UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE ODONTOLOGIA DE PIRACICABA

ANNA LUIZA DAMACENO ARAUJO

MACHINE LEARNING APLICADA ÀS INVESTIGAÇÕES NOS CAMPOS DA ESTOMATOLOGIA E DA PATOLOGIA ORAL E MAXILOFACIAL

MACHINE LEARNING APPLIED TO RESEARCH IN STOMATOLOGY AND ORAL AND MAXILLOFACIAL PATHOLOGY FIELDS

Piracicaba
2022

ANNA LUIZA DAMACENO ARAUJO

MACHINE LEARNING APLICADA ÀS INVESTIGAÇÕES NOS CAMPOS DA ESTOMATOLOGIA E DA PATOLOGIA ORAL E MAXILOFACIAL

MACHINE LEARNING APPLIED TO RESEARCH IN STOMATOLOGY AND ORAL AND MAXILLOFACIAL PATHOLOGY FIELDS

Tese apresentada à Faculdade de Odontologia de Piracicaba da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Doutora em Estomatopatologia, na Área de Estomatologia.

Thesis presented to the Piracicaba Dental School of the University of Campinas in partial fulfillment of the requirements for the degree of Doctor in Stomatopathology, in Stomatology area

Orientador: Prof. Dr. Alan Roger dos Santos Silva

Coorientador: Prof. Dr. Matheus Cardoso Moraes

Este exemplar corresponde a versão final da tese defendida pela aluna Anna Luiza Damaceno Araujo e orientada pelo Prof. Dr. Alan Roger dos Santos Silva e Prof. Dr. Matheus Cardoso Moraes.

Piracicaba
2022

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Faculdade de Odontologia de Piracicaba
Marilene Girello - CRB 8/6159

Ar15m

Araujo, Anna Luiza Damaceno, 1990-
Machine Learning aplicada às investigações nos campos da Estomatologia e da Patologia Oral e Maxilofacial / Anna Luiza Damaceno Araujo. – Piracicaba, SP : [s.n.], 2022.

Orientador: Alan Roger dos Santos Silva.
Coorientador: Matheus Cardoso Moraes.
Tese (doutorado) – Universidade Estadual de Campinas, Faculdade de Odontologia de Piracicaba.
Em regime interinstitucional com: Instituto de Ciência e Tecnologia da Universidade Federal de São Paulo.

1. Microscopia. 2. Análise. 3. Boca - Doenças. I. Santos-Silva, Alan Roger, 1981-. II. Moraes, Matheus Cardoso. III. Universidade Estadual de Campinas. Faculdade de Odontologia de Piracicaba. IV. Título.

Informações Complementares

Título em outro idioma: Machine Learning applied to research in Stomatology and Oral and Maxillofacial Pathology fields
Palavras-chave em inglês:
Microscopy
Analysis
Mouth - Diseases
Área de concentração: Estomatologia
Titulação: Doutora em Estomatopatologia
Banca examinadora:
Alan Roger dos Santos Silva [Orientador]
Márcio Ajudarte Lopes
Luiz Paulo Kowalski
César Andrés Rivera Martínez
Gleyson Kleber do Amaral Silva
Data de defesa: 26-09-2022
Programa de Pós-Graduação: Estomatopatologia

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-3725-8051
- Currículo Lattes do autor: http://lattes.cnpq.br/0633932030080115

UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Odontologia de Piracicaba

A Comissão Julgadora dos trabalhos de Defesa de Tese de Doutorado, em sessão pública realizada em 26 de setembro de 2022, considerou a candidata ANNA LUIZA DAMACENO ARAUJO aprovada.

PROF. DR. ALAN ROGER DOS SANTOS SILVA

PROF. DR. CÉSAR ANDRÉS RIVERA MARTÍNEZ

PROF. DR. GLEYSON KLEBER DO AMARAL SILVA

PROF. DR. LUIZ PAULO KOWALSKI

PROF. DR. MÁRCIO AJUDARTE LOPES

A Ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

AGRADECIMENTOS

À Universidade Estadual de Campinas, na pessoa do Magnífico Reitor, Prof. Dr. Antônio José de Almeida Meirelles.

À Faculdade de Odontologia de Piracicaba, na pessoa de seu Diretor, Prof. Dr. Flávio Henrique Baggio Aguiar.

Ao Prof. Dr. Valentim Adelino Ricardo Barão, Coordenador Geral da Pós-Graduação da Faculdade de Odontologia de Piracicaba.

Ao Coordenador do Programa de Pós-Graduação em Estomatopatologia, Prof. Dr. Pablo Agustin Vargas, que apóia incondicionalmente todos os alunos em suas jornadas acadêmicas, viabilizando o acesso à infraestrutura necessária, apoiando redes de interação entre pesquisadores de múltiplos centros de ensino e estimulando a independência do pesquisador.

Ao meu orientador, Prof. Dr. Alan Roger dos Santos Silva, responsável direto pela minha evolução como pesquisadora e ser humano. Prof. Alan acreditou em mim quando me faltavam motivos para acreditar em mim mesma e essa atitude me tornou capaz de realizar trabalhos e atingir metas que me pareceriam impossíveis em outras situações. Agradeço pela confiança cega, preocupação genuína, dedicação desmedida e inabalável paciência durante todos os anos que estive sob sua orientação. Sinto que poderei contar sempre com seu apoio, que se estende além dos limites da Academia e fortalece geração após geração de orientandos que se unem em torno de um mesmo objetivo: impulsionar a Pesquisa Brasileira. Ao Prof. Alan, agradeço também as possibilidades inigualáveis de interação profissional que culminaram na realização de trabalhos inovadores e parcerias importantes com as equipes da University of Sheffield na pessoa do Prof. Dr Ali Khurram e da University of Chicago, na pessoa do Prof. Dr. Alex T. Pearson, duas personalidades proeminentes no campo de pesquisa relacionada à Inteligência Artificial para diagnóstico de câncer de cabeça e pescoço. À eles sou grata pela oportunidade de aprendizado, interação e parceria.

Ao meu co-orientador Prof. Dr. Matheus Cardoso Moraes, que me acolheu sem ressalvas e dedicou parte importante do seu tempo ao meu treinamento por meio de transferência de aprendizado. A partir de suas aulas, eu pude ajustar os conceitos que aprendi de forma não supervisionada e, a partir do erro, aprendi a correlacionar corretamente a Patologia e a Engenharia, potencializando a minha abilidade de dicernimento. Poucos profissionais possuem a visão única que o Prof Matheus ostenta e eu sou muito grata por

poder interagir e aprender tanto com ele e sua equipe, nas pessoas de Viviane Mariano, Maíra Kudo, Giovanna Calabrese e Erin Crespo, que são profissionais dedicadíssimas e atentas, sempre dispostas a desenvolver soluções para nossas propostas.

Ao Prof. Dr. Márcio Ajudarte Lopes, que é uma figura paterna para todos nós, sempre preocupado com o bem estar de todos e sempre disposto a nos acolher nos momentos em que ele sabe que nossa família seria importante. Ao Prof. Márcio também agradeço pelo apoio e confiança que são tão característicos dele.

Ao Prof. Dr. Luiz Paulo Kowalski que tem nos apoiado incondicionalmente nos últimos anos e que, numa iniciativa sem precedentes, reuniu um proeminente grupo de pesquisa em torno de um objetivo maior: a produção de pesquisa e disseminação de conhecimento a cerca de sistemas baseados em Inteligência Artificial. Sou extremamente grata por poder continuar fazendo pesquisa sob sua supervisão.

Aos amigos da Pós-graduação, que me apóiam frente a todos os obstáculos e me incentivam desproporcionalmente. Obrigada por serem minha motivação diária para seguir em frente e minha rede de apoio. Graças aos nossos momentos juntos, a vida se enche de amor, amenizando a ansiedade e a saudade.

Ao meus pais, José de Araújo Filho e Terezinha Damaceno Araújo, e minhas irmãs, Ana Cláudia e Ana Paula, que me apoiam de forma incondicional e suportam a ausência por anos. Sem o apoio de vocês e de nossas famílias eu não estaria aqui hoje.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) - Código de Financiamento 001, com apoio da Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), processo nº 159067/2018-9, e com o apoio do Programa Institucional de Internacionalização (DSE-CAPES/PrInt), processo nº 88887.369713/2019-00.

RESUMO

Introdução: A incorporação de métodos de processamento de imagem e inteligência artificial (IA) na área médica está se desempenhando um papel transformador na medicina personalizada. Portanto, é importante reconhecer a natureza multidisciplinar das equipes envolvidas no desenvolvimento e implementação de novas abordagens de Machine Learning (ML) para câncer de cabeça e pescoço (CCP) e promover uma comunicação eficiente entre patologistas, médicos, cirurgiões e cientistas da computação. Em patologia oral e maxilofacial, pesquisadores enfatizam o desenvolvimento de modelos de ML para diagnóstico e prognóstico de desordens orais potencialmente malignas (DOPM) e carcinoma espinocelular (CEC) oral, um desafio frente aos sistemas subjetivos de classificação de displasia epitelial oral (DEO). Além disso, Modelos Preditivos (MPs) baseados em características radiômicas, clínicas, patológicas, dosimétricas para prever toxicidades em pacientes submetidos ao tratamento de CCP têm sido propostos, e alguns estudos demonstram que MPs baseados em imagem não apresentam desempenho superior aos modelos convencionais sem biomarcadores de imagem (BMIs). *Objetivos:* O presente estudo produziu evidência acerca de quatro objetivos principais: 1) ampliar a compreensão de patologistas orais, estomatologistas e cirurgiões de cabeça e pescoço sobre as abordagens diagnósticas baseadas em IA, com foco especial em Convolutional Neural Networks (CNNs), sintetizando fundamentos teóricos e conceituais; 2) implementar sete arquiteturas de Deep Learning (DL) para gradação de DEO em imagens histopatológicas de lâminas digitalizadas; 3) avaliar a percepção de clínicos ao avaliar DOPM que possa prejudicar desenvolvimento de modelos de DL; e 4) fazer uma revisão sistemática (SR) sobre modelos de ML atualmente usados para prever toxicidades relacionadas ao tratamento de CCP de modo a avaliar as evidências sobre o impacto de BMIs em PMs. *Métodos:* 1) Para a presente revisão conceitual, os autores resumiram e contextualizaram conceitos importantes comuns ao campo da engenharia para melhorar a compreensão de tais conceitos por patologistas e clínicos. 2) Para o estudo primário com foco em DEO, uma coorte de 82 pacientes (98 WSI corados por H&E) de três centros brasileiros foi recuperada, anotada, segmentada em patches e augmentada, gerando um total de 81.786 patches para treinar sete CNNs (ResNet50, InceptionV3, VGG16, Xception, MobileNet, DenseNet e EfficientNetB0) para a classificação de baixo risco (LR) e alto risco (HR) de malignização de acordo com o Sistema Binário. 3) Para avaliar a percepção dos clínicos sobre DOPM, 46 imagens de leucoplasias foram classificadas às cegas e anotadas por três observadores, e a concordância interobservador foi avaliada por meio de Fleiss Kappa e a média pixel-wise Intersection Over Union (IoU). 4) Para a RS, o acrônimo PICOS foi usado para orientar a questão da RS (os MPs podem prever com precisão as toxicidades do tratamento do CCP e os critérios de elegibilidade. A busca eletrônica em banco de dados abrangeu PubMed, EMBASE, Scopus, Cochrane Library, Web of Science e LILACS. O risco de viés (RoB) foi avaliado por meio da ferramenta PROBAST e os resultados foram sintetizados com base no formato dos dados (com e sem BMIs) para permitir a comparação das duas modalidades de dados. *Resultados: 1) Uma* revisão conceitual detalhada sobre terminologias e metodologia foi apresentada. *2)* Os resultados do estudo primário demonstraram que quase todos os modelos apresentaram uma alta taxa de aprendizado, mas um potencial de generalização muito baixo. No desenvolvimento do modelo, a VGG16 teve o melhor desempenho, mas apresentou *overfitting*. A EfficientB0 possui métricas comparáveis e a menor *loss* entre todas as CNNs, sendo ótima candidata para novos estudos. 3) A concordância interobservador para a análise de OPMD foi substancial e moderada, e a média de IoU foi de 0,675 (±0,030 std). 4) Um total de 28 estudos (4.713 pacientes) foram incluídos da RS. Um alto risco de viés foi identificado em 23 estudos. A meta-análise (MA) não mostrou diferença entre os modelos com e sem BMI. *Conclusão:* 1) O presente trabalho apresenta uma visão interdisciplinar privilegiada sobre técnicas de ML para processamento de imagens. A implementação de modelos de IA e visão computacional para reconhecimento de padrões na análise de imagens clínicas e histopatológicas de CCP tem o potencial de auxiliar no diagnóstico e na previsão prognóstica. 2) A avaliação comparativa das arquiteturas para classificação de DEO indicou que os modelos não foram capazes de generalizar o aprendizado usando a presente metodologia de anotação. 3) A percepção dos clínicos pode introduzir viés nas anotações utilizadas para treinar modelos de DL. 4) A heterogeneidade dos estudos incluídos na RS, bem como as métricas não padronizadas, impedem a comparação adequada dos estudos, e a ausência de um teste independente/externo não permite avaliar a capacidade de generalização do modelo.

Palavras-chave: Microscopia; Análise; Boca - Doenças.

ABSTRACT

Introduction: The incorporation of image processing methods and artificial intelligence (AI) in the medical field is revolutioning the personalized medicine. Therefore, it is important to recognize the multidisciplinary nature of teams involved in the development and implementation of new Machine Learning (ML) image-based approaches for head and neck cancer (HNC) and to promote efficient communication among oral pathologists, oral medicinists, head and neck surgeons and computer scientists. In oral and maxillofacial pathology, researchers give special emphasis to the development of ML models for oral potentially malignant disorders (OPMD) and oral squamous cell carcinoma (OSCC) diagnosis and prognosis, which is a great challenge since dysplasia grading systems for oral epithelial dysplasia (OED) are a source of disagreement among pathologists. Additionally, multivariable prediction models (PMs) based on radiomic features and non-imaging data (clinical, pathological, dosimetric features) to predict toxicities in patients that underwent HNC treatment have been proposed, and some studies demonstrates that image-based PMs does not outperform conventional models without image biomarkers (IBMs). Aims: The present study aim to accomplish four main objectives: 1) to enhance the comprehension of oral pathologists, oral medicinists and head and neck surgeons regarding to AI-based diagnostic approaches, with a special focus on CNNs, by summarizing theoretical and conceptual foundation; 2) to implement seven state-of-the-art Deep Learning (DL) architectures for oral epithelial dysplasia grading in histopathological images from wholes slide images (WSI); 3) to assess clinician's perception on OPMD that can limit DL approaches; and 4) to do a systematic review (SR) on Machine Learning (ML) models currently used to predict HNC treatment-related toxicities to assess the evidence regarding the impact of IBMs in PMs. Methods: 1) For the present conceptual review, the authors summarized and contextualized important concepts common to the engineering field to enhance the understanding of such concepts by pathologists and clinicians. 2) For the primary study, a cohort of 82 patients (98 H&E-stained WSI) with OED from three brazilian centers were retrieved, annotated, segmented in patches and augmented. A total of 81,786 patches were used to train seven CNNs (ResNet50, InceptionV3, VGG16, Xception, MobileNet, DenseNet, and EfficientNetB0) for the classification of OED in low risk (LR) and high risk (HR) according to the Binary System for malignization risk. 3) To assess clinician's perception, 46 images were blindly classified and annotated by three observers, and interobserver concordance was evaluated through Fleiss Kappa and mean pixel-wise Intersection Over Union (IoU). 4) For the SR, the acronym PICOS was used to orientate the focused review question (PMs can accurately predict HNC treatment toxicities?) and the eligibility criteria. Electronic database search encompassed PubMed, EMBASE, Scopus, Cochrane Library, Web of Science, and LILACS. Risk of Bias (RoB) was assessed through PROBAST and the results were synthesized based on the data format (with and without IBMs) to allow comparison. Results: 1) A conceptual review focusing on terminology and methodologies is presented. 2) Results from the primary study demonstrated that almost all of the models presented a high learning rate, yet very low generalization potential. At the model development, VGG16 performed the best, but with massive overfitting. EfficientB0 has comparable metrics and the lowest loss among all CNNs, being a great candidate for further studies. 3) The interobserver agreement was substantial and moderate, with mean IoU of 0.675 (±0.030 std). 4) A total of 28 studies and 4,713 patients were included in the SR. Xerostomia was the most frequently investigated toxicity. A high RoB was identified in 23 studies. Meta-analysis (MA) showed no difference among IBM- and non-IBM-based models. Conclusion: 1) The present work presents a privileged interdisciplinary view on ML techniques for image processing. The implementation of AI models for clinical and histopathological image analysis of HNC has the potential to aid diagnosis and prognostic prediction. 2) The comparative assessment of DL architectures for OED classification indicated that the models were not able to generalize using the present annotation methodology, due an overlapping of features between the two classes, which can be a confounding factor for the CNN training. 3) Clinicians' perception can introduce bias in the ground truth annotations used to train a DL model. Additionally, 4) the heterogeneity of the studies included in the SR, as well as non-standardized metrics, prevent proper comparison of studies, and the absence of an independent/external test does not allow the evaluation of the model's generalization ability.

Keywords: Microscopy; Review; Mouth - Diseases.

SUMÁRIO

1 INTRODUÇÃO

Inteligência Artificial (IA) é uma definição guarda-chuva para sistemas que podem reproduzir a inteligência humana e que tem sido utilizados com sucesso na análise de imagens médicas. Nesse domínio, o desenvolvimento de algoritmos para análise de imagens tem motivado a utilização desses sistemas para diagnóstico objetivo, precoce e preciso do câncer. Machine Learning (ML) é uma subárea da IA que desenvolve e aplica algoritmos treinados para resolver problemas de reconhecimento de padrões sem serem explicitamente programados. Dentro desta subárea, as modalidades de treinamento são definidas de acordo com a rotulagem dos dados como aprendizado supervisionado (dados rotulados), *weekly supervised*, e aprendizado não supervisionado (Krohn et al., 2019; Zhang et al.,2021).

No pool de algoritmos de aprendizado supervisionado podem ser enquadrados diversos algoritmos clássicos lineares e não lineares (alguns reposicionados da estatística convencional) quanto algoritmos mais modernos que se encaixam no conceito de *Representation Learning* (RL), que é uma modalidade de aprendizado na qual algoritmos identificam representações consistentes diretamente nos dados (do inglês *feature learning*). O *Perceptron,* neurônio artificial na sua forma mais primitiva, a *Multilayer Perceptron* e as Redes Neurais Convolucionais (do inglês *Convolutional Neural Network* (CNN)] são exemplos de algoritmos de RL referenciados como modernos (Krohn et al., 2019; Zhang et al.,2021; McCulloch and Pitts, 1990; Rosenblatt, 1998).

A *Multilayer Perceptron* (MLP) é uma rede neural artificial [do inglês *Artificial Neural* network (ANN)] composta por uma camada de entrada (composta essencialmente pelos dados) e duas camadas de nós *Perceptron* (uma camada oculta e uma camada de saída), na qual a camada de entrada está ligada ao número de atributos necessários para definir as classes, que são orientadas pela rotulação dos dados. Dessa forma, se duas características - formato e tamanho nuclear – são atributos para a classificação, são necessárias duas camadas de entrada conectadas aos neurônios da camada oculta, os quais conduzirão interações cruzadas entre os atributos nas camadas de entrada e ocultas para compor o valor de saída que aponta a probabilidade de uma amostra pertencer à uma determinada classe.

