, estéticas, psicológicas e sociais causadas aos indivíduos acometidos, exigindo acompanhamento multidisciplinar e especializado ao longo de diferentes fases da vida [2,3].

Nesse contexto, a Ortodontia desempenha papel fundamental no tratamento e na reabilitação funcional e estética desses pacientes. Entre os principais instrumentos utilizados na avaliação ortodôntica destaca-se o índice de GOSLON (*Great Ormond Street, London and Oslo*), proposto por Mars *et al.* (1987) [4], amplamente utilizado para classificação da severidade da má oclusão e definição do prognóstico de crescimento maxilofacial em pacientes com FLP unilateral completa. O índice categoriza os pacientes em cinco grupos prognósticos a partir da análise das relações interarcos nos sentidos anteroposterior, transversal e vertical, permitindo auxiliar no planejamento terapêutico e na definição da complexidade do tratamento ortodôntico [4]. Estudos posteriores demonstraram a relevância clínica e a confiabilidade do índice na avaliação de resultados ortodônticos em pacientes com FLP [5,6]. Apesar de sua ampla utilização em centros especializados, sua aplicação ainda depende expressivamente da experiência clínica do avaliador, podendo apresentar variabilidade interexaminadores [5].

Paralelamente, os avanços recentes da inteligência artificial (IA), principalmente nas áreas de visão computacional e aprendizado profundo (*deep learning*), têm ampliado expressivamente as possibilidades de aplicação de métodos automatizados no contexto da saúde [7]. Aplicações de IA em imagens médicas têm apresentado resultados promissores em tarefas como diagnóstico, segmentação anatômica, classificação de doenças e apoio à decisão clínica [8,9]. Na odontologia, redes neurais convolucionais (*convolutional neural networks*, CNNs) vêm sendo utilizadas na detecção de cáries, segmentação de estruturas anatômicas e identificação automatizada de alterações craniofaciais [8].

No contexto das FLP, revisões recentes da literatura apontam crescimento progressivo das aplicações de IA voltadas à avaliação clínica, planejamento terapêutico e predição prognóstica [10,11]. Estudos também demonstraram o potencial da utilização de escaneamentos digitais e modelos tridimensionais na avaliação ortodôntica de pacientes com FLP, favorecendo maior padronização e reprodutibilidade diagnóstica [5,12].

Apesar desses avanços, ainda existe escassez de estudos direcionados especificamente à classificação prognóstica automatizada baseada no índice de GOSLON utilizando imagens derivadas de escaneamentos intraorais e técnicas de aprendizado profundo [10,11]. Além disso, em aplicações clínicas de IA, torna-se cada vez mais necessária a utilização de métodos que proporcionem maior interpretabilidade e transparência das decisões automatizadas [13]. Nesse cenário, técnicas de explicabilidade vêm sendo amplamente utilizadas para destacar as regiões das imagens que mais influenciam a decisão dos modelos, contribuindo para maior confiabilidade clínica e compreensão dos padrões aprendidos pelas redes neurais [13].