Em *computer vision* aplicada à Patologia, o desenvolvimento e implementação de uma MLP requer uma etapa prévia de engenharia de atributos, na qual cientistas da computação extraem os dados necessários (ex., circunferência do núcleo, distância das margens do núcleo para as margens da membrana celualr) que serão então uttilizados para o treinamento da MLP. As CNNs, ao contrário, são ANN com a adição de uma camada convolucional para a extração

automática desses atributos. Durante o treinamento, ANN classificam o conjunto de dados de treinamento (imagens, no presente contexto) repetidamente durante um número predeterminado de épocas (número de vezes que o conjunto de dados passa pela CNN). Os elementos estruturais predefinidos pelo cientista de dados antes do treinamento são chamados de hiperparâmetros (ex., número de neurônios, o número de *kernels* e a camada convolucional) e os elementos que são reajustados pela ANN durante o treinamento, como os pesos e a função *kernel*, são chamados de parâmetros. O processo de treinamento da CNN é dividido em 3 etapas explicadas a seguir.

A etapa *Feed-forward* é definida pela imagem de entrada sendo apresentada à rede. As camadas convolucionais são compostas por combinações de operações de convolução e ativação (conv+activ), agrupamento [do inglês *polling* (Pol)] e achatamento [do inglês *flatenning* (Flat)] na última camada convolucional. A camada convolucional é uma função de correlação que calcula semelhanças entre características específicas relacionadas à classe das imagens e alguns filtros ou *kernels* (k), potencializando o reconhecimento de características viáveis através da correlação entre uma imagem e um *kernel*. O cálculo resulta em coeficientes de correlação cruzada junto com o sinal ou imagem. Quanto maiores os coeficientes, maior a correlação (semelhança) entre a máscara e a função. Enquanto a correlação é usada para extração de recursos, a convolução é essencialmente uma função de filtragem. Em resumo, enquanto a camada convolucional extrai as características mais discriminativas, a camada *fully connected* atualiza os pesos. A operação de *polling* reduz o tamanho do mapa de ativação resultante (redução de dimensionalidade), mantendo as características importantes da imagem. Sequencialmente, a operação de nivelamento decompõe os mapas de ativação filtrados e redimensionados em vetores. Os vetores são distribuições de valores das feições extraídas que podem variar, por exemplo, de 0 a 1. Por fim, esses vetores passam pela MLP para serem classificados (Krohn et al., 2019; Zhang et al.,2021). Na etapa de *Backpropagation*, um processo de otimização calcula o erro e=|y-yd | entre a saída (y) e o rótulo (yd). Em uma visão cartesiana, esse processo de treinamento é representado pelo ajuste do hiperplano no espaço de atributos. A configuração preenche o espaço de recursos com amostras de conjuntos de dados e o hiperplano os separa de acordo com a classe determinada (por exemplo, classe 1 = câncer, classe 0 = não câncer). (Krohn et al., 2019; Zhang et al.,2021).

Em termos práticos, MLP e CNNs aprendem a partir do erro e=|y-yd |, que representa a diferença entre a saída (y) e o rótulo (yd). No início do treinamento, uma ANN não treinada

classifica incorretamente as amostras e o ajuste do gradiente dos erros durante o treinamento é usado para a terceira etapa, que envolve a atualização dos parâmetros da rede – especificamente, os pesos (ωi) e o *bias* (θ) para MLP e a função *kernel* (k) para as camadas convolucionais. O ajuste dos pesos ocorre para equilibrar o contribuição de cada atributo, melhorando a classificação. Essa capacidade de atualizar os parâmetros a partir dos erros (verificando rótulos) até atingir uma classificação ótima é o que interpretamos como aprendizado supervisionado (Rosenblatt, 1998). Em outras palavras, a função de *loss* mede o quão errado o modelo está classificando, permitindo atualizar os pesos na direção correta do gradiente desta função para o menor valor da função de loss, otimizando o treinamento. Ao final de cada época, é realizada uma etapa de validação para verificar os parâmetros e hiperparâmetros e avaliar o potencial de aprendizado. Ao final do treinamento, é criado um mapa de ativação contendo recursos para discriminar classes. Essa correspondência entre máscara e imagem explica por que as CNNs são classificadas como modelos de aprendizado profundo [do inglês *Deep Learning* (DL)], pois aprendem "diretamente a partir dados".

A revisão de literatura apresentada refere-se à uma interpretação adaptada ao conceito de IA aplicado à patologia (Krohn et al., 2019; Zhang et al.,2021), que embasa os próximos capítulos da presente tese, que tem como objetivos principais: 1) ampliar a compreensão de patologistas orais, médicos orais e cirurgiões de cabeça e pescoço sobre as abordagens diagnósticas baseadas em IA, com foco especial nas CNNs, sintetizando fundamentos teóricos e conceituais; 2) implementar sete arquiteturas de Deep Learning (DL) de última geração para gradação de displasia epitelial oral em imagens histopatológicas de lâminas digitalizadas; 3) avaliar a percepção de clínicos ao avaliar DOPM que possa prejudicar desenvolvimento de modelos de DL, e 4) fazer uma revisão sistemática (SR) sobre modelos de ML atualmente usados para prever toxicidades relacionadas ao tratamento de HNC de modo a avaliar as evidências sobre o impacto de BMIs em PMs. Este estudo primário está de acordo com o *Guidelines for less biased data collection and algorithm evaluation* (Marée, 2017) e o TRIPOD statement (Collins et al., 2015; Moons et al., 2015). A presente RS foi conduzida seguindo as diretrizes do *Guidance for defining review question* (CHARMS Checklist) (Moons et al., 2019), do *Guide for SR and meta-analysis of Prediciton Model Studies* (Debray et al., 2017), do *Preferred Reporting Items for Systematic Reviews and Meta-Analysis* (PRISMA) (Page, McKenzie, Bossuyt, et al., 2021; Page, Bossuyt, Boutron, et al., 2021) e da lista de verificação PRISMA-P (Moher et al., 2015; Shamseer et al., 2015).

2 ARTIGOS

2.1 Artigo: Machine Learning Concepts applied to Oral Pathology and Oral Medicine: A Convolutional Neural Networks Approach

Artigo submetido para publicação no periódico Journal of Oral Pathology and Medicine

Anna Luíza Damaceno Araújo[a], Viviane Mariano da Silva[b], Maíra Suzuka Kudo[b], Eduardo Santos Carlos de Souza[h], Marcio Ajudarte Lopes[a], Pablo Agustin Vargas[a], Syed Ali Khurram[c], Alexander T. Pearson[d,e], Luiz Paulo Kowalski[f,g], André Carlos Ponce de Leon Ferreira de Carvalho[h], Alan Roger Santos-Silva[a], Matheus Cardoso Moraes[b].

Affiliation:
[a] Oral Diagnosis Department, Piracicaba Dental School, University of Campinas (UNICAMP), Piracicaba, São Paulo, Brazil.
[b] Institute of Science and Technology, Federal University of São Paulo (ICT-Unifesp), São José dos Campos, São Paulo, Brazil.
[c] Unit of Oral and Maxillofacial Pathology, School of Clinical Dentistry, University of Sheffield, Sheffield, UK.
[d] Section of Hemathology/Oncology, Department of Medicine, University of Chicago, Chicago, IL, USA.
[e] University of Chicago Comprehensive Cancer Center, Chicago, IL, USA.
[f] Department of Head and Neck Surgery and Otorhinolaryngology, A.C. Camargo Cancer Center, São Paulo, Brazil.
[g] Head and Neck Surgery Department and LIM 28, University of São Paulo Medical School, São Paulo, Brazil.
[h] Institute of Mathematics and Computer Sciences of University of São Paulo (ICMC-USP), São Carlos, São Paulo, Brazil

Corresponding Author: Alan Roger Santos-Silva
Oral Diagnosis Department, Piracicaba Dental School, UNICAMP
Adress: Av. Limeira, nº 901, Areião, Piracicaba, São Paulo - Brazil
Postal code: 13414- 903
Phone number: +55 19 21065320
E-mail: alan@unicamp.br

Acknowledgments

The authors would like to gratefully acknowledge Fabiana Facco Casarotti and Fabiane de Athayde for technical processing and image acquisition.

Competing Interests statement: None to declare.

Abstract

Introduction: The incorporation of image processing methods and artificial intelligence (AI) is shaping up to play a transformative role in personalized medicine. These AI models and networks can learn and process dense information in a short time, leading to an efficient, objective and accurate clinical and histopathological analysis, which can inform treatment modalities and improve prognostic outcomes. *Methods:* This paper targets oral pathologists, oral medicinists and head and neck surgeons to provide them with a theoretical and conceptual foundation of AI-based diagnostic approaches, with a special focus on Convolutional Neural Networks (CNNs), the state-of-the-art in AI and Deep Learning (DL). *Conclusion:* The development of these models and computer vision methods for pattern recognition in clinical and histopathological image analysis of head and neck cancer (HNC) has the potential to aid diagnosis and prognostic prediction.

Keywords: Artificial Intelligence; Deep Learning; Artificial Neural Network; Supervised Learning; Oral Cancer; Head and Neck Cancer; Oral Potentially Malignant Disorders

1. Artificial Intelligence basic concepts and approaches

Artificial Intelligence (AI) is an umbrella definition for systems that can reproduce human intelligence and have been successfully used in the analysis of medical imagens. In this domain, the development of algorithms for image analysis have motivated the use of these systems for objective, early and accurate cancer diagnosis.

Machine Learning (ML) is a subarea of AI that develops and applies algorithms to solve pattern recognition problems by learning from the data without being explicitly coded. These algorithms can be used in a data science pipeline, which include other data analysis steps, such as data transformation, data cleaning and feature engineering, which includes the extraction of relevant and measurable features through techniques that can be handcrafted to extract particular aspects of a dataset or learned directly form the data. The information about which feature is relevant or not for classification is not provided for the model, so it undergoes self-learning with better generalization ability than humans when dealing with similar amounts of data[1,2].

When the dataset is composed by images, relevant and measurable "low level" features (i.e., texture, morphometry, color, and histogram) can be manually extracted by the engineer and used as input data to train either a "traditional" ML algorithm (e.g., Random Forest, Support Vector Machine) or a "modern" Artificial Neural Network (ANN), which are based on the structure and functioning of the human nervous system[1,2].

Deep Networks (DNs) are ANNs with more than 2 layers of neurons with added convolutional layers for automatic feature extraction. DNs are trained by Deep Learning (DL) algorithms that automatically identifies consistent representations within the data (Representation Learning). DN models learn "high-level" features directly from the data and improve accuracy from previous "experience" (i.e., extensive data visualization), which is the ultimate emulation of the human brain learning function[1].

This review targets oral pathologists, oral medicinists and head and neck surgeons and aims to provide them with a conceptual explanation of ML algorithms usually employed in human pathology with focus on Convolutional Neural networks (CNN).

2. Artificial neuron, Multilayer Perceptron, and Convolutional Neural Networks

McCulloch and Walter Pitts[3] in 1943 proposed the first unit of an ANN, the McCulloch-Pitts (MCP) neuron, a supervised classifier that emulates the structure and functioning of a biological neuron. The MCP neuron takes Boolean values as input {0,1} and

returns Boolean values as output {0,1} (e.g., 0 = noncancer, 1 = cancer), which gives a binary nature to the decision boundary suitable for binary classification problems. The sum is a linear function, which means that this neuron can only distinguish linearly separable examples.

In 1958, Rosenblatt[4] proposed an artificial neuron unit formally known as Perceptron (**Figure 1),** a mathematical operation that, from a range of input values {-∞ → +∞}, returns Boolean values as output {0,1}. In a practical image context, input values range from 0 to 1 or 0 to 255, which allows classification from a correspondent combination of inputs ($_n$) (e.g., nuclear morphology, size, and color). Essentially, the output values of 0 or 1 are due to the intrinsic Boolean activation function of the Perceptron.

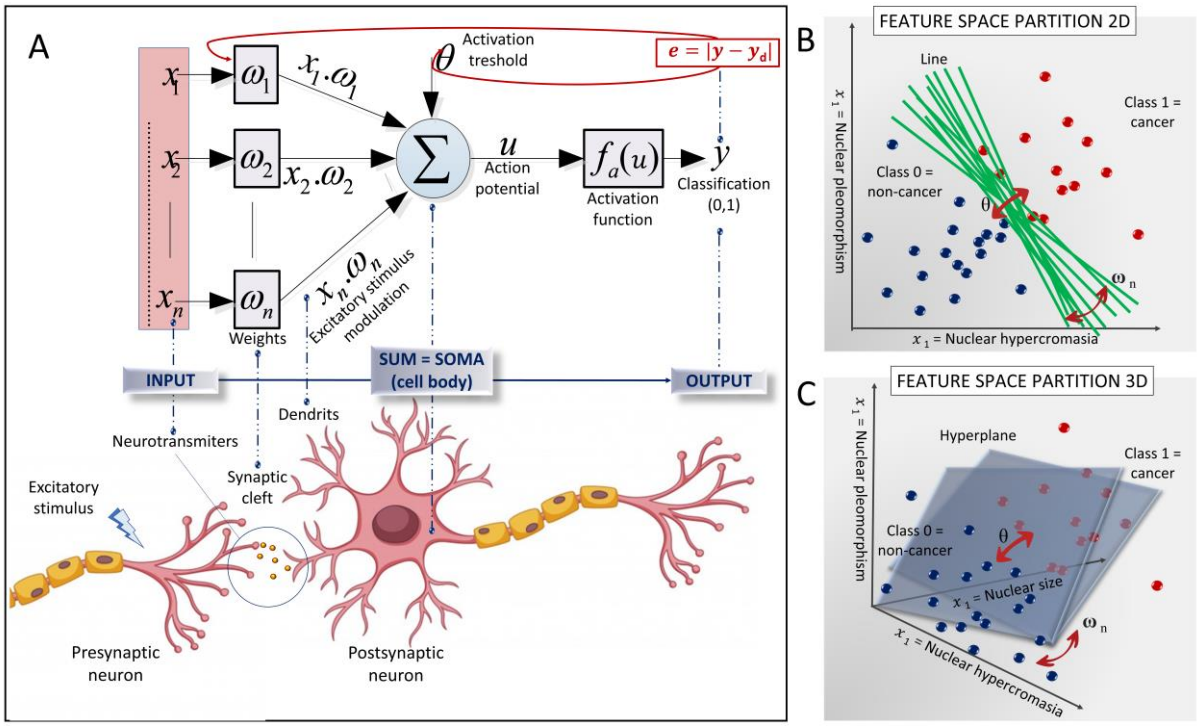


Figure 1: Artificial neuron (Perceptron), biological neuron correlation and feature space partition. (A) Neurotransmitters or input data ($x1$, $x2$…) are presented to the neuron, which is activated by the sum of excitatory stimuli. This means that the neuron node processes image information that results in an output value for classification. The classification error adjusts the weights (ω1, ω2, ω3…) and bias (θ) (B, C) by repositioning the line (2-dimensional) or the hyperplane (3 or n-dimensional). From a mathematical visual perspective, while the weights alter the inclination of the hyperplane, the bias (θ) adjusts the level, dislocating it forward and backward, setting the threshold to better address the class separability.

The input data ($\mathbf{x}_1$, $\mathbf{x}_{2...}$) are equivalents to the neurotransmitters released in the synaptic cleft, which are represented by the weights (ω_1, ω_2, $\omega_{3...}$) in the artificial computer model. To activate the neuron, a sum (Σ) of excitatory stimulus ($\mathbf{x}_i \cdot \boldsymbol{\omega}_i$) – interaction between the neurotransmitters and the gamma-aminobutyric acid (GABA) receptors in the synaptic cleft – leads to the activation threshold or "bias" (θ) to reach a critical stage, which corresponds to the opening of calcium channels in the postsynaptic membrane. The neuron is activated, resulting in the propagation of the action potential (u), an aggregation of excitatory stimulus from the dendrites that flows within the neuron axon. The activation function [$f\alpha(u)$] modulates the received stimulus to the output, which is essentially a value for classification **(Figure 1A)**.

A Multilayer Perceptron (MLP) is a classical ANN composed of an input layer containing the data input and two more layers of perceptron nodes: a hidden layer and an output layer **(Figure 2)**.

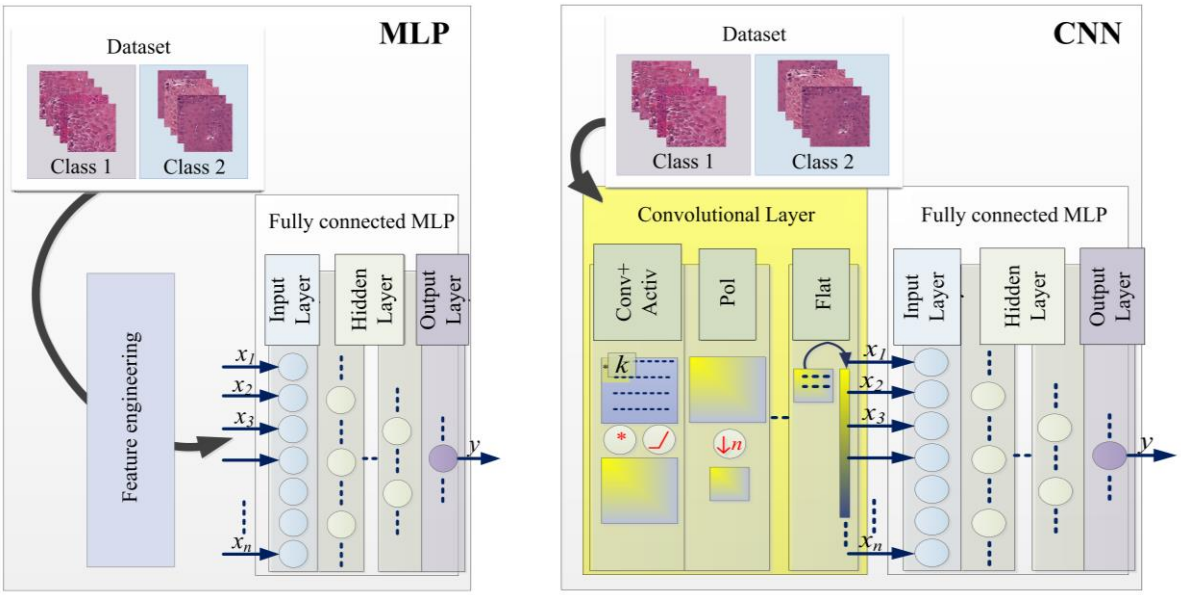


Figure 2: Multilayer Perceptron and Convolutional Neural Network.