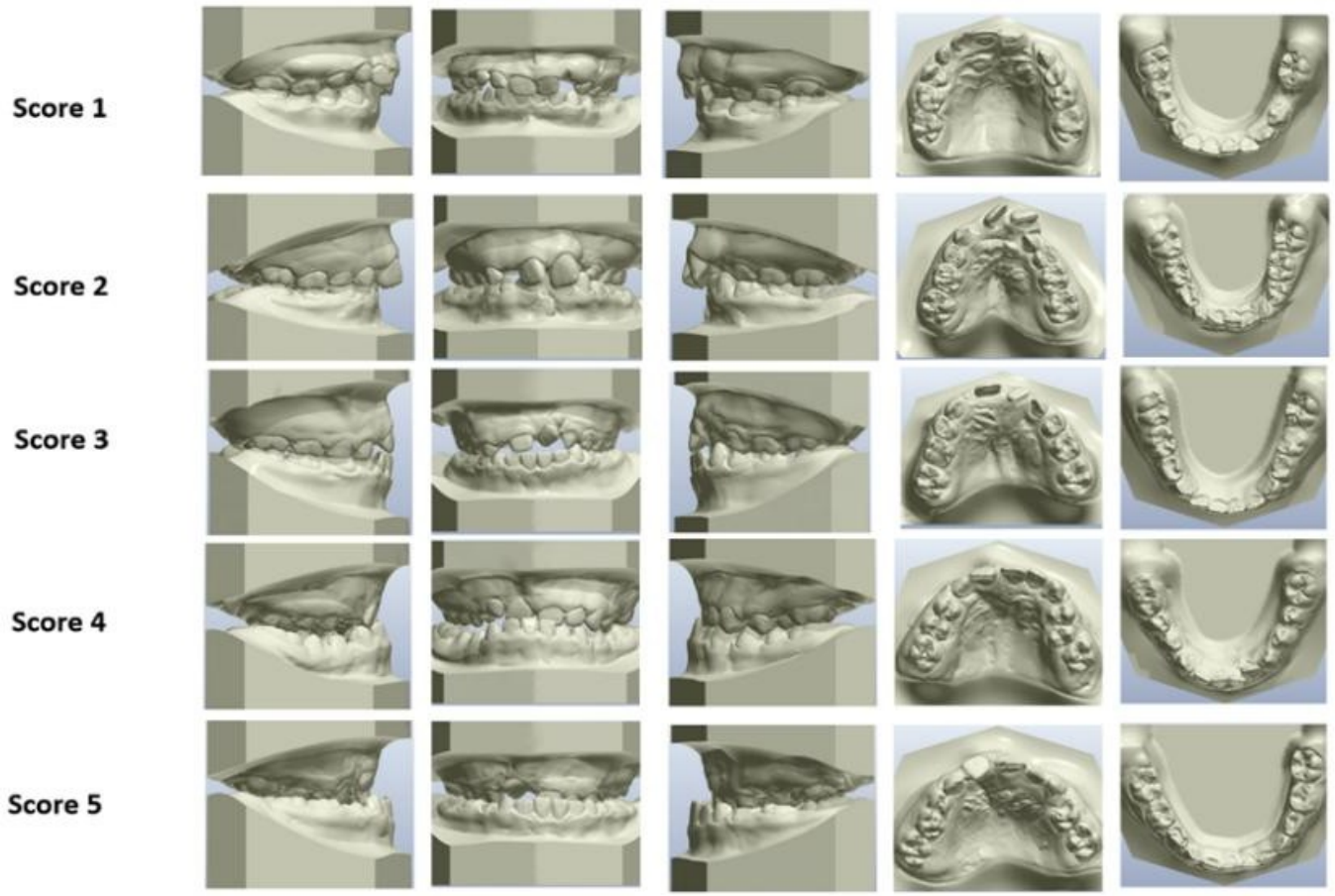
Diante da necessidade de maior padronização diagnóstica, da escassez de profissionais especializados em FLP e dos avanços recentes em IA aplicada à saúde, este estudo propõe o desenvolvimento e a avaliação da plataforma GoslonAI, um sistema inteligente de apoio à decisão clínica baseado em CNN para classificação prognóstica automatizada de pacientes com FLP unilateral. Para isso, foram utilizadas imagens derivadas de escaneamentos intraorais e arquiteturas de CNN pré-treinadas no conjunto de dados ImageNet, posteriormente ajustadas utilizando a técnica de *fine-tuning*. Adicionalmente, empregou-se uma técnica permitindo maior interpretabilidade das previsões realizadas pelos modelos. Dessa forma, busca-se contribuir para maior padronização, reprodutibilidade e apoio ao diagnóstico ortodôntico em pacientes com FLP, além de ampliar a integração entre IA e Saúde Digital no contexto das anomalias craniofaciais.

2 MÉTODOS

2.1 Delineamento do estudo

Trata-se de um estudo metodológico, observacional e retrospectivo, com abordagem quantitativa, voltado ao desenvolvimento e à avaliação da plataforma GoslonAI para classificação prognóstica automatizada de pacientes com FLP por meio de CNNs. O estudo foi aprovado pelo Comitê de Ética em Pesquisa, com Certificado de Apresentação para Apreciação Ética (CAAE) nº 81982724.6.0000.5183.

Figura 1. Exemplos de escaneamentos intraorais e suas respectivas classificações segundo o índice de GOSLON.



Fonte: Imagens fornecidas pelo HULW.

2.2 Base de dados

A coleta de dados foi realizada no Serviço de Fissuras Labiopalatinas do Hospital Universitário Lauro Wanderley (HULW), no contexto de um estudo transversal observacional com pacientes cadastrados entre 2015 e 2023, totalizando aproximadamente 2.650 indivíduos. Foram incluídos pacientes de ambos os sexos, com FLP não sindrômica, idade mínima de seis anos, palato operado há pelo menos um ano e disponibilidade de modelos digitais escaneados previamente ao tratamento ortodôntico. Foram excluídos indivíduos com idade superior a 20 anos, síndromes associadas, múltiplas perdas dentárias ou casos considerados complexos pelos especialistas. Todos os participantes concordaram em participar da pesquisa mediante assinatura do Termo de Consentimento Livre e Esclarecido (TCLE) e, quando aplicável, do Termo de Assentimento. Também foram reutilizados dados de um estudo multicêntrico conduzido sob os mesmos parâmetros metodológicos em centros brasileiros de referência no tratamento de FLP, incluindo o Centro de Atenção aos Defeitos da Face do Instituto de Medicina Integral Prof. Fernando Figueira (IMIP), em Recife-PE, e o Hospital de Reabilitação de Anomalias Craniofaciais da Universidade de São Paulo (USP), em Bauru-SP. Foram utilizados escaneamentos intraorais dos arcos dentários superior e inferior para composição da base de dados. Os escaneamentos foram avaliados independentemente por três especialistas em Ortodontia com experiência no tratamento de FLP e classificados de acordo com o índice oclusal de GOSLON [4-6]. A concordância interexaminadores foi avaliada pelo coeficiente Kappa de Cohen [5]. O índice de GOSLON classifica as relações oclusais em cinco categorias: excelente (escore 1), bom (escore 2), moderado (escore 3), pobre (escore 4) e muito pobre (escore 5). Para o treinamento do modelo de IA, os escores foram agrupados em três classes prognósticas: favorável (escores 1 e 2), moderada (escore 3) e desfavorável (escores 4 e 5). Ao todo, foram obtidos 587 escaneamentos intraorais, distribuídos em 173 amostras da Classe 1, 143 da Classe 2 e 271 da Classe 3. A partir desses escaneamentos, foram geradas imagens bidimensionais contendo vistas laterais dos arcos dentários posicionadas frontalmente entre si, armazenadas em formato JPEG e organizadas conforme as respectivas classes prognósticas.

Foi realizada a técnica de *undersampling* para balanceamento das classes, de modo que cada classe passou a conter 143 amostras, totalizando 429 registros. Em seguida, a base de dados foi dividida de forma estratificada em dois subconjuntos: treinamento/validação (80%) e teste (20%), preservando-se a

distribuição balanceada entre as classes em ambos os conjuntos. Para assegurar a reprodutibilidade dos experimentos, os subconjuntos foram documentados em arquivos no formato CSV contendo os identificadores das imagens e seus respectivos rótulos.

2.3 Aquisição e pré-processamento das imagens

Após a coleta e organização da base de dados, foi realizada a etapa de pré-processamento das imagens utilizando as bibliotecas *TensorFlow* e *Keras*, com o objetivo de padronizar as entradas e garantir maior estabilidade durante o treinamento dos modelos. As imagens foram organizadas em lotes contendo 32 amostras e submetidas às funções de pré-processamento específicas de cada arquitetura, responsáveis pela normalização dos pixels e adequação dos canais de cor conforme os padrões utilizados no pré-treinamento das CNNs. Durante o treinamento, foram aplicadas técnicas de aumento artificial de dados (*data augmentation*) utilizando o módulo *ImageDataGenerator* da biblioteca Keras 2.15.0, visando ampliar a variabilidade das amostras e reduzir os efeitos de *overfitting* [8]. Entre as transformações utilizadas destacam-se rotações aleatórias de até 15 graus, deslocamentos horizontais e verticais, operações de cisalhamento, zoom, inversões horizontais e variações de brilho, aplicadas estocasticamente durante o treinamento. Para os conjuntos de validação e teste não foram aplicadas técnicas de aumento artificial de dados. Por fim, a semente aleatória foi fixada em 42 para garantir a reprodutibilidade experimental.

2.4 Modelo de aprendizagem de máquina

As CNNs constituem uma classe de redes neurais artificiais amplamente empregadas em tarefas de visão computacional e classificação de imagens médicas. Neste estudo, foram avaliadas arquiteturas clássicas e recentes de CNNs, incluindo VGG16 [11], VGG19 [11], ResNet50, InceptionV3, Xception, DenseNet201, MobileNetV3Large e EfficientNetV2L [12], selecionadas por apresentarem diferentes estratégias de extração de características, eficiência computacional e desempenho preditivo. Enquanto modelos como VGG16 e VGG19 destacam-se pela simplicidade estrutural, arquiteturas como ResNet50, DenseNet201, InceptionV3 e Xception introduzem mecanismos que favorecem o treinamento de redes profundas e a extração multiescala de características. Já EfficientNetV2L e MobileNetV3Large foram desenvolvidas com foco em otimização computacional e eficiência de desempenho.

2.5 Avaliação do modelo

Para avaliação do desempenho das CNNs, foram utilizadas métricas amplamente empregadas em problemas de classificação multiclasse na área da saúde. Durante o treinamento, foram monitoradas as métricas de acurácia e AUC (*Area Under the ROC Curve*), calculadas automaticamente pelo método fit da biblioteca Keras, além da análise das curvas de perda (*loss*) e desempenho ao longo das épocas para identificação de possíveis comportamentos de *overfitting* e *underfitting*. A acurácia foi calculada conforme a expressão:

$$\text{Acurácia} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}} \quad (1)$$

em que VP corresponde aos verdadeiros positivos, VN aos verdadeiros negativos, FP aos falsos positivos e FN aos falsos negativos. Na etapa de testes, além da acurácia, foram utilizadas as métricas de precisão, revocação (*recall*) e F1-score [14], permitindo análise mais detalhada da capacidade de generalização dos modelos e do desempenho em cada classe prognóstica do índice de GOSLON. As métricas foram calculadas utilizando funções da biblioteca Scikit-Learn 1.6.1, complementadas por matrizes de confusão para identificação de padrões de erro e possíveis dificuldades de discriminação entre categorias clinicamente semelhantes.

2.6 Treinamento dos modelos

O treinamento das arquiteturas de CNN foi realizado no ambiente *Google Colaboratory* utilizando GPUs NVIDIA T4. Para padronização experimental, todas as arquiteturas receberam uma camada de classificação composta por uma camada totalmente conectada com 256 neurônios e ativação ReLU, seguida por uma camada de saída com três neurônios e função Softmax, correspondente às classes prognósticas definidas no estudo. O processo de otimização foi conduzido utilizando o algoritmo Adam [15] com taxa de aprendizado inicial de 10^{-3} e função de perda *categorical cross-entropy*, apropriada para classificação multiclasse. Durante o treinamento, foram utilizados os *callbacks* *EarlyStopping*, *ReduceLROnPlateau* e *ModelCheckpoint* para monitoramento do desempenho e redução de *overfitting*. A avaliação dos modelos foi realizada utilizando validação cruzada estratificada (*Stratified K-Fold*) com K=5 e 10 execuções independentes, implementada com a biblioteca *Scikit-Learn* 1.6.1. As métricas foram agregadas por média e desvio-padrão. Após a seleção do modelo com melhor desempenho médio, realizou-se etapa adicional de *fine-tuning* e avaliação no conjunto de testes utilizando as métricas de acurácia, precisão, revocação, F1-*score* e matrizes de confusão.