The input layer is connected to the number of features necessary to define a class in a due problem. In a computational context, it also defines the feature space partition (FSP), in which the number of descriptive characteristics for a specific classification problem is equivalent to the number of dimensions **(Figures 2B and 2C)**. For a better understanding of the FSP, the following explanation will consider a 2-class problem (e.g., cancer, noncancer) explored within a 2-dimensional FSP (e.g., two features will be taken into account: nuclear

pleomorphism and hyperchromatism) **(Figure 1B)**. Since the features are essentially diverse (morphological and coloring), during the training stage, the weights of each feature are also widely variable, requiring updating and balancing to an optimal value **(Figure 1B – green line)**. The more pleomorphic and hyperchromatic the nuclei are, the more likely it is to be classified as a cancer cell. However, some cells with hyperchromatic nuclei may not represent cancerous cells, as shown in **Figure 1B.** Expanding this example, if the two features are not ideal for proper class separation, a third feature can be used, resulting in a 3-dimensional FSP, as seen **in Figure 1C.** It is worthwhile to note that an MLP for histological image recognition can process numerous features and readjust thousands of parameters (i.e., weights, internal kernel function) during training. Moreover, cross-interactions among all neurons in input and hidden layers form the output value, and since a formal MLP would also return a binary classification, the number of classes of a problem determines the number of neurons that should be used in the output layer's nodes. This means that if there is a 3-class problem, at least 2 output neurons should be used); hence, three combined values (0, 0), (0, 1), and (1, 0) can be used to define each class. However, the most used form of designing the output is using one-hot, one output neuron dedicated to each class; thus, the corresponding three combined values become (0, 0, 1), (0, 1, 0), and (1, 0, 0). In addition, this output structure permit showing the probability of the sample belonging to each class, ex. (0.01, 0.02, 0.97).

The addition of a convolutional layer in an MLP replaces the feature engineering step and composes the basic structure of a CNN **(Figure 2).** The number of neurons, number of kernels, and the convolutional layer are the predefined structural elements called hyperparameters.

3. *CNN training*

During training, the CNN model (convolutional layers and MLP) classifies the training dataset repeatedly during a predetermined number of epochs (number of times the dataset passes through the CNN). The CNN training process is divided into 3 steps, which are further explained.

The ***Feed-forward step*** is defined by the input image being presented to the network. The convolutional layers are composed of combinations of convolution and activation (conv+activ), polling (Pol), and flattening (Flat) operations in the last convolutional layer. The convolutional layer is a correlation function that calculates similarities between specific features related to the class from the images and some filters or kernels (*k*), enhancing the

recognition of feasible features through the correlation between an image and a kernel. The calculation results in coefficients of the cross-correlation along with the signal or image. The higher the coefficients are, the higher the correlation (similarity) among the mask and function. While correlation is used for feature extraction, convolution is essentially a filtering function. In summary, while the convolutional layer extracts the most discriminative features, the fully connected layer update the weights, both occurs during training stage. The polling operation reduces the resultant activation map size (dimensionality reduction) while maintaining the important image features. Sequentially, the flattening operation decomposes the filtered and resized activation maps into vectors. The vectors are distributions of values of the extracted features that can range, for example, from 0 to 1. Finally, these vectors pass through the MLP to be classified.

In the ***Backpropagation step***, an optimization process computes the error $e = |y - y\mathrm{d}|$ between the output (y) and ground truth label ($y\mathrm{d}$)[5]. In a cartesian view, this training process is represented by hyperplane adjustment in the FSP (**figures 2B-green line)** and **(Figure 1C-blue plane**). The setup populates the space of features with dataset samples and the hyperplane separates them according to the determined class (e.g., class 1 = cancer, class 0 = noncancer).

In practical terms, at the beginning of the training, a "naïve" CNN is expected to misclassify samples. From misclassification, the gradient of the errors during training is used to guide the ***update parameters step*** – specifically, weights (ω_i) and bias (θ) for MLP, and kernel function (k) for convolutional layers (**Figure 3A**) – toward balancing the contribution of each extracted feature and weights of neurons and improving classification. This ability to update the parameters from the errors (by checking labels) until reaching an optimum classification is what we interpret as supervised learning[4]. Additionally, at the end of each epoch, a validation step is carried out to check the parameters and/or hyperparameters and to evaluate the potential of learning. During training, the loss function measures how wrong the model is classifying, allowing it to update the weights in the right direction of the gradient of this function towards the lowest value of the loss function.

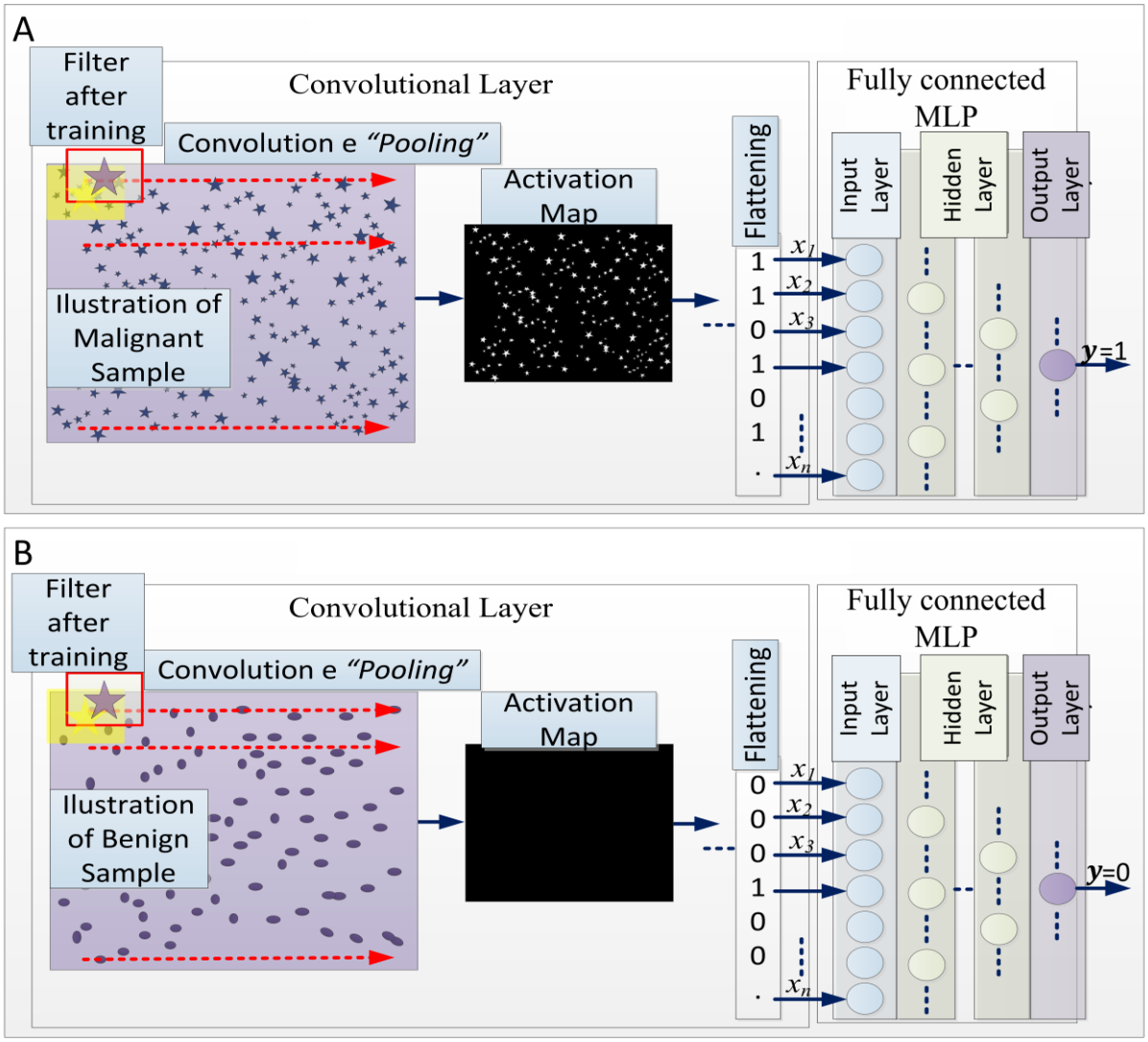


Figure 3: Correlation and convolution. A CNN is trained with 2-class labelled images with benign and malignant histopathological representations. From classification error, the CNN updates its parameters until fully trained and generates a stellated mask that corresponds to malignant features. The mask is automatically generated according to relevant features to maximize class separation, which is the basic process for image feature recognition. (A) A malignant histopathological image being presented to a trained CNN that has a star as a mask. The stellated mask sweeps the images searching for correspondence highlighting cells with similar features. An activation map is generated from this similarity. (B) A benign histopathological image paired with the same mask. Since there is no similarity, the map is not activated. In flattening (B and C), the activation map is vectorized, and the given coefficients are higher for the image corresponding to the mask that identify relevant features of malignancy (star) than when compared to coefficients from the benign image.

At the end of the training, an activation map containing features to discriminate classes is created, hereby represented by a stellated mask. **Figures 4 B and C** exemplify the correlation between a mask generated by a trained CNN and two different input images: a stellated malignant and a rounded benign image. At the end of this interaction, the activation map will be only "highlighted" if the corresponding malignant cells are presented in the image. This correspondence between mask and image explains why CNNs are classified as DN models that learn "directly from the data".

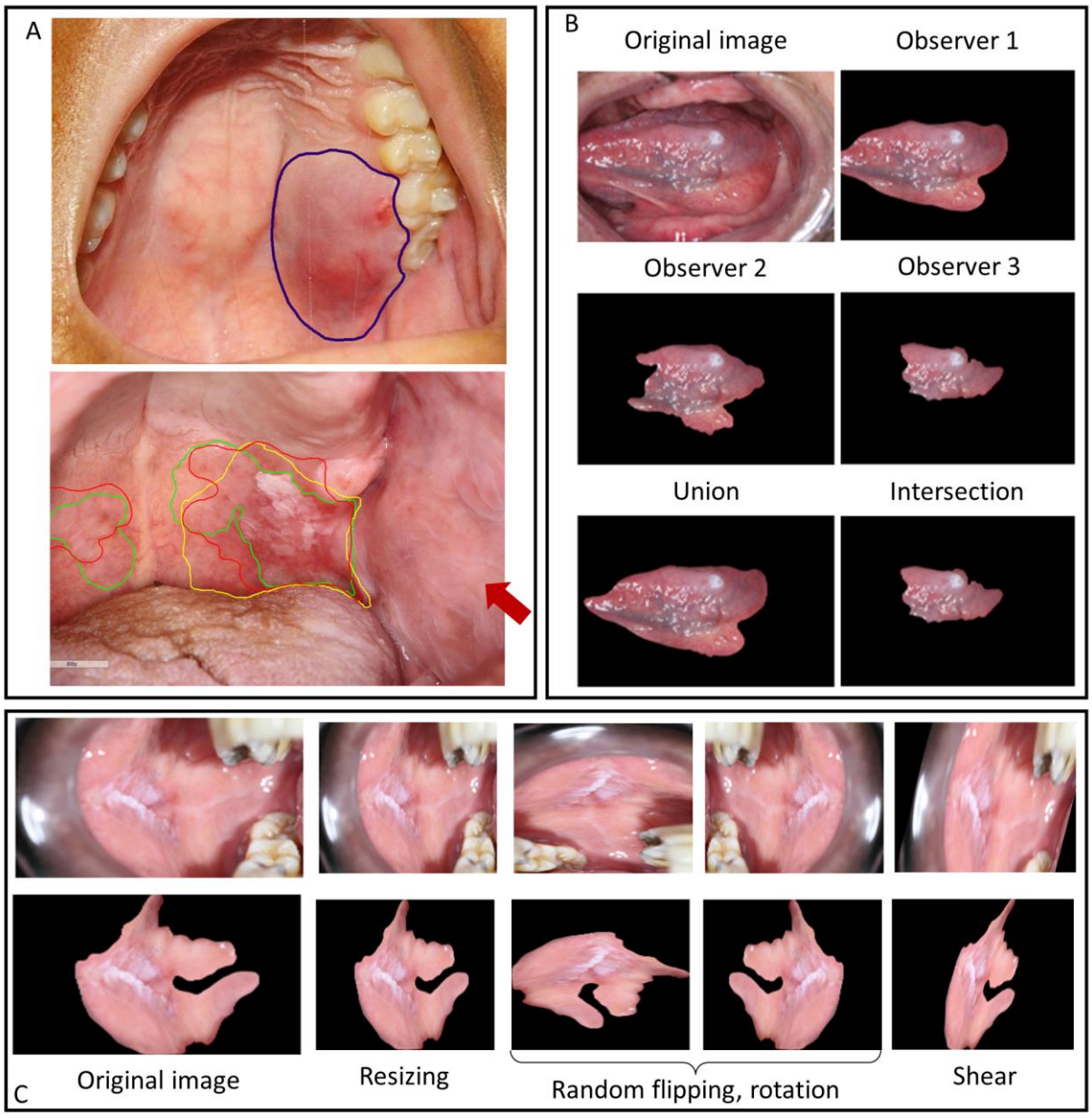


Figure 4. Image annotation, segmentation and data augmentation. (A) Clinical photograph annotations of actinic cheilitis and a salivary gland tumor in the palate (first row). The white plaque is adjacent to areas with flash reflex in the lower lip vermilion, which could lead to

misclassification; the tumor in the palate presents saliva strings in the oral mucosa surface. Clinical photograph annotations of oral potentially malignant disorders (OPMD) in the tongue and soft palate according to three observers (second row). The red arrow indicates a leucoedema lesion susceptible to being misclassified as a premalignant lesion. (B) Segmentation masks, union and intersection of all annotations. The union is commonly used as the ground truth to train the model, minimizing the false negatives. (C) Data augmentation of resized clinical images and masks. (To be continued).

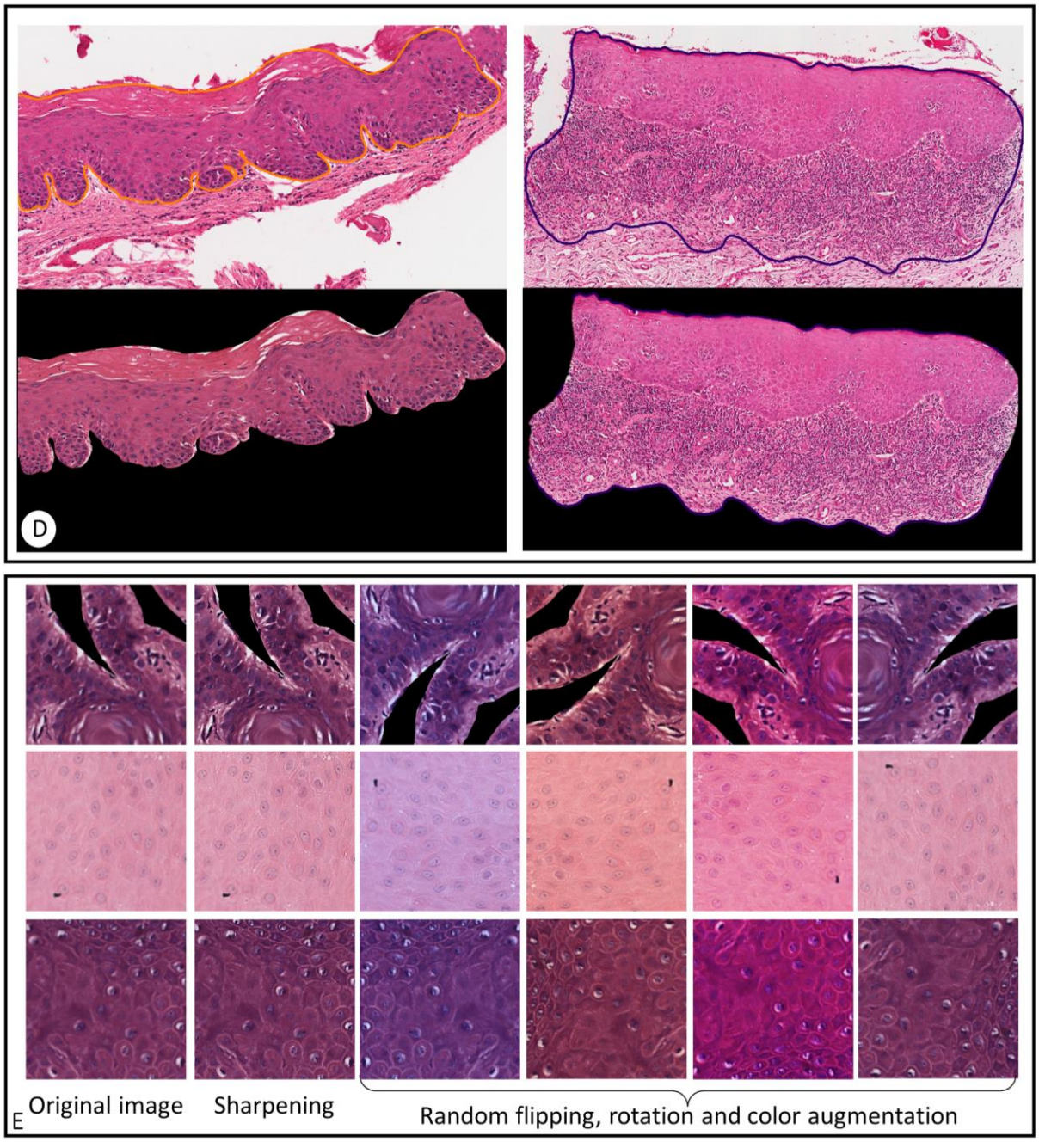


Figure 4. (Continuation) (D) Annotation and segmentation of the region of interest (ROI) in a histopathological digital slide. In the present case, the inclusion of normal connective tissue is not feasible but should be considered in lichenoid lesions, for example. If the classification problem is not simply tumor versus nontumour, it is important to include sufficient features that compound the diagnosis, and this can be variable according to the disease of interest. (E) Random data and color augmentation of histopathological patches.

4. Application of image-based AI models in oral medicine and head and neck oncology

The prospect of obtaining new clinicopathological correlations from image biomarkers through pattern recognition of subtle information in histopathological images has the potential to improve diagnosis, guide treatment, and improve prognosis[6]. Additionally, image analysis has the potential to reduce the need for immunohistochemical staining to reach a conclusive differential diagnosis of histologically similar lesions[7].

Traditional ML approaches for classification and segmentation of histopathological images are the most used thus far in oral medicine and pathology[8] with a special emphasis on oral potentially malignant disorders (OPMD) and oral squamous cell carcinoma (OSCC) diagnosis. These classical models presented good results for the proposed image recognition task, with an accuracy of up to 95%[9-14]. However, these studies have limited context considering the amount of data variability, sampling strategies and non-standardized reported metrics.

DL-based models for histopathological image analysis were not widely used until recently[15,16] and should be investigated. Moreover, models for the clinical diagnosis of H&N cancer[17] and approaches that use radiographs for the context of CCP to provide decision support for cancer treatment and outcome prediction have been proposed[18,19].

Articles using DN models for the early diagnosis of OPMD based on clinical images demonstrate high-performance metrics. Shamin[17] recently reported a screening DL-based model to detect OPMD with a mean accuracy of 0.98 using Vgg19. Camalan[20] performed transfer learning on Inception-ResNet-V2 to classify photographic images as suspicious or normal (accuracy of 90.9%), and class activation maps were further generated to evaluate the most discriminative regions for the classification task. Jubair[21] proposed a pretrained CNN (EfficientNet-B0) to differentiate normal images from premalignant/malignant images. This light CNN achieved an accuracy of 85.0%. Tanriver[22] compared 2 segmentation methods (semantic and instance) and performed an object detection experiment and further 3-class

problem classification (benign, OPMD, and malignant). Classification experiments compared five “state-of-the-art” architectures with encouraging results. A multi-institutional study[23] to detect oral squamous cell carcinoma (OSCC) using DL achieved 0.93 in the external test set, demonstrating a satisfactory generalization ability of the model. The final goal of these approaches is to develop an automated method to support the clinical decision of oral medicine doctors in oral cancer screening and early identification.

Additionally, prognostic models based on low-level imaging features (nuclear shape and texture)[24] and clinical and genomic markers[25,26] have been proposed. Kim[27] compared traditional classifiers and a DL model for survival prediction of oral cancer patients, with DeepSurv out-performing random survival forest (RSF) and the Cox proportional hazard model (CPH). Mahmood[28] correlated cytological features of oral epithelial dysplasia in glass slides with transformation and recurrence using Cox proportional hazard regression and Kaplan–Meier curves with a fair prediction power.

Treatment-related toxicity prediction models are usually based on the inclusion of image biomarkers known as radiomic features or non-imaging data (clinical, pathological, dosimetric features)[29-38]. These multivariable models provide a wide range of interpretive possibilities and results of the area under the receiver operation curve (AUROC) reaches values between 0.43 and 0.98. Despite these encouraging metrics, further conclusions on the reliability of these models to predict radiation-induced toxicity in HNC are otherwise shallow[38]. Published studies are not externally validated, an important step to test the generalization ability of the model in unseen data. Additionally, the lack of standardization in research limits the interpretation of the results.

5. Methods for image-based DL model training for diagnosis

5.1 Clinical actions

Image acquisition and data collection

Image acquisition refers to the digitalization of a glass slide and clinical photography. The quality of imaging datasets is conditioned on the early construction phases (case retrieval, review, and selection), as well as on the equipment used. According to the *Guidelines for less biased data collection and algorithm evaluation*[39], technical and biological variabilities, particularities of training set, class definition, sampling, evaluation, and quality control protocols, reproducibility, traceability, and software variations should be considered to construct realistic ground-truth datasets.

Technical variations on glass slide preparations or different scanner equipment can generate a variety of color matrices and intensity. Ignoring color variation in dataset construction can affect the performance of a DL model affecting its capacity to generalize the data. It is important to construct a training dataset composed of a realistic, rich, and balanced variation of possible features and colors, since a varied but realistic dataset potentializes generalization during learning, as a test is performed in a blind dataset[40].

Technical variability is important to reflect the clinician's reality and to build a robust model for image classification. Therefore, a previous computer processing step of color normalization, color augmentation and conversion to grayscale can be applied[41]. Nevertheless, the utilization of samples from different institutions and obtained from different equipment is encouraged[39].

Despite not having a consensus on the proper number of patients/images, the recommendation is to include a wide sample, but most importantly, the representation of each class should cover all possible biological variations (e.g., histological variations, subtypes, or patterns of a tumor in WSI and race representation in clinical photographs). This can be challenging at the onset of some pools of lesions, especially for rarer tumors.

A certain imbalance feature variation is, therefore, expected but should be minimized by data augmentation, a technique that aims to generate more examples with a diversity of characteristics, increasing the size and variety of the sample for training, which is one of the most important aspects to consider when training a DL model.

Annotation

Image annotation is one of the most important and challenging steps of computer vision research and requires experienced professionals' involvement to label the data. Learning modalities are defined according to the data labeling as i) Supervised, ii) Semi-Supervised and iii) Unsupervised. Random Forest, Support Vector Machines (SVM), and ANN (including DNs) are models trained with labeled data (annotated images) in which a previous determination of each class was assigned by a panel of experts to provide a reference standard, a ground truth for the model to learn (Supervised Learning).

Annotating WSI is a time-consuming task, and despite some attempts to automatize regions of interest (ROI) detection[42,43], manual annotation remains the best strategy to set a reference standard for the model, providing examples of representative areas of tumors, improving the training time and increasing the classification accuracy[44]. From our

perspective, the quality of the annotation relies on the gray area of human interpretation, and this subjectivity can be transferred to the model since the classification provided by digital systems seeks to quantify expert perception. Otherwise, this entanglement does not affect the reliability of the artificial model but mimics human perception and decision-making.

When not crucial for classification, the stroma (connective tissue) should not be included since the inclusion of common image information for both classes can either be disregarded or even generate overoptimistic performance results. To overcome methodological bias, WSI should be reviewed by an expert panel to decide and supervise ROI annotation with at least three observers being enrolled in radiographic and clinical photography annotation, which should be biopsy proven. The subjective and highly variable limits of some lesions, particularly thin white plaques and striated lesions (e.g., leucoplakias and lichenoid lesions), are particularly challenging, as are overlapping histopathological features among distinct grades (e.g., dysplasia grading) **(Figure 4).**

5.2 Computational actions

Pre-processing

Pre-processing aims to remove image noise, which is virtually any information not wanted for the classification task. A classic example is a misclassification due the presence of hair along with the skin lesion[45].

In the domain of oral clinical photographs, saliva and humidity-associated flash reflex can be present according to the location of the lesion and light incidence, which can be misclassified as a white lesion according to the pixel value. The same applies to regions with augmented vascularization, such as the soft palate, floor of mouth and red lesions **(Figure 4)**.

Conversely, when dealing with WSIs, it is impractical to remove artifacts, which could lead to an unrealistic biased sample. According to the guidelines for generating less biased datasets for algorithm evaluation[39], one of the recommendations highlights the importance of including real-life and equally balanced technical variations as focus-related artifacts, dark spots, fingerprints, HE-staining scheme[46], bubbles, tissue tears/folds, autolysis artifacts, and formalin pigments[47].

Image segmentation

Image segmentation is the contour definition and the separation of relevant information from nonrelevant background. In WSI, it identifies and separates the ROI from

nontumour areas according to the reference annotation[48]. In clinical photographs, it is useful to separate the lesion from surgical retractors, teeth, gauze and gloved fingers retracting the oral mucosa, which are usually framed with the lesion, leaving the ROI with only relevant information for further classification[17,21,22] **(Figure 4).**

Segmentation can be manual, in which an experienced professional manually delineates the ROI (image annotation); semiautomatic, with seed or approximate contour; and automatic, defined by the format variability and image quality**.** A practical example is pixel classification given a set of discriminant features (e.g., color matrix, intensity) within a range of values given by an activation function (semantic classification).

Patch generation (fragmentation)

WSI fragmentation generates smaller image samples (patches) with dimensions according to the CNN-model input structure and kernel size (e.g., 2562x256 for AlexNet; 299x299 for InceptionV3). Ideally, the input size is fitted to pursue a minimally efficient template for learning, respecting the median minimum size of the most important structures (e.g., acini). This strategy reduces the computational complexity while preventing image information loss[49].

WSI can exceed gigabytes per image, and since the majority of CNNs have limitations regarding larger image sizes, fragmentation is required. However, small patch sizes can interfere with the accuracy[50], and this is particularly true for histological images that should have sufficient information to map the input to the output in one patch. An intermediate patch size can be arbitrarily chosen to avoid losing information when resizing the patches to fit in several CNNs but a drastic resizing can distort the image information. Additionally, fragmentation of segmented images creates patches from both classes separately while fragmenting a no segmented image generates patches from the transitional zone, which could improve training efficiency[51].

Fragmentation in clinical photographs was reported by Camalan[20] in an approach that created heatmaps based on the patch predicted class to identify regions in the image most significant in classification. Class activation map interpretation was more expressive for the fragmented images, providing a better understanding. For clinical images, the most common reduction strategy reported is resizing[17,21,22].

Data augmentation

Data augmentation comprises strategies to increase the dataset size and variability, either because there is a limited dataset regarding the variation of important features or because the dataset does not have an even distribution among classes. An unbalanced sample affects the performance of ML-based classifiers since it can induce the CNN to classify according to the most common class, especially in a small dataset[52]. The imbalance of the dataset classes is an important factor when evaluating the sensitivity of the proposed method.

Data augmentation involves several techniques to artificially generate random variations in original images (rotation, zoom, flipping, mirroring, resizing, blurring, noising sharpening, elastic deformation, brightness, contrast and color variation) to increase the number of samples available for training and, therefore, improve the models' performance[39-54]. Color augmentation is a specific augmentation technique focused on color variation – red, green, and blue combination (RGB) – of the underlying image. Additionally, perturbation of the RGB color space automatically generates stain variation, mimicking technical variations in glass slide processing[55-57] **(Figure 4).** The variation in feature characteristics provided by augmentation provides better learning, thus improving the generalization ability of the model. This is important, as the differences among datasets are considered the main cause for inter-dataset dissimilarities[40]. Data augmentation can also be used to eliminate site-specific signatures[58].

Sampling

Sampling methods are strategies to split the full dataset into training:validation:test sets **(Figure 5).** Usually, the validation set is reserved to test the parameters, while the test set is used to assess the performance of the model. Proportion is chosen according to sample size and aims to provide a representative and large enough dataset for training, as well as sufficient samples for validation and testing (e.g., 80%:10%:10% or 70%:20%:10%). When dealing with WSIs, special attention is required for patient distribution to prevent the inclusion of patches from a single patient in both the training and test sets to prevent data leakage.

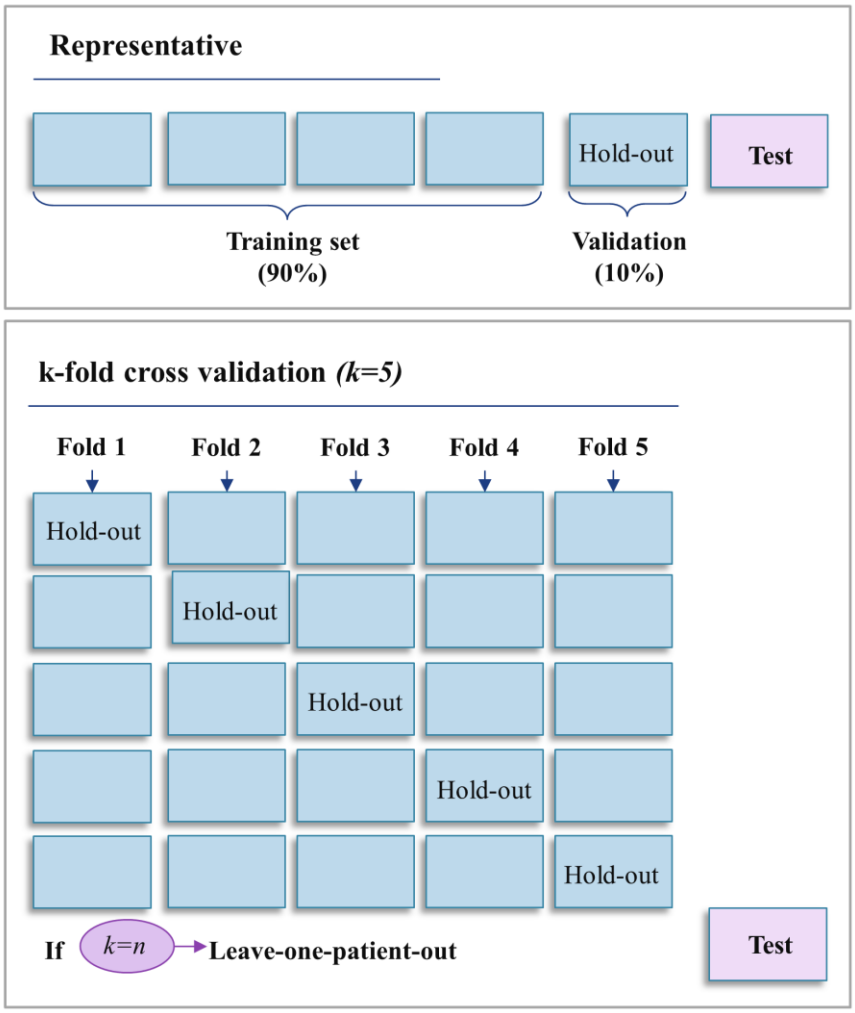


Figure 5. Sampling methods. Training and validation data are used to build, train the model and fit the hyperparameters. Test data remains unseen and are further used to evaluate the generalization ability of the model (class discrimination power).

k-fold cross validation is a sampling method that allows training ML models with limited data in a mutually exclusive way, resulting in less biased/overestimated performance metrics. In this procedure, the full dataset is randomly split into k groups in which one group is held out for validation while the remaining is used for training with each fold representing the group shifting. This strategy eventually allows the model to be trained with the entire dataset. When the number of folds is equal to the number of samples (*k=n)*, the cross-validation method is called leave-one-patient-out (LoPo) **(Figure 5).**

Transfer Learning

Transfer Learning is a strategy in which parameters (pre-trained weights) from a previously trained model are reused as initialization in other CNN for the same, similar, or different classification tasks, and Since DN models do not use tissue-specific information, it is feasible to apply the same model for different problems[51]. This strategy is usually applied when the current dataset is insufficient for training a model for a given task or for improving the computing time since it requires less training time. However, if there is a large amount of data and proper hyperparameters the recommendation is not to use transfer learning, since learning from scratch provides better performance if taking into account sample-dependent features[6,44].

DN models have better generalization ability when compared to classical ML-based classifiers, and this tends to improve once the dataset increases. On the other hand, increasing the dataset size does not seem to affect classical ML-based classifier performances. Ultimately, there is no difference between the SVM and DL performances when using a small dataset[49].

Conclusion

The incorporation of ML-based models in the diagnostic process has the potential to reveal new correlations between clinical-pathological characteristics and image processing for an early and accurate diagnosis. This contributes to personalized treatment planning, which consequently improves survival rates, reducing the need for invasive and mutilating surgeries and the risk of recurrence or metastatic disease. The distinct complexion of the multidisciplinary teams involved in the development of new approaches can delay or generate inefficient communication between individual teams specialized in the medical and biomedical engineering fields. The present work presents a privileged interdisciplinary view that aims to unify the perception resulting from the performance of oral pathologists, oral medicinists, head and neck surgeons and computer scientists, focusing on promoting the development and implementation of new ML image-based approaches for H&N cancer diagnosis.

Conflict of interest

We declare that the authors have no financial relationship with any commercial associations, current and within the past five years, that might pose a potential, perceived or real conflict of

interest. These include grants, patent-licensing arrangements, consultancies, stock or other equity ownership, advisory board memberships, or payments for conducting or publicizing our study.

Ethics Approval

Images exhibited in this publication are from patients who are part of a study performed in accordance with the Declaration of Helsinki and approved by the Piracicaba Dental Ethical Committee, Registration number CAAE: 42235421.9.0000.5418.

Author contribution

All authors have made substantial contributions to the conception (MCM, ALDA), draft and design (ALDA, VMDS, MSK, ESCS, MCM), and review (ESCS, MAL, PAV, SAK, ATP, LPK, ACPLFC, ARSS, MCM) the final version of the work. The authors agree to be accountable for all aspects of the work and ensures that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Funding

This study was funded by the Coordination for the Improvement of Higher Education Personnel (CAPES/PROEX, Brazil) process number 001, the National Council for Scientific and Technological Development (CNPq, Brazil), by the grants from São Paulo Research Foundation (FAPESP, Brazil) process number: 2009/53839-2, and the Fund to support teaching, research and extension (FAEPEX, Brazil) process number: 2597/21, which supported the acquisition of the equipment used.

References

1. Krohn J, Beyleveld G, Bassens A. Deep Learning Illustrated: A Visual, Interactive Guide to Artificial Intelligence. 2019. ISBN 10: 0135121728; 13: 9780135121726.
2. Zhang A, Lipton ZC, Li M, Smola AJ. Dive into Deep Learning. 2021 doi: doi.org/10.48550/arXiv.2106.11342
3. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. 1943. Bull Math Biol. 1990;52(1-2):99-115; discussion 73-97. PMID: 2185863.
4. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev. 1958 Nov;65(6):386-408. doi: 10.1037/h0042519.
5. Zaheer R, Shaziya H. A Study of the Optimization Algorithms in Deep Learning. Third International Conference on Inventive Systems and Control (ICISC), 2019. 536-539. doi: 10.1109/ICISC44355.2019.9036442.
6. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, Moreira AL, Razavian N, Tsirigos A. Classification and mutation prediction from non-small cell lung

cancer histopathology images using deep learning. Nat Med. 2018 Oct;24(10):1559-1567. doi: 10.1038/s41591-018-0177-5.

7. Kubach J, Muhlebner-Fahrngruber A, Soylemezoglu F, Miyata H, Niehusmann P, Honavar M, Rogerio F, Kim SH, Aronica E, Garbelli R, Vilz S, Popp A, Walcher S, Neuner C, Scholz M, Kuerten S, Schropp V, Roeder S, Eichhorn P, Eckstein M, Brehmer A, Kobow K, Coras R, Blumcke I, Jabari S. Same but different: A Web-based deep learning application revealed classifying features for the histopathologic distinction of cortical malformations. Epilepsia. 2020 Mar;61(3):421-432. doi: 10.1111/epi.16447.
8. Mahmood H, Shaban M, Rajpoot N, Khurram SA. Artificial Intelligence-based methods in head and neck cancer diagnosis: an overview. Br J Cancer. 2021 Jun;124(12):1934-1940. doi: 10.1038/s41416-021-01386-x. Epub 2021 Apr 19. PMID: 33875821; PMCID: PMC8184820.
9. Muthu Rama Krishnan M, Pal M, Bomminayuni SK, Chakraborty C, Paul RR, Chatterjee J, Ray AK. Automated classification of cells in sub-epithelial connective tissue of oral sub-mucous fibrosis-an SVM based approach. Comput Biol Med. 2009 Dec;39(12):1096-104. doi: 10.1016/j.compbiomed.2009.09.004.
10. Muthu Rama Krishnan M, Choudhary A, Chakraborty C, Ray AK, Paul RR. Texture based segmentation of epithelial layer from oral histological images. Micron. 2011 Aug;42(6):632-41. doi: 10.1016/j.micron.2011.03.003.
11. Krishnan MM, Acharya UR, Chakraborty C, Ray AK. Automated diagnosis of oral cancer using higher order spectra features and local binary pattern: a comparative study. Technol Cancer Res Treat. 2011 Oct;10(5):443-55. doi: 10.7785/tcrt.2012.500221.
12. Krishnan MM, Venkatraghavan V, Acharya UR, Pal M, Paul RR, Min LC, Ray AK, Chatterjee J, Chakraborty C. Automated oral cancer identification using histopathological images: a hybrid feature extraction paradigm. Micron. 2012 Feb;43(2-3):352-64. doi: 10.1016/j.micron.2011.09.016.
13. Muthu Rama Krishnan M, Shah P, Chakraborty C, Ray AK. Statistical analysis of textural features for improved classification of oral histopathological images. J Med Syst. 2012 Apr;36(2):865-81. doi: 10.1007/s10916-010-9550-8.
14. Mahmood H, Shaban M, Indave BI, Santos-Silva AR, Rajpoot N, Khurram SA. Use of artificial intelligence in diagnosis of head and neck precancerous and cancerous lesions: A systematic review. Oral Oncol. 2020 Nov;110:104885. doi: 10.1016/j.oraloncology.2020.104885.
15. Shaban M, Khurram SA, Fraz MM, Alsubaie N, Masood I, Mushtaq S, Hassan M, Loya A, Rajpoot NM. A Novel Digital Score for Abundance of Tumour Infiltrating Lymphocytes Predicts Disease Free Survival in Oral Squamous Cell Carcinoma. Sci Rep. 2019 Sep 16;9(1):13341. doi: 10.1038/s41598-019-49710-z
16. Fraz MM, Shaban M, Graham S, Khurram SA, Rajpoot NM. Uncertainty Driven Pooling Network for Microvessel Segmentation in Routine Histology Images. In: Computational Pathology and Ophthalmic Medical Image Analysis. OMIA COMPAY 2018. Lecture Notes in Computer Science. Springer International Publishing; 2018; 11039:156–64. doi: doi.org/10.1007/978-3-030-00949-6_19
17. Shamim MZM, Syed S, Shiblee M, Usman M, Ali S. Automated Detection of Oral Pre-Cancerous Tongue Lesions Using Deep Learning for Early Diagnosis of Oral Cavity Cancer. The Computer Journal, Jan 2020. 65(1):91–104. doi: doi.org/10.1093/comjnl/bxaa136
18. Sher DJ, Godley A, Park Y, Carpenter C, Nash M, Hesami H, Zhong Z, Lin MH. Prospective study of artificial intelligence-based decision support to improve head and neck radiotherapy plan quality. Clin. Transl. Radiat. Oncol.2021; 29:65-70. ISSN 2405-6308. doi: doi.org/10.1016/j.ctro.2021.05.006.