2.7 Explicabilidade dos modelos

A explicabilidade e a interpretabilidade das previsões representam aspectos fundamentais em aplicações de IA na área da saúde, especialmente em sistemas de apoio à decisão clínica. Nesse contexto, foi utilizada a técnica Grad-CAM (*Gradient-weighted Class Activation Mapping*) [16], responsável pela geração de mapas de ativação (*heatmaps*) capazes de destacar visualmente as regiões da imagem que exerceram maior influência sobre a decisão final do modelo. A utilização dessa

abordagem permite análise qualitativa do comportamento das CNNs, favorecendo maior transparência, interpretabilidade e confiabilidade clínica das predições realizadas. Além disso, os mapas de ativação auxiliam na identificação de possíveis vieses e padrões espúrios aprendidos pelos modelos, contribuindo para maior robustez e segurança da aplicação no contexto ortodôntico.

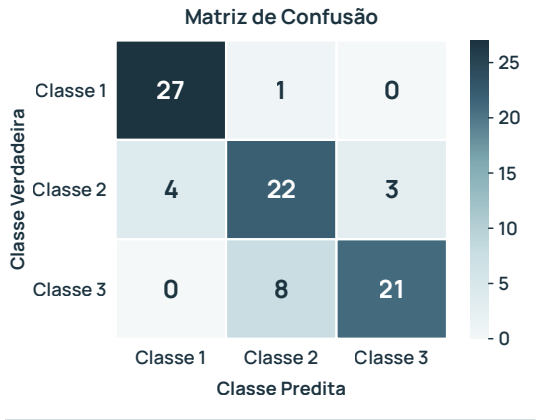
2.8 Arquitetura da ferramenta

O modelo de IA foi incorporado a uma solução estruturada como uma aplicação web *full-stack* responsiva, compatível com dispositivos móveis e computadores, permitindo acesso simplificado por meio de navegadores web. A arquitetura do sistema foi organizada em diferentes camadas funcionais, contemplando interface do usuário, *backend* de processamento, banco de dados e serviço externo de inferência por IA. O *frontend* da aplicação foi desenvolvido utilizando React, Reactstrap e Bootstrap, enquanto o *backend* foi implementado em Python com *framework* Flask e autenticação baseada em JSON Web Token (JWT). O armazenamento dos dados estruturados ocorre em banco PostgreSQL gerenciado pelo Supabase.

2.9 Fluxo da ferramenta

O fluxo operacional da plataforma inicia-se com o envio das imagens relacionadas à avaliação ortodôntica do paciente. Após o *upload*, o sistema realiza automaticamente etapas de validação e pré-processamento. O processamento ocorre por meio de comunicação REST segura com o modelo de IA, utilizando imagens codificadas em Base64. Após a inferência, a plataforma retorna automaticamente a classe prognóstica predita e gera um relatório digital em formato PDF contendo os resultados da análise ortodôntica. Além da funcionalidade de classificação automatizada, a plataforma incorpora mecanismos relacionados à experiência do usuário e à segurança da informação, incluindo autenticação segura, recuperação de senha, histórico de análises, coleta de *feedback* e adequação aos princípios da Lei Geral de Proteção de Dados (LGPD) [17]. Outro aspecto relevante refere-se à arquitetura modular da solução [13], na qual o modelo de IA encontra-se desacoplado do *backend* principal da aplicação, favorecendo escalabilidade, manutenção e futuras atualizações do sistema.

Figura 2. Matriz de confusão da VGG19 no conjunto de testes.



Fonte: Elaborado pelos autores (2026).

3 RESULTADOS

De modo geral, as arquiteturas da família VGG apresentaram os melhores desempenhos entre os modelos avaliados (Tabela 1). A VGG19 obteve os melhores resultados globais, alcançando 83,5% de acurácia média, F1-*score* de 0,84 e AUC de 0,93, seguida pela VGG16 com 83,3% de acurácia e F1-*score* de 0,83. Apesar da elevada complexidade computacional dessas arquiteturas, os resultados sugerem maior capacidade de captura de padrões relacionados às relações interarcos utilizadas na classificação prognóstica. A MobileNetV3Large também apresentou desempenho competitivo, atingindo 82,8% de acurácia e F1-*score* de 0,83, destacando-se pela elevada eficiência computacional e reduzido número de parâmetros.

As arquiteturas DenseNet201 e ResNet50 apresentaram desempenho intermediário, enquanto InceptionV3, Xception e EfficientNetV2L obtiveram resultados inferiores no conjunto de dados utilizado. Embora a EfficientNetV2L tenha apresentado elevado desempenho no *benchmark* ImageNet, registrou apenas 72% de acurácia neste estudo, sugerindo possível limitação relacionada ao tamanho reduzido da base de

Tabela 1. Resultados de avaliação das arquiteturas.

Arquitetura	Acurácia	+/-	Precisão	+/-	Revocação	+/-	F1-*score*	+/-	AUC	+/-
VGG19	83,52%	0,010	0,84	0,011	0,83	0,010	0,84	0,009	0,93	0,005
VGG16	83,34%	0,011	0,84	0,008	0,83	0,011	0,83	0,011	0,93	0,008
MobileNetV3Large	82,82%	0,012	0,84	0,011	0,83	0,012	0,83	0,012	0,93	0,007
DenseNet201	81,59%	0,006	0,83	0,009	0,82	0,006	0,82	0,006	0,92	0,005
ResNet50	79,59%	0,009	0,81	0,007	0,80	0,009	0,80	0,010	0,92	0,007
InceptionV3	74,57%	0,019	0,76	0,020	0,75	0,019	0,74	0,021	0,87	0,010
Xception	72,48%	0,011	0,74	0,014	0,72	0,011	0,72	0,011	0,85	0,009
EfficientNetV2L	72,01%	0,011	0,73	0,011	0,72	0,011	0,72	0,012	0,85	0,013

Fonte: Elaborado pelos autores (2026).