19. Kearney V, Chan JW, Valdes G, Solberg TD, Yom SS. The application of artificial intelligence in the IMRT planning process for head and neck cancer, Oral Oncol. 2018; 87:111-116. ISSN 1368-8375.
20. Camalan S, Mahmood H, Binol H, Araújo ALD, Santos-Silva AR, Vargas PA, Lopes MA, Khurram SA, Gurcan MN. Convolutional Neural Network-Based Clinical Predictors of Oral Dysplasia: Class Activation Map Analysis of Deep Learning Results. Cancers (Basel). 2021 Mar 14;13(6):1291. doi: 10.3390/cancers13061291.
21. Jubair F, Al-Karadsheh O, Malamos D, Al Mahdi S, Saad Y, Hassona Y. A novel lightweight deep convolutional neural network for early detection of oral cancer. Oral Dis. 2021 Feb 26. doi: 10.1111/odi.13825.
22. Tanriver G, Soluk Tekkesin M, Ergen O. Automated Detection and Classification of Oral Lesions Using Deep Learning to Detect Oral Potentially Malignant Disorders. Cancers (Basel). 2021 Jun 2;13(11):2766. doi: 10.3390/cancers13112766.
23. Fu Q, Chen Y, Li Z, Jing Q, Hu C, Liu H, Bao J, Hong Y, Shi T, Li K, Zou H, Song Y, Wang H, Wang X, Wang Y, Liu J, Liu H, Chen S, Chen R, Zhang M, Zhao J, Xiang J, Liu B, Jia J, Wu H, Zhao Y, Wan L, Xiong X. A deep learning algorithm for detection of oral cavity squamous cell carcinoma from photographic images: A retrospective study. EClinicalMedicine. 2020 Sep 23;27:100558. doi: 10.1016/j.eclinm.2020.100558.
24. Lu C, Lewis JS Jr, Dupont WD, Plummer WD Jr, Janowczyk A, Madabhushi A. An oral cavity squamous cell carcinoma quantitative histomorphometric-based image classifier of nuclear morphology can risk stratify patients for disease-specific survival. Mod Pathol. 2017 Dec;30(12):1655-1665. doi: 10.1038/modpathol.2017.98.
25. Chang SW, Abdul-Kareem S, Merican AF, Zain RB. Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods. BMC Bioinformatics. 2013 May 31;14:170. doi: 10.1186/1471-2105-14-170.
26. Chang S. A Hybrid Prognostic Model for Oral Cancer based on Clinicopathologic and Genomic Markers. Sains Malaysiana. 2014. 43:567 – 573.
27. Kim DW, Lee S, Kwon S, Nam W, Cha IH, Kim HJ. Deep learning-based survival prediction of oral cancer patients. Sci Rep. 2019 May 6;9(1):6994. doi: 10.1038/s41598-019-43372-7.
28. Mahmood H, Bradburn M, Rajpoot N, Islam NM, Kujan O, Khurram SA. Prediction of malignant transformation and recurrence of oral epithelial dysplasia using architectural and cytological feature specific prognostic models. Mod Pathol. 2022 Mar 31. doi: 10.1038/s41379-022-01067-x.
29. Nardone V, Tini P, Nioche C, Mazzei MA, Carfagno T, Battaglia G, Pastina P, Grassi R, Sebaste L, Pirtoli L. Texture analysis as a predictor of radiation-induced xerostomia in head and neck patients undergoing IMRT. Radiol Med. 2018 Jun;123(6):415-423. doi: 10.1007/s11547-017-0850-7.
30. Dean JA, Welsh LC, Wong KH, Aleksic A, Dunne E, Islam MR, Patel A, Patel P, Petkar I, Phillips I, Sham J, Schick U, Newbold KL, Bhide SA, Harrington KJ, Nutting CM, Gulliford SL. Normal Tissue Complication Probability (NTCP) Modelling of Severe Acute Mucositis using a Novel Oral Mucosal Surface Organ at Risk. Clin Oncol (R Coll Radiol). 2017 Apr;29(4):263-273. doi: 10.1016/j.clon.2016.12.001.
31. Dean J, Wong K, Gay H, Welsh L, Jones AB, Schick U, Oh JH, Apte A, Newbold K, Bhide S, Harrington K, Deasy J, Nutting C, Gulliford S. Incorporating spatial dose metrics in machine learning-based normal tissue complication probability (NTCP) models of severe acute dysphagia resulting from head and neck radiotherapy. Clin Transl Radiat Oncol. 2018 Jan;8:27-39. doi: 10.1016/j.ctro.2017.11.009.

32. Ursino S, Giuliano A, Martino FD, Cocuzza P, Molinari A, Stefanelli A, Giusti P, Aringhieri G, Morganti R, Neri E, Traino C, Paiar F. Incorporating dose-volume histogram parameters of swallowing organs at risk in a videofluoroscopy-based predictive model of radiation-induced dysphagia after head and neck cancer intensity-modulated radiation therapy. Strahlenther Onkol. 2021 Mar;197(3):209-218. doi: 10.1007/s00066-020-01697-7.
33. Wentzel A, Hanula P, van Dijk LV, Elgohari B, Mohamed ASR, Cardenas CE, Fuller CD, Vock DM, Canahuate G, Marai GE. Precision toxicity correlates of tumor spatial proximity to organs at risk in cancer patients receiving intensity-modulated radiotherapy. Radiother Oncol. 2020 Jul;148:245-251. doi: 10.1016/j.radonc.2020.05.023.
34. De Araujo Faria V, Azimbagirad M, Viani Arruda G, Fernandes Pavoni J, Cezar Felipe J, Dos Santos EMCMF, Murta Junior LO. Prediction of Radiation-Related Dental Caries Through PyRadiomics Features and Artificial Neural Network on Panoramic Radiography. J Digit Imaging. 2021 Oct;34(5):1237-1248. doi: 10.1007/s10278-021-00487-6.
35. Humbert-Vidan L, Patel V, Oksuz I, King AP, Guerrero Urbano T. Comparison of machine learning methods for prediction of osteoradionecrosis incidence in patients with head and neck cancer. Br J Radiol. 2021 Apr 1;94(1120):20200026. doi: 10.1259/bjr.20200026.
36. Smyczynska U, Grabia S, Nowicka Z, Papis-Ubych A, Bibik R, Latusek T, Rutkowski T, Fijuth J, Fendler W, Tomasik B. Prediction of Radiation-Induced Hypothyroidism Using Radiomic Data Analysis Does Not Show Superiority over Standard Normal Tissue Complication Models. Cancers (Basel). 2021 Nov 8;13(21):5584. doi: 10.3390/cancers13215584.
37. Abdollahi H, Mostafaei S, Cheraghi S, Shiri I, Rabi Mahdavi S, Kazemnejad A. Cochlea CT radiomics predicts chemoradiotherapy induced sensorineural hearing loss in head and neck cancer patients: A machine learning and multi-variable modelling study. Phys Med. 2018 Jan;45:192-197. doi: 10.1016/j.ejmp.2017.10.008.
38. Carbonara R, Bonomo P, Di Rito A, Didonna V, Gregucci F, Ciliberti MP, Surgo A, Bonaparte I, Fiorentino A, Sardaro A. Investigation of Radiation-Induced Toxicity in Head and Neck Cancer Patients through Radiomics and Machine Learning: A Systematic Review. J Oncol. 2021 Jun 9;2021:5566508. doi: 10.1155/2021/5566508.
39. Marée R. The Need for Careful Data Collection for Pattern Recognition in Digital Pathology. J Pathol Inform. 2017 Apr 10;8:19. doi: 10.4103/jpi.jpi_94_16.
40. Somaratne U, Wong KW, Parry J, Sohel F,Wang X, Laga H, Hamid. Improving Follicular Lymphoma Identification using the Class of Interest for Transfer Learning. *2019 Digital Image Computing: Techniques and Applications (DICTA)*, 2019;1-7. doi: 10.1109/DICTA47822.2019.8946075.
41. Komura D, Ishikawa S. Machine Learning Methods for Histopathological Image Analysis. Comput Struct Biotechnol J. 2018 Feb 9;16:34-42. doi: 10.1016/j.csbj.2018.01.001.
42. Mercan E, Aksoy S, Shapiro LG, Weaver DL, Brunyé TT, Elmore JG. Localization of Diagnostically Relevant Regions of Interest in Whole Slide Images: a Comparative Study. J Digit Imaging. 2016 Aug;29(4):496-506. doi: 10.1007/s10278-016-9873-1.
43. Kothari S, Phan JH, Osunkoya AO, Wang MD. Biological Interpretation of Morphological Patterns in Histopathological Whole-Slide Images. ACM BCB. 2012 Oct;2012:218-225. doi: 10.1145/2382936.2382964.
44. Hart SN, Flotte W, Norgan AP, Shah KK, Buchan ZR, Mounajjed T, Flotte TJ. Classification of Melanocytic Lesions in Selected and Whole-Slide Images via

Convolutional Neural Networks. J Pathol Inform. 2019 Feb 20;10:5. doi: 10.4103/jpi.jpi_32_18.
45. Munir K, Elahi H, Ayub A, Frezza F, Rizzi A. Cancer Diagnosis Using Deep Learning: A Bibliographic Review. Cancers (Basel). 2019 Aug 23;11(9):1235. doi: 10.3390/cancers11091235.
46. Schömig-Markiefka B, Pryalukhin A, Hulla W, Bychkov A, Fukuoka J, Madabhushi A, Achter V, Nieroda L, Büttner R, Quaas A, Tolkach Y. Quality control stress test for deep learning-based diagnostic model in digital pathology. Mod Pathol. 2021 Dec;34(12):2098-2108. doi: 10.1038/s41379-021-00859-x.
47. Taqi SA, Sami SA, Sami LB, Zaki SA. A review of artifacts in histopathology. J Oral Maxillofac Pathol. 2018 May-Aug;22(2):279. doi: 10.4103/jomfp.JOMFP_125_15.
48. Krishnan MM, Acharya UR, Chakraborty C, Ray AK. Automated diagnosis of oral cancer using higher order spectra features and local binary pattern: a comparative study. Technol Cancer Res Treat. 2011 Oct;10(5):443-55. doi: 10.7785/tcrt.2012.500221.
49. Lai Z, Deng H. Medical Image Classification Based on Deep Features Extracted by Deep Model and Statistic Feature Fusion with Multilayer Perceptron. Comput Intell Neurosci. 2018 Sep 12;2018:2061516. doi: 10.1155/2018/2061516.
50. Mishra R, Daescu O, Leavey P, Rakheja D, Sengupta A. Convolutional Neural Network for Histopathological Analysis of Osteosarcoma. J Comput Biol. 2018 Mar;25(3):313-325. doi: 10.1089/cmb.2017.0153.
51. Nirschl JJ, Janowczyk A, Peyster EG, Frank R, Margulies KB, Feldman MD, Madabhushi A. Deep learning tissue segmentation in cardiac histopathology images. Deep learning for medical image analysis, Elsevier. 2017. 179-195.
52. Wei Q, Dunbrack RL Jr. The role of balanced training and testing data sets for binary classifiers in bioinformatics. PLoS One. 2013 Jul 9;8(7):e67863. doi: 10.1371/journal.pone.0067863.
53. Xu H, Lu C, Berendt R, Jha N, Mandal M. Automated analysis and classification of melanocytic tumor on skin whole slide images. Comput Med Imaging Graph. 2018 Jun; 66:124-134. doi: 10.1016/j.compmedimag.2018.01.008.
54. Gómez Hernández, Eduardo & García, José. Parallel Computing: Technology Trends, 2019. 36: 35. doi: 10.3233/APC200022.
55. Tellez D, Litjens G, Bándi P, Bulten W, Bokhorst JM, Ciompi F, van der Laak J. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Med Image Anal. 2019 Dec;58:101544. doi: 10.1016/j.media.2019.101544.
56. Mikołajczyk A, Grochowski M. Data augmentation for improving deep learning in image classification problem. 2018 International Interdisciplinary PhD Workshop (IIPhDW), 2018, pp. 117-122, doi: 10.1109/IIPHDW.2018.8388338.
57. Faryna K, van der Laak J, Litjens G. Tailoring automated data augmentation to H&E-stained histopathology. In Medical Imaging with Deep Learning. 2021 Feb. 143:168-178 Available from https://proceedings.mlr.press/v143/faryna21a.html.
58. Howard FM, Dolezal J, Kochanny S, Schulte J, Chen H, Heij L, Huo D, Nanda R, Olopade OI, Kather JN, Cipriani N, Grossman RL, Pearson AT. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. Nat Commun. 2021 Jul 20;12(1):4423. doi: 10.1038/s41467-021-24698-1.

2.2 Artigo: The use of Deep Learning State-of-the-Art Architectures for Oral Epithelial Dysplasia Grading: a comparative assessment

[a]Anna Luíza Damaceno Araújo*; [b]Viviane Mariano*; [b]Matheus Cardoso Moraes; [b]Henrique Alves de Amorim, [c]Felipe Paiva Fonseca; [c]Maria Sissa Sant'Ana; [c]Ricardo Alves de Mesquita, [a]Bruno Augusto Linhares Almeida Mariz; [d]Hélder Antônio Rebelo Pontes; [a]Lucas Lacerda de Souza; [a]Cristina Saldivia Siracusa; [e]Syed Ali Khurram; [f,g]Alexander T. Pearson; [h]Manoela Domingues Martins; [a]Márcio Ajudarte Lopes; [a]Pablo Agustin Vargas; [i,j] Luiz Paulo Kowalski; [k]Saman Warnakulasuriya; [a]Alan Roger Santos-Silva.

Affiliation:
[a]Oral Diagnosis Department, Piracicaba Dental School, University of Campinas (UNICAMP), Piracicaba, São Paulo, Brazil.
[b]Institute of Science and Technology, Federal University of São Paulo (ICT-Unifesp), São José dos Campos, São Paulo, Brazil.
[c]Department of Oral Surgery and Pathology, School of Dentistry, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Minas Gerais, Brazil.
[d]Service of Oral Pathology, João de Barros Barreto University Hospital, Federal University of Pará, Belém, Brazil.
[e]Unit of Oral and Maxillofacial Pathology, School of Clinical Dentistry, University of Sheffield, S10 2TA, Sheffield, UK.
[f]Section of Hematology/Oncology, Department of Medicine, University of Chicago, Chicago, IL, USA.
[g]University of Chicago Comprehensive Cancer Center, Chicago, IL, USA.
[h]Department of Oral Pathology, School of Dentistry, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil
[i]Department of Head and Neck Surgery and Otorhinolaryngology, A.C. Camargo Cancer Center, São Paulo, Brazil.
[j]Head and Neck Surgery Department and LIM 28, University of São Paulo Medical School, São Paulo, Brazil.
[k]Emeritus Professor, King's College London, UK; WHO Collaborating Centre for Oral Cancer, London, UK.
***These authors contributed equally to the manuscript**

Corresponding Author: Alan Roger Santos-Silva
Oral Diagnosis Department, Piracicaba Dental School, UNICAMP
Adress: Av. Limeira, nº 901, Areião, Piracicaba, São Paulo, Brazil.
Postal code: 13414- 903
Phone number: +55 19 21065320
E-mail: alan@unicamp.br

Competing Interests statement: None to declare.

Abstract

The proper assessment of oral potentially malignant disorders (OPMD) enrolls the oral epithelial dysplasia (OED) grading. In this context, patients with higher grades of OED have increased risk for cancer progression. Dysplasia grading systems for OED are a source of disagreement among pathologists and Machine Learning (ML) can aid in the oral cancer screening, improve early detection of OPMD and, consequently, improve diagnostic and treatment. This is a cross-sectional study is composed by a cohort of 82 patients with OPMD and correspondent 98 H&E-stained whole slide images (WSI) with biopsied-proven dysplasia. All WSI were manually annotated based on the Binary System for OED. The annotated regions of interest (ROI) were segmented and fragmented into small patches (299x299 pixels) and non-randomly sampled into development data and test data. The development data was color augmented, resulting in a total of 81,786 (32,608 LR patches and 49,178 HR patches) for training. The holded-out test set enrolled a total of 4486 patches (HR=2.724; LR=1.762). State-of-the-art convolutional neural networks (CNNs) as ResNet50, InceptionV3, VGG16, Xception, MobileNet, DenseNet, and EfficientNetB0 were trained, validated and tested with the same datasets. Results: Almost all of the models presented a high learning rate, yet very low generalization potential. At the model development, VGG16 performed the best, but with massive overfitting. In the test set, VGG16 presented the best accuracy, sensitivity, specificity, and AUROC (62%, 62%, 66%, and 65%, respectively), associated with the higher loss among all CNNs tested. EfficientB0 has comparable metrics and the lowest loss among all CNNs, being a great candidate for further studies. Using the present annotation methodology, the models were not able to generalize enough to be applied in real-life datasets due an overlapping of features between the two classes, which can be a confounding factor for the CNN training.

Keywords: Artificial Intelligence; whole slide images; leukoplakia; erythroleukoplakia; dysplasia grading.

Introduction

Oral potentially malignant disorders (OPMD) are defined by the World Health Organization[1] (WHO) as a group of lesions that present an increased risk to develop oral squamous cell carcinoma (OSCC), the most common cancer in head and neck, and are clinically characterized by leukoplakia and its variants (verrucous proliferative leukoplakias, erythroplakias, etc.). The malignization risk ranges from 3% in homogeneous leukoplakias to 14.5% non-homogeneous leukoplakias within a period of five years[2] with some authors indicating percentages ranging from 2.6 to 29.2%, depending on the dysplasia grade and with malignant transformation occurring in the dysplastic site[3]. In this context, patients with OPMD requires proper follow-up to manage biopsy-proven dysplastic lesions accordingly (i.e., surgical excision, CO_2 laser therapy, photodynamic therapy), and professionals should be aware of recurrency rate of 30%[4].