Tabela 2. Relatório de classificação da VGG19 no conjunto de testes.

	Precisão	Revocação	F1-*score*	Nº	Acurácia
Classe 1	0,87	0,96	0,92	28	0,81
Classe 2	0,71	0,76	0,73	29	
Classe 3	0,88	0,72	0,79	29	
Média	0,82	0,81	0,81	86	

Fonte: Elaborado pelos autores (2026).

Figura 3. Visualizações Grad-CAM sobre previsões corretas no conjunto de testes.



Fonte: Elaborado pelos autores (2026).

dados e à especificidade do problema clínico analisado. Após a etapa de validação cruzada, a VGG19 foi selecionada como modelo preditivo final. Durante o treinamento, a arquitetura apresentou comportamento estável, mantendo valores de acurácia entre 80% e 90% nos conjuntos de treino e validação. No conjunto de testes, o modelo atingiu 81% de acurácia global e F1-*score* médio de 0,81. A Classe 1 apresentou os melhores resultados, com *recall* de 0,96 e F1-*score* de 0,92, indicando elevada capacidade de identificação de casos com prognóstico favorável (Tabela 2). Já as Classes 2 e 3 apresentaram maior sobreposição de características, resultando em maior confusão entre categorias prognósticas intermediárias e desfavoráveis (Tabela 2).

A análise da matriz de confusão (Figura 2) demonstrou ausência de erros entre as Classes 1 e 3 no conjunto de testes, aspecto particularmente relevante do ponto de vista clínico, uma vez que erros entre prognósticos extremos poderiam impactar diretamente a definição terapêutica e o planejamento ortodôntico. Complementarmente, as visualizações obtidas por meio da técnica Grad-CAM (Figura 3) evidenciaram que os modelos concentraram sua atenção principalmente sobre os arcos dentários e regiões relacionadas às relações interarcos, indicando coerência clínica entre os padrões utilizados pela rede neural e os critérios ortodônticos empregados na classificação do índice de GOSLON.

A plataforma GoslonAI foi desenvolvida como uma interface web para classificação prognóstica automatizada baseada no índice de GOSLON. A solução permite o envio de imagens derivadas de escaneamentos intraorais diretamente pelo navegador, executando automaticamente etapas de validação, pré-processamento e inferência por IA. Após a análise, o sistema retorna a classe prognóstica predita e gera um relatório digital em formato PDF contendo os resultados ortodônticos. A plataforma apresenta arquitetura responsiva e desacoplada, possibilitando atualização contínua dos modelos e adequação aos princípios da LGPD, sem armazenamento permanente das imagens submetidas.

4 DISCUSSÃO

Os achados deste estudo evidenciam que CNNs, quando aplicadas a imagens derivadas de escaneamentos intraorais, apresentam bom desempenho clínico e são capazes de classificar prognosticamente pacientes com FLP unilateral com base no índice de GOSLON. De modo geral, as arquiteturas avaliadas apresentaram desempenho satisfatório, com destaque para os modelos da família VGG, particularmente a VGG19, alcançando acurácia de 83,5%, F1-score de 0,84 e AUC de 0,93. Esses achados estão de acordo com evidências científicas atuais, as quais demonstram o potencial das CNNs em tarefas de classificação de imagens ortodônticas derivadas de escaneamentos intraorais [6-8].

A superioridade observada das arquiteturas VGG em relação a modelos mais recentes, como EfficientNetV2L e Xception, pode estar relacionada às características do

conjunto de dados analisado. Embora essas arquiteturas mais modernas apresentem resultados expressivos em *benchmarks* amplamente utilizados, como o ImageNet, os achados deste estudo sugerem que modelos mais complexos nem sempre representam a melhor escolha para bases de dados clínicas reduzidas e altamente específicas. Isso pode ocorrer porque, em cenários com menor número de dados, a complexidade do modelo torna-se um fator decisivo para a capacidade de generalização. Estudos prévios demonstram que redes de menor complexidade podem apresentar desempenho comparável ou até superior ao de arquiteturas mais sofisticadas quando aplicadas a bases pequenas, especialmente em contextos médicos, nos quais a limitação amostral, a especificidade do domínio e a variabilidade anatômica podem dificultar a validação e favorecer o sobreajuste do modelo [14,18].

Nesse contexto, a VGG19 pode ter apresentado melhor desempenho justamente por sua arquitetura mais direta e estável, baseada em convoluções sequenciais com filtros pequenos de 3×3. Essa característica pode favorecer a identificação de padrões locais e graduais nas relações interarcos [11], que são fundamentais para a classificação pelo índice de GOSLON. Esse comportamento já foi descrito na literatura de IA aplicada à saúde, especialmente em cenários marcados por tamanho amostral limitado, elevada especificidade do domínio e variabilidade anatômica importante. Nessas condições, arquiteturas mais sofisticadas podem apresentar maior risco de *overfitting*, enquanto modelos mais simples e bem estabelecidos podem oferecer desempenho mais consistente e melhor adaptação ao problema investigado [7].

Outro aspecto importante refere-se ao desempenho diferencial entre as classes prognósticas. A Classe 1 (escores 1 e 2 do GOSLON) apresentou *recall* de 0,96 e F1-*score* de 0,92, demonstrando a capacidade do modelo de identificar casos com prognóstico favorável. Esse resultado é clinicamente relevante, pois a identificação precoce e confiável de pacientes com prognóstico favorável pode orientar condutas conservadoras, evitando intervenções cirúrgicas desnecessárias em fases críticas do desenvolvimento craniofacial. Por outro lado, a maior sobreposição observada entre as Classes 2 e 3 reflete uma dificuldade inerente ao próprio problema clínico do prognóstico moderado e desfavorável, pois frequentemente envolve julgamentos subjetivos, mesmo entre especialistas experientes. Estudos de confiabilidade com o índice de GOSLON reportam variabilidade interexaminadores com coeficiente Kappa variando de moderado a substancial, especialmente entre as categorias intermediárias [4,5,19,20], o que reforça a dificuldade de discriminar esses grupos limítrofes. Nesse sentido, o comportamento do modelo espelha, de certa forma, a variabilidade diagnóstica humana. Estudos em IA médica demonstram que modelos de aprendizado profundo tendem a apresentar maior incerteza e erros precisamente nas mesmas regiões de interface entre classes onde examinadores humanos também divergem, fenômeno atribuído à ambiguidade intrínseca do problema clínico e não a falhas exclusivas do sistema automatizado [15].

Um dado particularmente relevante entre as classes prognósticas foi a ausência completa de erros entre as Classes 1 e 3 na etapa de testes. Em sistemas de classificação ordinal aplicados à saúde, o custo clínico dos erros é intrinsecamente assimétrico: confundir categorias extremas representa um risco qualitativamente superior ao de confundir categorias adjacentes, pois implica condutas terapêuticas radicalmente divergentes [21]. Confundir um prognóstico excelente com um muito desfavorável ou vice-versa poderia levar a intervenções desnecessárias ou à omissão de tratamentos em momentos críticos do desenvolvimento craniofacial do paciente. A preservação dessa discriminação nos extremos prognósticos representa, portanto, um atributo de segurança clínica do sistema, alinhado com o que a literatura denomina de “erros clinicamente plausíveis” como critério de avaliação de sistemas de IA em saúde [15,21].

A análise por meio da técnica Grad-CAM acrescenta uma dimensão qualitativa fundamental aos resultados quantitativos. Os mapas de ativação demonstraram que o modelo direciona sua atenção predominantemente para os arcos dentários e para as regiões de relação interarco, compatíveis com os critérios clínicos que fundamentam o índice de GOSLON. Esse alinhamento entre as regiões de interesse do modelo e os critérios diagnósticos estabelecidos não pode ser interpretado de maneira comum, pois os sistemas de IA que apresentam alta acurácia, mas baseiam suas predições em regiões anatomicamente irrelevantes, constituem um risco real em aplicações clínicas [10]. Estudos recentes em IA aplicada à Ortodontia têm destacado o Grad-CAM como uma boa ferramenta para validação da plausibilidade clínica das predições automatizadas [22, 23], e sua adoção vem sendo recomendada como requisito para a tradução de modelos de aprendizado profundo ao contexto assistencial [7,24]. A coerência observada nos mapas Grad-CAM da GoslonAI indica que o sistema não apenas classifica corretamente, mas o faz por razões clinicamente interpretáveis, indicando sua credibilidade e potencializando a aceitação por parte dos profissionais de saúde. Achados semelhantes foram descritos em estudos que utilizaram Grad-CAM para classificação de anomalias alveolares em pacientes com FLP [10] e para detecção de cáries em radiografias panorâmicas [24].

Do ponto de vista translacional, a GoslonAI avança além do que a maioria dos estudos de IA em FLP tem proposto. Revisões da literatura apontam que grande parte das iniciativas nessa área permanece restrita à validação computacional de modelos, sem avançar

para soluções efetivamente integráveis ao contexto assistencial [7,8]. A plataforma desenvolvida neste estudo, ao incorporar interface web responsiva, geração automatizada de relatórios em PDF, autenticação segura e adequação à LGPD, configura-se como uma ferramenta funcional e pronta para aplicação em cenários reais. A arquitetura desacoplada adotada, na qual o modelo de IA opera de forma independente do *backend* principal, favorece ainda a atualização contínua dos modelos sem comprometimento da operação do sistema, sendo um aspecto fundamental para sustentabilidade tecnológica em plataformas de Saúde Digital [13].