Dysplasia grading systems for oral epithelial dysplasia (OED) are a source of disagreement among pathologists, and Machine Learning (ML) models to assist in the malignant transformation prediction are being developed[5]. These methods have great potential to overcome these limitations regarding assessment of OED, and can aid in the diagnosis and proper management of OPMD, allowing close surveillance for cancer progression. In the context of OPMD, previous work reported traditional ML-based approaches for segmentation[6] and classification[7–10]of cell components/dysplasia identification with accuracy values varying from 88,69% to 95.7% depending on the feature descriptor and classifier used. Baik[11] developed a new algorithm to identify nuclear phenotypic changes based on morphometric data and voting scores to discriminate normal and abnormal nuclei to differentiate OPMD at low risk from those at high risk, which showed 80% classification rate at cellular level. Moreover, histopathological DL-based models were reported for segmentation[12,13] and prognostication[5,14–17] of OSCC.

The present study is based on the TRIPOD[18,19], a checklist for reporting diagnostic and prognostic prediction modelling studies, and aimed to develop and validate a DL-based model for OED grading accordingly to the Binary Classification[1,20]. The present work is the first to provide DL results for OED grading based on histopathological images.

Methods

Dataset

This is a cross-sectional diagnostic modelling study in which the study cohort is composed by 82 patients sourced from the primary care of three institutions from three different regions in Brazil: 42 patients from the Piracicaba Dental School -FOP (Piracicaba, São Paulo, Brazil), 19 patients from the Federal University of Minas Gerais - UFMG (Belo Horizonte, Minas Gerais, Brazil), 21 patients from the Federal University of Pará - UFPA (Belém, Pará, Brazil). The dataset comprises a total of 98 H&E-stained whole slide images (WSI) of OPMD (i.e., leukoplakia, erythroleukoplakia, proliferative verrucous leukoplakia) with all grades of dysplasia retrieved between 2005 and 2020. All patients should have at least one year of follow-up and complete photographic documentation. Patients with oral lichen planus were not considered as OPMD given the controversial nature of this lesion. Two patients with a medical history of Lupus and Fanconi's anemia were also excluded. Regarding dysplasia grading, patients presenting mild, moderate and severe/carcinoma in situ)[21] were included. Patients with micro or frankly invasive carcinoma were not included.

The glass slides were scanned using the Aperio Digital Pathology System (Leica Biosystems, Wetzlar, Germany) with a spatial sampling of 0.47μm per pixel, with automated focusing and magnification at ×20. WSI were primarily diagnosed by several pathologists (OPA, PAV, FPF, RAM, HARP) with vast experience, and according to the three-level WHO 2005 system[21], which is based on the architectural and cythological changing and the

distribution of atypia within the epithelium. For the present research, all cases were blindly and independently reviewed by ALDA and BALAM to assess dysplasia grade according to the BS for OED grading[1,20]. The BS improves interobserver agreement over the three-level method by reducing the complexity of the grading scheme, limiting the experienced-based subjective interpretation and allowing reproducibility among pathologists[22–24].

The BS proposed by Kujan[20] and further recommended by the WHO 2017[1] evaluates: i) architecture changes (irregular epithelial stratification, loss of polarity of basal cells, drop-shaped rete ridges, increased number of mitotic figures, abnormally superficial mitoses, premature keratinization in single cells, presence of keratin pearls, and loss of epithelial cohesion); and ii) cytological changes (abnormal variation of nuclear and cell size and shape, increased nuclear-cytoplasm ratio, presence of atypical mitosis, increased number and size of nucleoli and hyperchromatism). The cut-off for LR lesions is to present less than four architectural and/or less than five cytological changes.

Pre-processing, data augmentation and sampling

Representative areas of both classes compose the region of interest (ROI), which were classified and manually annotated by an expert (ALDA) accordingly to the pre-defined architectural and cytological criteria, with some patients' biopsies having annotations of both high risk (HR) and low risk (LR) of malignization, and avoiding major artifacts (folds, tears). The H&E staining of slides was widely variable **(figure 1).**

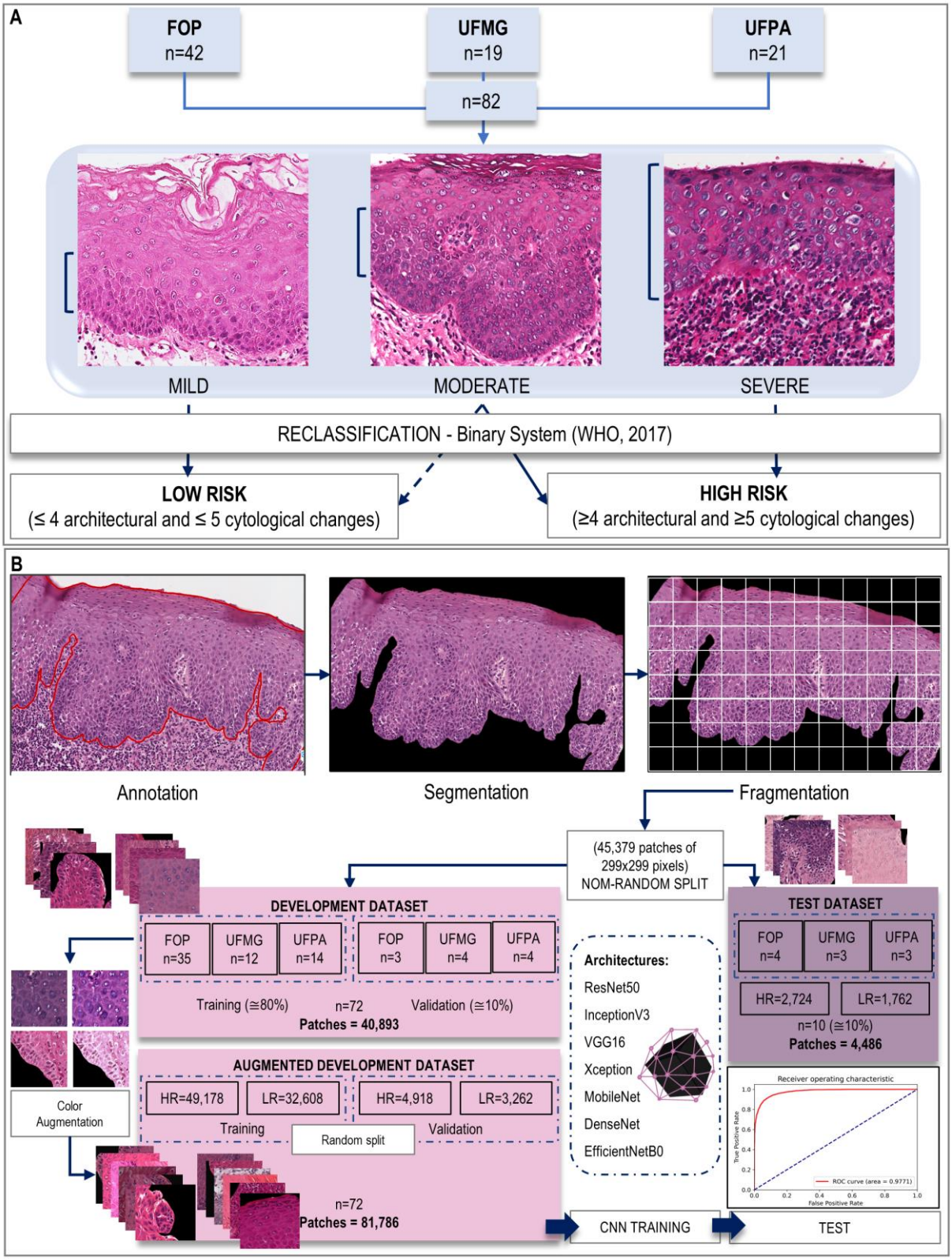


Figure 1. Methodology

The annotated ROI were segmented and fragmented into small patches (299x299 pixels). This image size is supported by the chosen architectures' inputs, and it is sufficient to provide important histopathological elements for classification. Initially, a total of 45,379 patches were non-randomly split into, 90% (40,893 patches) for models' development, and 10% (4,486 patches) for the testing set, with special attention to preserve the test set (10 patients) unseen to evaluate models' diagnostic performances. This sampling method is considered an intermediary between "internal" and "external validation" and characterizes the present study as of type 2b[18,19]. Additionally, in accordance with the *Guidelines for less biased data collection and algorithm evaluation*[25], this nom-random strategy aims to provide an equitable balance in terms of the institutions of origin (staining variations), population (biological variance), and classes (HR:LR) for each set while maintaining the right pre-determined proportion of patches.

In the next step, the 40,893 patches reserved for developing the model were randomly split into 36,804 for training and 4,089 for validation. To increase the robustness of the predictive model and account for staining variations influence in models' performances, the images of the training and validation set were submitted to color augmentation by randomly altering up to 50% of the image intensity (saturation) and brightness, and by adding variations in the color distribution of red, green, and blue (RGB) channels. This approach reproduces real-life technical variations (stainning, image acquisition) and diminishes the possibility of the staining scheme being outstandingly accountable for classification. Therefore, each patch generated one correspondent augmented patch with random variable parameters, resulting in a total of 81,786 (32,608 LR patches and 49,178 HR patches used for training and validation. The high number of samples of both classes is important to provide a rich variety of complementary clinical information, minimizing the possible effect of the HR and LR

unbalancing **(figure 1)**. Accordingly, given the imbalance of LR and HR, patches, a corresponding normalization was performed.

The holded-out test set enrolled a total of 4486 patches (HR=2.724; LR=1.762).

Architectures and implementation

State-of-the-art CNNs widely validated for several applications as ResNet50[26], InceptionV3[27,28], VGG16[29], Xception[30], MobileNet[31], DenseNet[32], and EfficientNetB0[33] were trained, validated and tested with the same datasets **(figure 1).** To provide an even comparison among models, architectures were implemented based on their original publication; hence, the activation functions, number of convolutional layers, kernels, neurons, layers, and sublayers of the fully connected (FC) layer were kept as the original publications. Accordingly, during training procedure, the same optimizer (Adam) with a learning rate lr=0.0001, and 75 epochs were used, and predefined weights such as “imageNet” were not applied; thus, a unique input size (299x299 pixels) could be used to every structure.

The algorithm was implemented using Python 3.6 and several open-source libraries specific to machine learning and image processing (TensorFlow, Keras, Scikit-Learn, and OpenCV). The training was carried out until accuracy stabilized and validation loss reduced it variation.

CNN training was conducted using an Intel CORE i7 3.50 GHz computer processor with 32GB RAM and 1TB, available at the Signal and Image Processing Laboratory at the Institute of Science and Technology, Federal University of São Paulo (ICT-Unifesp). After training/validation, the corresponding CNN models were fed by the input patches of the test set.

Results

Training and validation

The training and validation accuracy and loss curves are illustrated in **figure 2.** Almost all of the models presented a high learning rate, yet very low generalization potential. Specifically, regarding learning capacity, ResNet50 and InceptionV3 behave similarly, increasing training accuracy over time and stabilizing around 30th epoch, with validation accuracy presenting more instable in ResNet50. The VGG16 and Xception quickly reached training accuracy over 95% before the 10th epoch, indicating that these two architectures have a great learning potential with fast improvement, but with similar unstable validation accuracy values. DenseNet improved accuracy slowly until reached stability around the 50th-60th epochs, while MobileNet and EfficientNetB0 did not stabilized their training accuracy curves within the given training time. Despite EfficientNetB0 having required a long time to learn (i.e., to improve accuracy), it also presented the most stable validation accuracy curves among all CNNs.

The training loss curves of ResNet50, InceptionV3, VGG16, Xception, and DenseNet behave as expected, but the gap between training and validation loss indicates overfitting. Additionally, validation loss curves of the aforementioned CNNs show severe instability with drastic ups and downs, which indicates unrepresentative validation data when compared to what the models were learning to interpret the training data.

MobileNet presented a divergent gap between training and validation loss curves, indicating significant overfitting. EfficientNetB0 took a long time to learn (improve accuracy) but also had the best validation and training loss curves among all CNNs, presenting the lowest loss curve since the beginning of the training, as well as the most stable validation loss curve with less variation and few overfitting.

In Summary, ResNet50, InceptionV3 and DenseNet performed similarly. VGG16, Xception and MobileNet presented expressive overfitting, with MobileNet showing the worst results. EfficientNetB0 may be promisor for the present classification task but requires further investigation with specific adaptations for our purpose. Among all CNNs, VGG16 performed the best, since it learned quickly and maintained the accuracy during training. However, the loss indicates a massive overfitting.

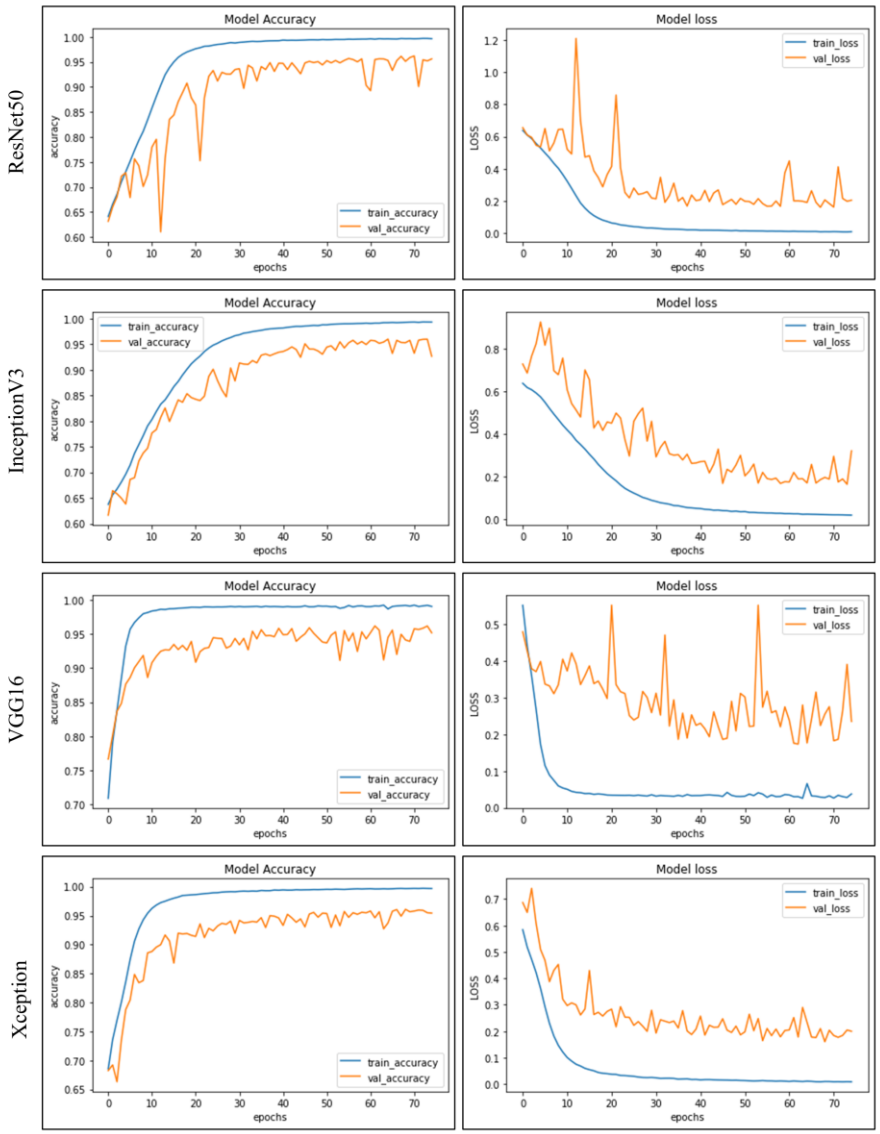


Figure 2. Training and validation metrics (accuracy and loss) (to be continued)

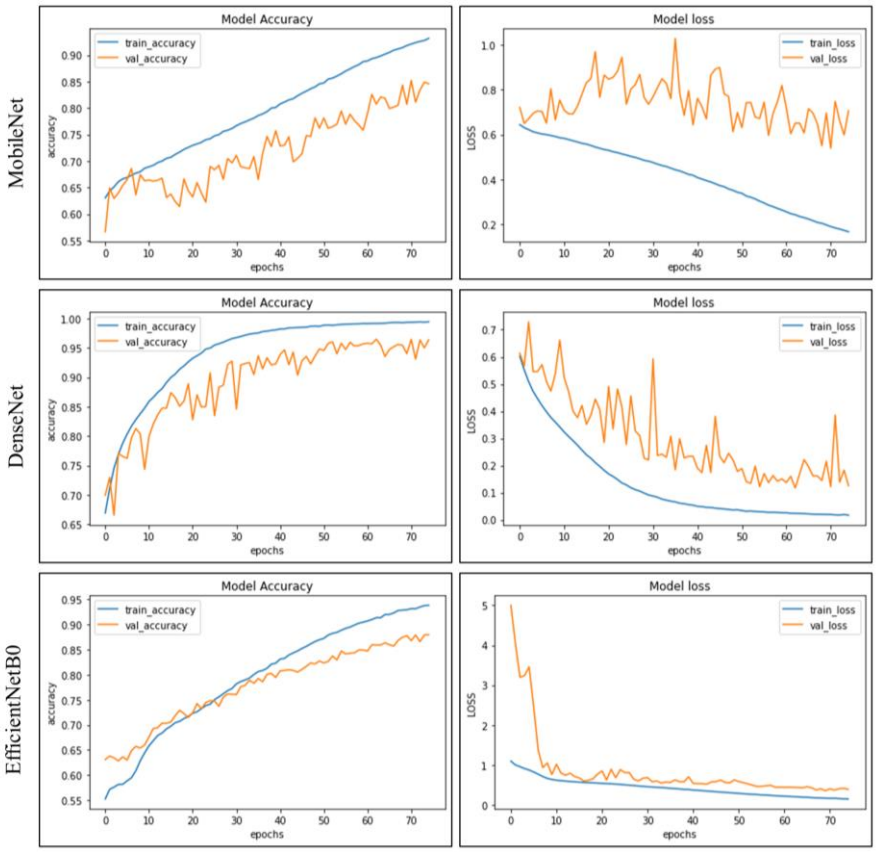


Figure 2. Training and validation metrics (accuracy and loss) (continuation)

Test

Test metrics are summarized in **table 1,** which specifically corroborates the low generalization potential, yet showing important differences. VGG16 presented the best accuracy, sensitivity, specificity, and AUROC (62%, 62%, 66%, and 65%, respectively), associated with the higher loss among all CNNs tested. EfficientB0 has comparable metrics and the lowest loss among all CNNs, being a great candidate for further studies. The confusion matrix displays how predictive is the model, allowing evaluation of false positives (FP)/type I errors and false negatives (FN)/type II errors, while the AUROC curves show the differentiation behavior in the classification task (**figure 3)**.

Table 2. Test performances metrics

	ResNet 50	**InceptionV3**	**VGG16**	**Xception**	**MobileNet**	**DenseNet**	**EfficientNetB0**
Time*	114s	128s	123s	116s	**49s**	146s	102s
Loss	3,60	3,90	5,00	3,60	3,00	3,50	**1,90**
Accuracy (Total)	0,54	0,55	**0,62**	0,55	0,57	0,54	0,58
Accuracy (Normalised)	0,55	0,59	**0,62**	0,55	0,59	0,56	0,56
TP	1400,00	1100,00	1700,00	1400,00	1300,00	1300,00	1500,00
TP%	0,53	0,41	0,62	0,50	0,49	0,47	0,55
FP	780,00	410,00	670,00	680,00	540,00	610,00	670,00
FP%	0,44	0,23	0,38	0,39	0,31	0,35	0,38
TN	980,00	1400,00	1100,00	1100,00	1200,00	1200,00	1100,00
TN%	0,56	0,77	0,62	0,61	0,69	0,65	0,62
FN	1300,00	1600,00	1000,00	1400,00	1400,00	1400,00	1200,00
FN%	0,47	0,59	0,38	0,50	0,51	0,53	0,45
Precision	0,65	**0,73**	0,71	0,67	0,71	0,68	0,69
Sensitivity	0,53	0,41	**0,62**	0,50	0,49	0,47	0,55
Specificity	0,56	0,77	0,62	0,61	**0,69**	0,65	0,62
F1Score	0,58	0,53	**0,66**	0,57	0,58	0,56	0,61
AUC	0,59	0,63	**0,65**	0,59	0,61	0,61	0,63

*To classify 4486 patches from the test set

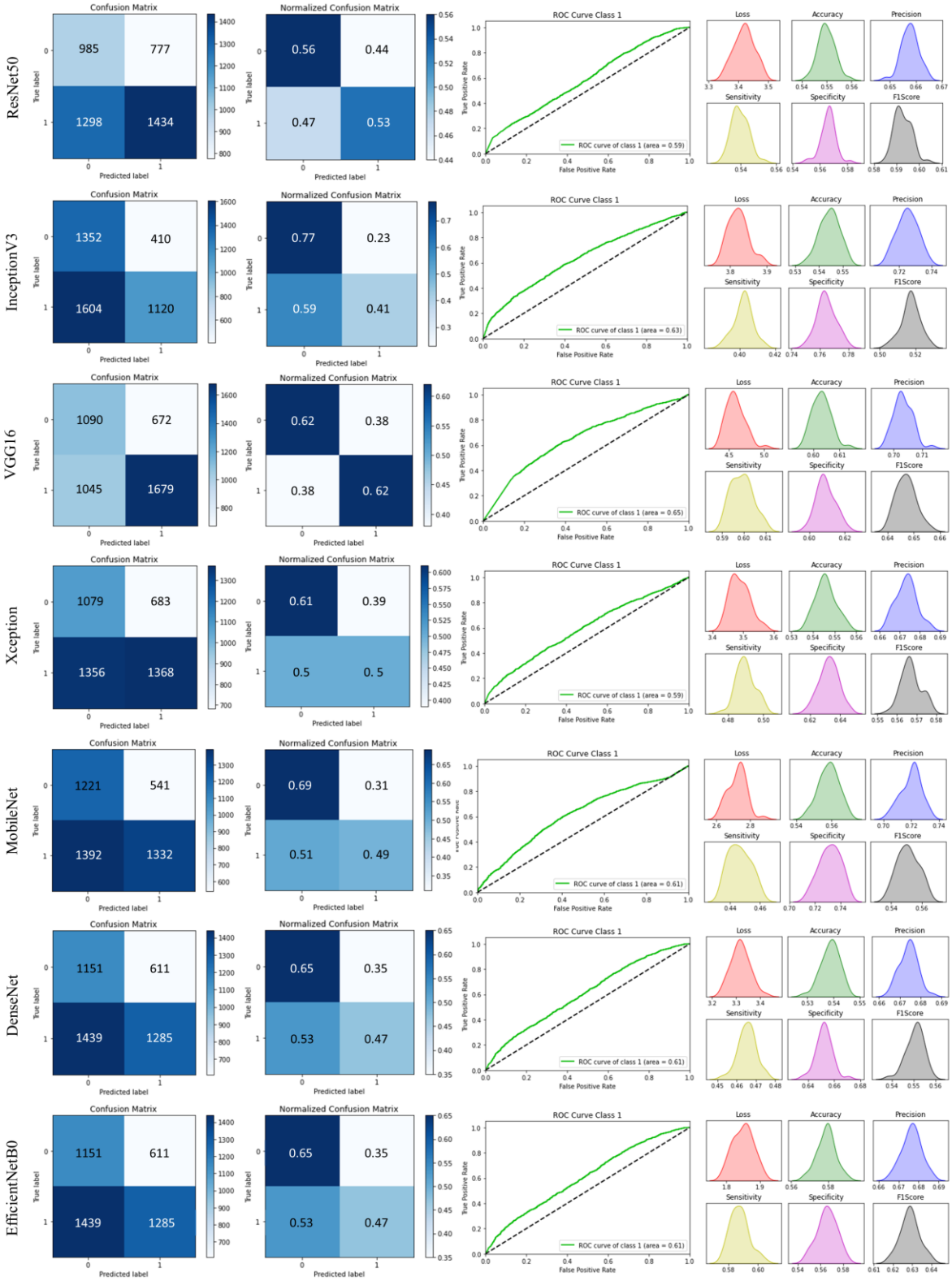


Figure 3. Test metrics (confusion matrix and AUROC curves), and metrics of an additional test with augmented data.

After the test, an additional analysis to test the variability of the augmentation was conducted by testing the models for 40 loops with augmented data. In each individual loop, the original test dataset is randomly transformed into a new augmented test dataset. Accordingly, a new test procedure is carried out; hence, the parameters of performance corresponding to the regarding loop are computed. As it can be seen from the regarding distributions (figure 3), the results were very similar to the one using the original data (table 1). From these results, it is possible to conclude that only the augmentation is not enough to prevent overfitting and improve generalization. In other words, this overfitting demonstrated that the learned characteristics to classify the training data were not probably relevant to potentially generalize the result.

Discussion

The diagnosis of HNC and OPMD by using classic ML-based models for classification and segmentation of histopathological image analysis have been reported by several groups[6–10,34–36], with only a few DL-based studies reported[12,14]. The present study is the first to apply Deep CNN for OED grading based on the BS.

The BS has sensitivity of 85%, specificity of 80% and accuracy of 82%, and greatly improved the grading of moderate dysplasia cases as high-risk progression cases that should be clinically supervised. Moreover, this system presents prognostic value, allowing outcome prediction (malignization) of 85% of patients[20]. The present protocol is based on the BS in accordance with evidence suggesting this system improves interobserver agreement and reproducibility among pathologists when compared to the TLS[22–24], but a recently published systematic review has demonstrated that this association remains inconclusive.

To improve reproducibility and interobserver agreement, Nankinvell[37] suggested a cut-off point up to four architectural and cytological changes. In the present work, the authors

considered the original cut-off points of four architectural and five cytological changes to preserve a more balanced proportion of HR and LR, since a natural imbalance is already seen with the majority of cases being classified as HR. This bias occurs because stable clinical lesions that are usually present LR malignization risk are not commonly biopsied until clinical changes became drastic.

The separation between training and validation loss curves, especially if the gap has a divergent tendency as in MobileNet, indicates that the model is overfitting since it works well on the training data but not quite in the validation data. This separation means the model it is not generalizing enough to perform well in external data. The representation of the validation loss curve over the training loss curve also can indicate that validation data is harder for the model than training data. In this context we also can see up and down jumps in the validation loss curves in six of the seven tested models, which indicates that the validation data is not representative when compared to the training data. A possible solution to this is the addition of dropout by setting a certain percentage of the neurons to zero, not using all neurons during training, but effectively using all neurons during validation. This should lead the model to be more robust in validation, and as such, the validation loss curve would be lower than the training loss curve.

The proper assessment of cytological and architectural criteria aims to provide an objective OED grading, but authors report interobserver agreement among 62% and 90%[23,38]. The authors acknowledge that support systems for OED grading should overcome this percentage but particularities in the data curation and the investigated disorders should be taken into account. A statistically significant increase in nuclear volume density, nuclear-cell ratio, nuclear area and perimeter are found in dysplastic cells when compared within different dysplasia grades, with higher values in increasing grades[39]. This premise corroborates that shape and size of nuclei can help in the differentiation among dysplasia grading. The CNN

convolutional layer filters relevant features for the classification task and the flattening function provides values corresponding to each class. Therefore, a CNNs should be able to correlate the features and learn the differences that stand for HR and LR. However, the way annotations are made, enrolling all epithelium extension, and the fragmentation process may insert bias by providing a great intersection of features that characterize both OED classes when using the BS grading system. Since the cut-offs rely solely on the quantification of the cytological and architectural changes, some cases may present severe changes (attending to the cut-off criteria for high malignization risk) but yet, presenting alterations limited to the basal cell layer, with superior epithelium levels preserved (**figure 4**). In this scenario, patches originating from dysplastic areas at the basal layer will be correctly labelled as HR, while patches from the abovementioned area may display not enough changes to fit in HR classification. Ultimately, an annotated ROI can generate widely different patches according to the epithelium level. These patches may present characteristics common to both classes but annotated under the same label, which can generate confusing "gold standard" with possible correlated information, confusing the CNNs, impairing learning and favoring "memorization" and overfitting.

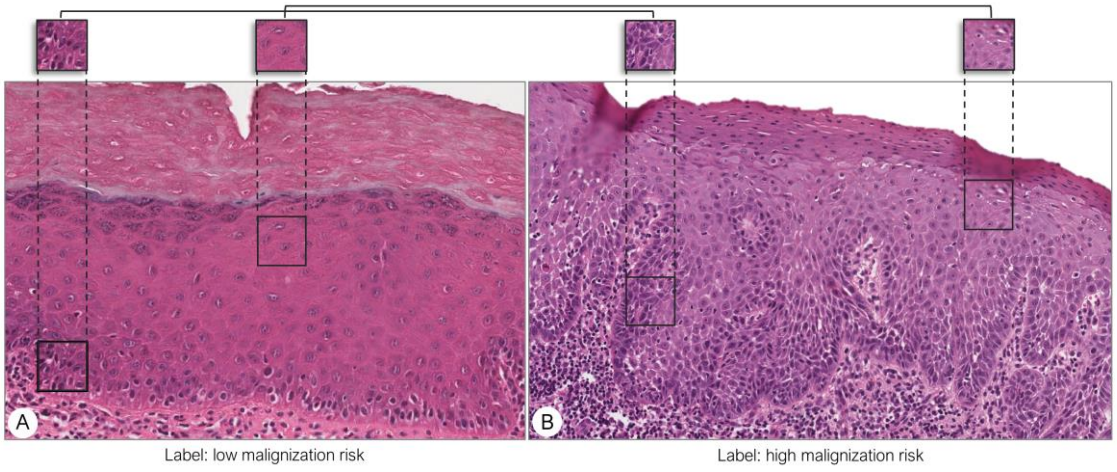


Figure 4. Pitfall explanation. A. Dysplastic lesion presenting mild dysplasia and labelled as low risk, and B. lesion presenting moderate dysplasia and labelled as high risk, according to

the Binary System. Note that generated patches are very similar and can be present in both classes according to the extension of dysplasia within the epithelium, despite the labels provided in the context of the architectural and cytological changes. These areas can be a confounding factor for the CNN training, resulting in overoptimistic results in training as the CNN "memorizes" patches. Additionally, since both patterns will be present at some level independently on the ground truth label, accuracy at patch-level may not be any better than chance (around 50%).

The authors also considered using the three-level grading[21], but this system is based on the distribution of dysplastic characteristics along the epithelium levels, requiring analysis of the full epithelium architecture to assess dysplasia extension. Since histopathological images are fragmented to fit the CNNs' kernels, the proper application of both BS (as conducted in the present work) and the three-level system is compromised. Maybe a more assertive methodology is to use the outcome as image labels, not the dysplasia grade. Alternatively, authors can include only the dysplastic area in annotations of ROIs and label the patient according to the BS to reduce inter-variability as much as possible. It is expected that the predominance of alterations should be enough to allow differentiation between LR and HR lesions.

The use of slides from different institutions and the technical differences in H&E preparation (e.g., section thickness and staining variation) represent real aspects of clinical pathology routine, which brings the developed model closer to a real-life condition of use and raises the challenge of achieving good results and increases the robustness of the model. Color augmentation has the potential of improving the model's generalization ability by creating samples with a wider variety of color distribution for training; hence, artificially reproducing

laboratory technical differences in slide preparation, our results show that augmentation alone was not enough to improve the performance of the tested model's system for our purpose.

We tested seven important DL-based systems for binary classification of epithelial dysplasia in histopathological images of OPMD. This is the first published work to apply and evaluate performance of known Deep CNN for OED grading based on the BS. The use of DL models is due to eliminating inter-pathologist variability in the analysis of OED, a known pitfall. Our test results indicate that, using the present annotation methodology (including all epithelium, with labels based on the BS) the models, in their original structure, were not able to generalize enough to be applied in real-life datasets, despite the great learning capacity of these models. Additionally, dysplasia grading alone does not predict malignant transformation[40], and requires clinical correlation. This may limit these models' applications, since the recognition of dysplasia are limited to the moment of the biopsy, not taking into account treatments and the outcome. Nonetheless, the proposed investigation and results provided are important to be evolved, and the accumulated knowledge and information may be the starting point for researchers on this subject. Future work will compare different annotations methodologies (based on the patient outcome and annotations limited to the dysplastic area), as well as to develop and implement a novel DL architecture to recognize dysplasia in histopathological images. Additionally, we will investigate the use of multi-data (clinical images, histopathological images and clinical data) to assess prognostication and ultimately predict if a patient will develop OSSC. Moreover, investigations on the ability of DL models to differentiate high and low malignization risk in clinical images from OPMD and differentiate OPMD clinical images from incipient lesions of oral squamous cell carcinoma are currently under investigation.

Acknowledgements

The authors would like to gratefully acknowledge the financial support of the Coordination for the Improvement of Higher Education Personnel (CAPES/PROEX, Brazil), the National Council for Scientific and Technological Development (CNPq, Brazil) and the grants from São Paulo Research Foundation (FAPESP, Brazil) process number: 2009/53839-2, which supported the acquisition of the equipment used.

Conflict of interest

We declare that the authors have no financial relationship with any commercial associations, current and within the past five years, that might pose a potential, perceived or real conflict of interest. These include grants, patent licensing arrangements, consultancies, stock or other equity ownership, advisory board memberships, or payments for conducting or publicizing our study.

Ethical Approval

This study was performed in accordance with the Declaration of Helsinki and was approved by the Piracicaba Dental Ethical Committee, Registration number CAAE: 42235421.9.0000.5418, which also comprised Material Transfer Agreements between co-participant Institutions to share digital slides.

Author contribution

All authors had substantial contributions to the conception (ALDA, VM, MCM, ARSS), draft and design (ALDA, VM, MCM), acquisition (FPF, MSS, RAM, BALAM, HARP, LCDS, CSS), analysis (ALDA, VM, MCM) and interpretation (ALDA, VM, MCM, HAA, SAK, ATP, MDM, MAL, PAV, LPK, ARSS) of data for the work.

Funding

This study was funded by the Coordination for the Improvement of Higher Education Personnel (CAPES/PROEX, Brazil) process number: 001, the National Council for Scientific and Technological Development (CNPq, Brazil), and the grants from São Paulo Research Foundation (FAPESP, Brazil) process number: 2009/53839-2, which supported the acquisition of the equipment used.

Data Availability Statement:

The datasets and used and/or analyzed during the current study are available from the corresponding author on reasonable request. All author agrees to be accountable for any aspects of the work and we ensure that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. El-Naggar AK, J. K. G. J. T. T. S. P. WHO Classification of Head and Neck Tumours. (Lyon: IARC, 2017).
2. Warnakulasuriya, S. Clinical features and presentation of oral potentially malignant disorders. Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology vol. 125 582–590 (2018).
3. Hankinson, P. M., Mohammed-Ali, R. I., Smith, A. T. & Khurram, S. A. Malignant transformation in a cohort of patients with oral epithelial dysplasia. British Journal of Oral and Maxillofacial Surgery 59, 1099–1101 (2021).
4. van der Waal, I. Potentially malignant disorders of the oral and oropharyngeal mucosa; present concepts of management. Oral Oncol 46, 423–425 (2010).
5. Mahmood, H., Bradburn, M., Rajpoot, N., Islam, N. M., Kujan, O. & Khurram, S. A. Prediction of malignant transformation and recurrence of oral epithelial dysplasia using

architectural and cytological feature specific prognostic models. Modern Pathology (2022) doi:10.1038/s41379-022-01067-x.

6. Muthu Rama Krishnan, M., Choudhary, A., Chakraborty, C., Ray, A. K. & Paul, R. R. Texture based segmentation of epithelial layer from oral histological images. Micron 42, 632–641 (2011).

7. Muthu Rama Krishnan, M., Pal, M., Bomminayuni, S. K., Chakraborty, C., Paul, R. R., Chatterjee, J. et al. Automated classification of cells in sub-epithelial connective tissue of oral sub-mucous fibrosis-An SVM based approach. Comput Biol Med 39, 1096–1104 (2009).

8. Krishnan MM, Acharya UR, Chakraborty C, Ray AK. Automated diagnosis of oral cancer using higher order spectra features and local binary pattern: a comparative study. Technol Cancer Res Treat. 2011 Oct;10(5):443-55. doi: 10.7785/tcrt.2012.500221. PMID: 21895029.

9. Krishnan, M. M. R., Venkatraghavan, V., Acharya, U. R., Pal, M., Paul, R. R., Min, L. C. et al. Automated oral cancer identification using histopathological images: A hybrid feature extraction paradigm. Micron 43, 352–364 (2012).

10. Krishnan, M. M. R., Shah, P., Chakraborty, C. & Ray, A. K. Statistical analysis of textural features for improved classification of oral histopathological images. Journal of Medical Systems vol. 36 865–881 (2012).

11. Baik, J., Ye, Q., Zhang, L., Poh, C., Rosin, M., MacAulay, C. et al. Automated classification of oral premalignant lesions using image cytometry and Random Forests-based algorithms. Cellular Oncology 37, 193–202 (2014).

12. Fraz, M. M., Shaban, M., Graham, S., Khurram, S. A. & Rajpoot, N. M. Uncertainty Driven Pooling Network for Microvessel Segmentation in Routine Histology Images. in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) vol. 11039 LNCS 156–164 (Springer Verlag, 2018).

13. Musulin, J., Štifanić, D., Zulijani, A., Ćabov, T., Dekanić, A. & Car, Z. An Enhanced Histopathology Analysis: An AI-Based System for Multiclass Grading of Oral

Squamous Cell Carcinoma and Segmenting of Epithelial and Stromal Tissue. Cancers (Basel) 13, 1784 (2021).

14. Shaban, M., Khurram, S. A., Fraz, M. M., Alsubaie, N., Masood, I., Mushtaq, S. et al. A Novel Digital Score for Abundance of Tumour Infiltrating Lymphocytes Predicts Disease Free Survival in Oral Squamous Cell Carcinoma. Sci Rep 9, (2019).

15. Lu, C., Lewis, J. S., Dupont, W. D., Plummer, W. D., Janowczyk, A. & Madabhushi, A. An oral cavity squamous cell carcinoma quantitative histomorphometric-based image classifier of nuclear morphology can risk stratify patients for disease-specific survival. Modern Pathology 30, 1655–1665 (2017).

16. Kim, D. W., Lee, S., Kwon, S., Nam, W., Cha, I. H. & Kim, H. J. Deep learning-based survival prediction of oral cancer patients. Sci Rep 9, (2019).

17. Bur, A. M., Holcomb, A., Goodwin, S., Woodroof, J., Karadaghy, O., Shnayder, Y. et al. Machine learning to predict occult nodal metastasis in early oral squamous cell carcinoma. Oral Oncol 92, 20–25 (2019).

18. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. Ann Intern Med 162, 55–63 (2015).