Apesar dos resultados positivos de desempenho da ferramenta, é importante destacar algumas limitações do presente estudo. O balanceamento da base de dados por *undersampling*, embora necessário para mitigar o desequilíbrio entre classes, reduziu o número total de amostras disponíveis para treinamento, o que pode limitar a capacidade de generalização dos modelos em populações externas com distribuições prognósticas distintas. Estudos de *transfer learning* em imagens médicas indicam que conjuntos de dados inferiores a 2.000 amostras impõem desafios relevantes à generalização, mesmo quando estratégias de regularização e aumento de dados são adotadas [25]. Além disso, a utilização de imagens bidimensionais geradas a partir de escaneamentos tridimensionais, ainda que metodologicamente justificável pela padronização, implica perda parcial de informações morfológicas espaciais potencialmente relevantes para a discriminação entre categorias intermediárias. Abordagens que processam diretamente modelos tridimensionais têm demonstrado resultados promissores na classificação de anomalias craniofaciais em pacientes com FLP [10,12], e sua incorporação futura à GoslonAI pode representar uma direção de evolução da plataforma. Destaca-se ainda que a base de dados, embora multicêntrica, está circunscrita a centros brasileiros de referência, o que pode não refletir a heterogeneidade fenotípica de populações com FLP em outros contextos geográficos e étnicos. Desse modo, como proposta para estudos futuros, indica-se a utilização de amostras mais amplas, validação externa prospectiva e exploração de arquiteturas capazes de processar diretamente representações tridimensionais.

Por fim, os achados deste estudo reforçam o potencial da IA como ferramenta auxiliar na padronização diagnóstica e no apoio à tomada de decisão ortodôntica em pacientes com FLP, visto que a variabilidade diagnóstica interexaminadores é um problema amplamente documentado na avaliação pelo índice de GOSLON [4,5,19,20]. A integração de ferramentas como a GoslonAI a estratégias de Saúde Digital voltadas ao Sistema Único de Saúde (SUS) representa, nesse sentido, uma direção promissora para o fortalecimento do cuidado longitudinal de pacientes com anomalias craniofaciais no Brasil, alinhando-se ao movimento global de incorporação responsável e explicável da IA na prática clínica [7,8,15].

5 CONCLUSÃO

Os resultados demonstraram que CNNs podem atuar como ferramenta de apoio à decisão clínica, contribuindo para maior padronização, agilidade e reprodutibilidade das avaliações ortodônticas. Além do desempenho preditivo satisfatório, a plataforma integra recursos translacionais voltados à Saúde Digital, incluindo sistema web responsivo, geração automatizada de relatórios e mecanismos de interpretabilidade por Grad-CAM. Apesar das limitações relacionadas ao tamanho e à heterogeneidade da base de dados, a solução apresenta potencial para aplicação em centros especializados, apoio remoto ao diagnóstico e futura integração a estratégias digitais voltadas ao SUS e ao cuidado de pacientes com anomalias craniofaciais.

Contribuição dos autores. Os autores contribuíram igualmente em todas as etapas de construção do artigo: concepção e delineamento do estudo; coleta, análise e interpretação dos dados; redação e revisão crítica do manuscrito; e aprovação da versão final.

Agradecimentos. Os autores agradecem aos funcionários do Serviço de Fissuras Labiopalatinas do HULW da UFPB, pelo suporte operacional e pela contribuição à coleta e organização dos dados que tornaram este estudo possível.

Financiamento. Este trabalho foi realizado com o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), processo nº 445115/2023-8, no âmbito da Convocatória CNPq nº 21/2023: Estudos Transdisciplinares em Saúde Pública.

Disponibilidade de dados e materiais. Os autores não declararam a disponibilidade dos dados e materiais da pesquisa.

Conflitos de interesse. Os autores não declararam conflitos de interesse.

Declaração de uso de IA. Ferramentas de IA generativa foram utilizadas exclusivamente para apoio à revisão textual, reorganização estrutural e sumarização do manuscrito. Todo o conteúdo foi revisado e validado pelos autores, que assumem integral responsabilidade pelas informações apresentadas.

REFERÊNCIAS

1. World Health Organization. Global strategies to reduce the health-care burden of craniofacial anomalies [Internet]. Geneva: WHO; 2002 [citado 2026 nov 24]. Disponível em: https://www.paho.org/en/documents/global-strategies-reduce-health-care-burden-craniofacial-anomalies-report-who-meetings
2. Marazita ML, Mooney MP. Current concepts in the embryology and genetics of cleft lip and cleft palate. Clin Plast Surg. 2004;31(2):125-40. doi: https://doi.org/10.1016/S0094-1298(03)00138-X
3. Dixon MJ, Marazita ML, Beaty TH, Murray JC. Cleft lip and palate: understanding genetic and environmental influences. Nat Rev Genet. 2011;12(3):167-78. doi: https://doi.org/10.1038/nrg2933
4. Mars M, Plint DA, Houston WJ, Bergland O, Semb G. The Goslon Yardstick: a new system of assessing dental arch relationships in children with unilateral clefts of the lip and palate. Cleft Palate J. 1987;24(4):314–322.
5. Haque S, Khamis MF, Alam MK, Wan Ahmad AWM. The Assessment of 3D Digital Models Using GOSLON Yardstick Index: Exploring Confounding Factors Responsible for Unfavourable Treatment Outcome in Multi-Population Children With UCLP. Front Pediatr. 2021;9:646830. doi: https://doi.org/10.3389/fped.2021.646830
6. Nollet PJ, Katsaros C, Van't Hof MA, Kuijpers-Jagtman AM. Treatment outcome in unilateral cleft lip and palate evaluated with the GOSLON yardstick: a meta-analysis of 1236 patients. Plast Reconstr Surg. 2005;116(5):1255-62. doi: https://doi.org/10.1097/01.prs.0000181652.84855.a3
7. Castiglioni I, Rundo L, Codari M, Di Leo G, Salvatore C, Interlenghi M, et al. AI applications to medical images: From machine learning to deep learning. Phys Med. 2021;83:9-24. doi: https://doi.org/10.1016/j.ejmp.2021.02.006
8. Huqh MZU, Abdullah JY, Wong LS, Jamayet NB, Alam MK, Rashid QF, et al. Clinical Applications of Artificial Intelligence and Machine Learning in Children with Cleft Lip and Palate-A Systematic Review. Int J Environ Res Public Health. 2022;19(17):10860. doi: https://doi.org/10.3390/ijerph191710860
9. Almoammar KA. Harnessing the Power of Artificial Intelligence in Cleft Lip and Palate: An In-Depth Analysis from Diagnosis to Treatment, a Comprehensive Review. Children (Basel). 2024;11(2):140. doi: https://doi.org/10.3390/children11020140
10. Miranda F, Choudhari V, Barone S, Anchling L, Hutin N, Gurgel M, et al. Interpretable artificial intelligence for classification of alveolar bone defect in patients with cleft lip and palate. Sci Rep. 2023;13(1):15861. doi: https://doi.org/10.1038/s41598-023-43125-7
11. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. 2014.
12. Tan M, Le Q. EfficientNetV2: smaller models and faster training. Proc ICML. 2021;10096-10106.
13. Newman S. Building Microservices. 2nd ed. O'Reilly Media; 2021.
14. Shaikhina T, Khovanova NA. Handling limited datasets with neural networks in medical applications: A small-data approach. Artif Intell Med. 2017;75:51-63. doi: https://doi.org/10.1016/j.artmed.2016.12.003
15. Litvinov A, Ushakov E, Senotrusova S, Lukianov K, Markin Y, Mikhailova L, et al. Clinically Aware Learning: Ordinal Loss Improves Medical Image Classifiers. J Clin Med. 2026;15(1):365. doi: https://doi.org/10.3390/jcm15010365
16. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Int J Comput Vis. 2020;128(2):336-59. doi: https://doi.org/10.1007/s11263-019-01228-7
17. Brasil. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD) [Internet]. Diário Oficial da União. 2018 ago 15 [citado 2026 out 18]. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm
18. Brigato L, Iocchi L. A Close Look at Deep Learning with Small Data. In: 2020 25th International Conference on Pattern Recognition (ICPR). [s.l.]: IEEE; 2021. p. 2490-7. doi: https://doi.org/10.1109/icpr48806.2021.9412492
19. Wiedel AP, Chaudhry M, Svensson H, Becker M. The 5-year-olds' index, the GOSLON Yardstick index, and the modified Huddart/Bodenham index among children with complete unilateral cleft lip and palate: A methodological study. Int J Orthod Rehabil. 2018;9(3):101. doi: https://doi.org/10.4103/ijor.ijor_11_18
20. Saele PK, Daltveit DS, Bjering R, Klepsland T, Torgersbråten NE, Kubon C, et al. Reliability of clinical orthodontic indicators in the Norwegian Registry of Cleft Lip and Palate. BMC Oral Health. 2025;25(1):1763. doi: https://doi.org/10.1186/s12903-025-07043-6
21. Lu C, Angelopoulos AN, Pomerantz S. Improving Trustworthiness of AI Disease Severity Rating in Medical Imaging with Ordinal Conformal Prediction Sets. In: Lecture Notes in Computer Science. Cham: Springer Nature Switzerland; 2022. p. 545-54. doi: https://doi.org/10.1007/978-3-031-16452-1_52
22. Kavousinejad S, Ameli-Mazandarani Z, Behnaz M, Ebadifar A. A Deep Learning Framework for Automated Classification and Archiving of Orthodontic Diagnostic Documents. Cureus. 2024;16(12):e76530. doi: https://doi.org/10.7759/cureus.76530
23. Surendran A, Daigavane P, Shrivastav S, Kamble R, Sanchla AD, Bharti L, et al. The Future of Orthodontics: Deep Learning Technologies. Cureus. 2024;16(6):e62045. doi: https://doi.org/10.7759/cureus.62045
24. Oztekin F, Katar O, Sadak F, Yildirim M, Cakar H, Aydogan M, et al. An Explainable Deep Learning Model to Prediction Dental Caries Using Panoramic Radiograph Images. Diagnostics (Basel). 2023;13(2):226. doi: https://doi.org/10.3390/diagnostics13020226
25. Romero M, Interian Y, Solberg T, Valdes G. Targeted transfer learning to improve performance in small medical physics datasets. Med Phys. 2020;47(12):6246-56. doi: https://doi.org/10.1002/mp.14507

NOTAS EDITORIAIS

O Journal of Health Informatics é uma publicação científica mantida pela Sociedade Brasileira de Informática em Saúde, associação científica sem fins lucrativos dedicada ao desenvolvimento da Informática em Saúde e da Saúde Digital no Brasil. O processo editorial é conduzido com independência científica, sem interferência institucional, comercial ou de patrocinadores, e segue as boas práticas internacionais de ética em publicação científica.



Este artigo foi avaliado por pares, em sistema duplo-cego, por especialistas da área. A equipe editorial usou ferramentas de inteligência artificial, com supervisão humana, na triagem inicial, na conferência do texto, das referências e dos números e, quando necessário, no redesenho de figuras, quadros e tabelas, aprovado pelos autores na prova. As ferramentas e as salvaguardas estão na Política de uso de inteligência artificial da revista. As decisões editoriais foram tomadas por pessoas.