19. Moons, K. G. M., Altman, D. G., Reitsma, J. B., Ioannidis, J. P. A., Macaskill, P., Steyerberg, E. W. et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Ann Intern Med 162, W1–W73 (2015).

20. Kujan, O., Oliver, R. J., Khattab, A., Roberts, S. A., Thakker, N. & Sloan, P. Evaluation of a new binary system of grading oral epithelial dysplasia for prediction of malignant transformation. Oral Oncol 42, 987–993 (2006).

21. World Health Organization classification of tumours. Pathology and genetics of tumours of the head and neck. (IARC Press, 2005).

22. Kujan, O., Khattab, A., Oliver, R. J., Roberts, S. A., Thakker, N. & Sloan, P. Why oral histopathology suffers inter-observer variability on grading oral epithelial dysplasia: An attempt to understand the sources of variation. Oral Oncol 43, 224–231 (2007).

23. Speight, P. M., Abram, T. J., Floriano, P. N., James, R., Vick, J., Thornhill, M. H. et al. Interobserver agreement in dysplasia grading: toward an enhanced gold standard for clinical pathology trials. Oral Surg Oral Med Oral Pathol Oral Radiol 120, 474-482.e2 (2015).

24. Khoury, Z. H., Sultan, M. & Sultan, A. S. Oral Epithelial Dysplasia Grading Systems: A Systematic Review & Meta-Analysis. Int J Surg Pathol 30, 499–511 (2022).

25. Marée, R. The need for careful data collection for pattern recognition in digital pathology. J Pathol Inform 8, (2017).

26. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) vols. 2016-December 770–778 (IEEE, 2016).

27. Szegedy, C., Wei Liu, Yangqing Jia, Sermanet, P., Reed, S., Anguelov, D. et al. Going deeper with convolutions. in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 1–9 (IEEE, 2015). doi:10.1109/CVPR.2015.7298594.

28. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the Inception Architecture for Computer Vision. (2015).

29. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. (2014).

30. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) vols. 2017-January 1800–1807 (IEEE, 2017).

31. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T. et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. (2017) doi:https://doi.org/10.48550/arXiv.1704.04861.

32. Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q. Densely Connected Convolutional Networks. in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) vols. 2017-January 2261–2269 (IEEE, 2017).

33. Tan, M. & Le, Q. v. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. (2019).

34. Mahmood, H., Shaban, M., Indave, B. I., Santos-Silva, A. R., Rajpoot, N. & Khurram, S. A. Use of artificial intelligence in diagnosis of head and neck precancerous and cancerous lesions: A systematic review. Oral Oncology vol. 110 (2020).

35. Mete, M., Xu, X., Fan, C. Y. & Shafirstein, G. A machine learning approach for identification of head and neck squamous cell carcinoma. in Proceedings - 2007 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2007 29–34 (2007). doi:10.1109/BIBM.2007.57.

36. Folmsbee, J., Liu, X., Brandwein-Weber, M. & Doyle, S. Active deep learning: Improved training efficiency of convolutional neural networks for tissue classification in oral cavity cancer. in 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018) 770–773 (IEEE, 2018). doi:10.1109/ISBI.2018.8363686.

37. Nankivell, P., Williams, H., Matthews, P., Suortamo, S., Snead, D., McConkey, C. et al. The binary oral dysplasia grading system: Validity testing and suggested improvement. Oral Surg Oral Med Oral Pathol Oral Radiol 115, 87–94 (2013).

38. Ranganathan, K. & Kavitha, L. Oral epithelial dysplasia: Classifications and clinical relevance in risk assessment of oral potentially malignant disorders. J Oral Maxillofac Pathol 23, 19–27 (2019).

39. Prema, V., Thomas, T., Harikrishnan, P., Viswanathan, M., Srichinthu, K. & Rajkumar, K. Morphometric analysis of suprabasal cell layer in oral epithelial dysplasia: A computer-assisted microscopic study. J Pharm Bioallied Sci 12, 204 (2020).

40. Leeky, M., Narayan, T., Sadhana, S., Saleha, J. & Geetha, K. Grading of oral epithelial dysplasia: Points to ponder. Journal of Oral and Maxillofacial Pathology 19, 198 (2015).

2.3 Artigo: Clinicians' perception of oral potentially malignant disorders: a pitfall for image annotation in Deep Learning

Anna Luíza Damaceno Araújo[a], Eduardo Santos Carlos de Souza[b], Isabel Schausltz Pereira Faustino[a], Cristina Saldivia Siracusa[a], Tamires Brito-Sarracino[b], Marcio Ajudarte Lopes[a], Pablo Agustin Vargas[a], Syed Ali Khurram[c], Alexander T. Pearson[d,e], Luiz Paulo Kowalski[f,g], André Carlos Ponce de Leon Ferreira de Carvalho[b], Alan Roger Santos-Silva[a].

Affiliation:
[a] Oral Diagnosis Department, Semiology and Oral Pathology Areas, Piracicaba Dental School, University of Campinas (UNICAMP), Piracicaba, São Paulo, Brazil.
[b] Institute of Science and Technology, Federal University of São Paulo (ICT-Unifesp), São José dos Campos, São Paulo, Brazil.
[c] Unit of Oral and Maxillofacial Pathology, School of Clinical Dentistry, University of Sheffield, Sheffield, UK.
[d] Section of Hemathology/Oncology, Department of Medicine, University of Chicago, Chicago, IL, USA.
[e] University of Chicago Comprehensive Cancer Center, Chicago, IL, USA.
[f] Department of Head and Neck Surgery and Otorhinolaryngology, A.C. Camargo Cancer Center, São Paulo, Brazil.
[g] Head and Neck Surgery Department and LIM 28, University of São Paulo Medical School, São Paulo, Brazil.
[h] Institute of Mathematics and Computer Sciences of University of São Paulo (ICMC-USP), São Carlos, São Paulo, Brazil

Corresponding Author: Alan Roger Santos-Silva

Oral Diagnosis Department, Piracicaba Dental School, UNICAMP

Adress: Av. Limeira, nº 901, Areião, Piracicaba, São Paulo, Brazil

Postal code: 13414- 903

Phone number: +55 19 21065320

E-mail: alan@unicamp.br

Competing Interests statement: None to declare.

Abstract

Introduction: The use of Artificial intelligence (AI) for image analysis based on Supervised Learning (SL) requires image annotation to provide ground truth reference to train Deep Learning (DL) algorithms. This step should be accurately performed to provide a good reference for the model. The present study aims to evaluate and quantify the clinician's perception on the oral potentially malignant disorder (OPMD) to understand the source of interobserver variability while assessing these disorders. *Study Design:* A dataset of 46 clinical images from 37 patients clinically diagnosed with leukoplakias were reviewed, classified, and manually annotated at pixel level by three labelers. For the clinical criteria, we assessed the κ statistics (Fleiss's Kappa) to establish the interobserver agreement, and annotations were compared using mean pixel-wise Intersection Over Union (IoU). *Results:* The interobserver agreement for homogeneous/non-homogeneous criteria was considered substantial (κ=63, with 95% CI, ranging from 0.47 to 0.80). For the subclassification of non-homogeneous lesions, the interobserver agreement was considered moderate (κ=43, with 95% CI, ranging from 0.34 to 0.53) with p value <0.001. A mean IoU of 0.53 (±0.22 std) was obtained and considered low. *Conclusion*: These results demonstrate that there is an important disagreement among clinicians while evaluating OPMDs using known clinical criteria. The analysis of annotated images corroborates these findings. From this analysis, the authors acknowledge that there is a substantial probability of transferring the subjectivity of human analysis to AI models during training and recommend managing this bias by including at least two experienced clinicians in the dataset construction and to use the image of the union as the ground truth reference for CNN training.

Keywords: Supervised Learning; Oral Potentially Malignant Disorders; Oral Cancer; Head and Neck Cancer;

Introduction

Leukoplakia is a clinical diagnosis for asymptomatic white lesions defined as a patch or plaque with well-defined borders that cannot be scrapped off, which occurrence is not linked to traumatic events, and has been biopsy-proven not to be any other known white disorders. These lesions can be classified as homogeneous (slightly elevated, predominantly smooth with or without crack fissures), and non-homogeneous (irregular surface and/or mixed in colour), represented by white plaques associated with erosive (red) areas (i.e., speckled), as well as texture alterations (i.e., verrucous/exophytic or nodular). Additionally, there is also a subset of epithelial dysplasia (ED) lesions presenting lichenoid features and reticular formation [1,2] (figure 1).

At the histopathological examination, leukoplakias can vary from presenting no dysplasia to some grade of dysplasia, with some having a squamous cell carcinoma (SCC) diagnosis, even when presenting a misleading clinically indolent appearance. Additionally, non-homogeneous leukoplakias tends to have higher epithelial dysplasia (ED) grade and malignization risk rates [1]. This motivates investigation regarding the application of computer vision approaches to develop and implement image-based machine learning (ML) models for detection and classification of such confounding lesions to support oral medicinists in the screening and diagnosis of oral potentially malignant disorders (OPMD).

Therefore, an important step of data construction is the manual annotation (figure 2A) of clinical images by clinicians to separate only relevant information from the background. The segmented images (figure 2B and 2C) will provide a ground truth reference to train a DL model through supervised learning for segmentation. This is not an easy task if we consider the widely variable and highly subjective criteria to classify OPMD lesions based on eye-ball observation and personal experience.

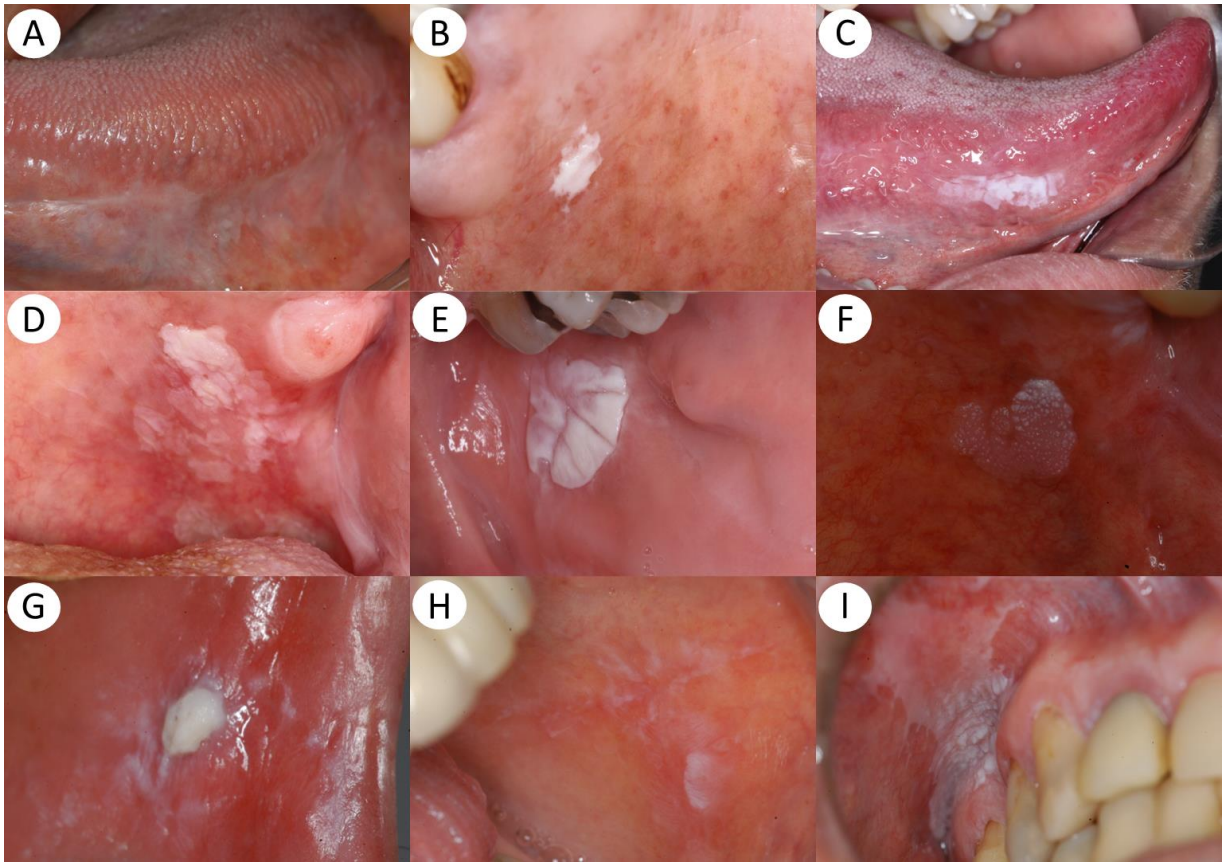


Figure 1. Clinical aspects of leukoplakia: thin smooth (A), thick smooth (B) and thick fissured (C) homogeneous leukoplakias. Non-homogeneous leukoplakias with speck-led (D), fissured (E), granular/with excrescencies (F), nodular (G), lichenoid (H), and verrucous (I) features.

The present study aims to evaluate and quantify the clinician's perception on the OPMD's clinical aspect, while evaluating, classifying and manually annotating photographic images of oral leukoplakias, as well as to understand the source of interobserver variability while assessing these lesions.

Methods

This cohort comprises 46 clinical photographs from 37 patients clinically diagnosed with oral leukoplakias between 2005 and 2020, and biopsied-proven as and presenting ED at the histopathological exam, retrieved from the Piracicaba Dental School (Piracicaba, São

Paulo, Brazil) archives. Patients with oral lichen planus were not included given the controversial nature of this lesion but ED with lichenoid features were included since it may represent a great pitfall in the present context. This study was performed in accordance with the Declaration of Helsinki and is a pilot study of the protocol approved by the Piracicaba Dental Ethical Committee, Registration number CAAE: 42235421.9.0000.5418, which also comprised Material Transfer Agreements between co-participant Institutions to share digital slides.

A total of 46 images were independently reviewed, classified and manually annotated at pixel-level by three labelers (ALDA, ISPF and CSS). Clinical characteristics and histological findings are shown in table 1.

Table 1. Clinical characteristics and dysplasia grade

Gender	
Male	35 (76%)
Female	11 (23%)
Location	
Tongue	21
Buccal mucosa	6
Floor of the mouth	3
Palate	9
Gingiva	2
Alveolar ridge	1
Tonsilary fossa	2
Vestibular fold	2
Histologic findings	
No dysplasia	0 (0%)
Mild	13 (28.2%)
Moderate	33 (71.7%)

The cohort was classified into homogeneous (slightly elevated white plaque, with discrete and shallow fissures) and non-homogeneous (more drastic alterations in texture and/or color). Lesions classified as non-homogeneous were clinically subclassified as speckled (red and white), fissured, nodular, verrucous, and lichenoid [1,2].

For the clinical criteria, we assessed the κ statistics (Fleiss's Kappa) to estab-lish the interobserver agreement, in which values of $\kappa < 0.00$ indicates poor agreement, 0.0–0.2 slight agreement, 0.2–0.4 fair agreement, 0.4–0.6 moderate agreement, 0.6–0.8 substantial or good agreement, and > 0.8 excellent or almost perfect agreement [3]. Statistical analyses were conducted using Real Statistics Resource Pack for Excell (Release 7.6) Copyright (2013 – 2021) Charles Zaiontz. www.real-statistics.com.

To evaluate the agreement between labelers' annotations, the mean pixel-wise Intersection Over Union (IoU) was calculated comparing three-paired annotation for each image and dividing Intersection Pixel Count by the union Pixel Count [4]. The mean IoU represents the coincident areas in all annotations, the area that all labelers considered as "lesion".

Results

The interobserver agreement for homogeneous/non-homogeneous criteria was considered substantial (κ=63, with 95% CI, ranging from 0.47 to 0.80). For the sub-classification of non-homogeneous lesions, the interobserver agreement was consid-ered moderate (κ=43, with 95% CI, ranging from 0.34 to 0.53). A mean IoU of 0.53 (±0.22 std) was obtained, which corroborates the Fleiss kappa analysis, ultimately pointing to a significative discrepancy among the three labelers' annotations.

Table 2. Fleiss kappa analysis

Metric	**Classification**[a]	**Subclassification**[b]
kappa	0.639	0.436
s.e.	0.085	0.048
z-stat	7.510	9.014
p-value	1.767	**<0.001***
lower	0.472	0.341
upper	0.806	0.531

[a]homogeneous/non-homogeneous; [b]speckled (red and white), fissured, nodular, verrucous, and lichenoid; *statistically significant

Table 3. Analysis of annotations' interobserver agreement.

Metrics	**Intersection Over Union**			
	1st and 2nd	**1st and 3rd**	**2nd and 3rd**	**All observers**
count	96.000000	97.000000	79.000000	272.000000
mean	0.691998	0.604695	0.636715	0.644808
std	0.195882	0.222336	0.208343	0.211753
min	0.063777	0.000000	0.000000	0.000000
25%	0.610898	0.491974	0.535257	0.550713
50%	0.730431	0.639990	0.670263	0.678382
75%	0.831222	0.770541	0.793060	0.803272
max	0.961309	0.943191	0.937528	0.961309

Count: cases counting; std: standard deviation; min: minimum; ma: maximum; 1st quartile (25%); 2nd quartile (50%); 3rd quartile (75%);

Discussion

In the domain of oral and maxillofacial diseases, there are a few previously published studies that aimed to develop an automated method to support oral medicine doctors in the screening and early identification of OPMD and OSCC in clinical photographs [5-15]. All approaches begin with detection and classification of objects, known as image segmentation. Automatic segmentation techniques are defined as unsupervised, in which the so called "classic" algorithms explores the data sets and drawn inferences, or supervised (i.e., based on image annotation, which is the core of ML-based algorithms).

Manual image annotation conducted by experienced medicinists is, therefore, an important step of supervised learning, which compares ground-truth annotation with the predicted segmented image. A pitfall identified in the present study was the difficulty in delimiting the borders of the lesions, especially in out-of-focus areas and locations were a poor vascularization gives a paler aspect to the mucosa (e.g., lateral border of the tongue, when the clinicians is pulling the tongue to take the photo) (figure 2), which could be a confounding factor. The present application faces not only the challenging identification of sometimes subtle white plaques, but also the identification of reddish nuances discriminative of erosive areas with great chances of presenting higher dysplasia grades in non-homogeneous leukoplakias. Camalan [5] considers manual annotation prone to error due inter and intra-labeller variability and proposed that automatic segmentation techniques can be valid alternatives to overcome this limitation. Our research was able to quantify variability in assessing OPMDs by measuring how differently such lesions are perceived by oral medicinists through interobservers' agreement of clinical criteria and by calculating the IoU of segmented images. The present appraisal provides a unique evaluation perspective of annotations' discrepancies and corroborates this already known subjectivity.

Unsupervised segmentation approaches as Clustering-based algorithms (e.g., k-means) partition the image by grouping the similar pixels in a pre-defined number of clusters (Figure 3A), while threshold-based algorithms (e.g., Otsu) groups the pixels according to intensity values by assessing the histogram in an automatic or static way (Figure 3B). Unsupervised segmentation approaches are usually associated with texture descriptors since relying only on the pixel value could exclude important areas from the segmented image, requiring extensive calibration steps. Additionally, the presence of "noise" as surgical retractors, gloved fingers, gauze, lips/skin/hair, teeth, and other lesions and oral conditions not related to the classification task, greatly impacts the use of automatic segmentation compromising the use of such algorithms are not ideal for the intended clinical application.

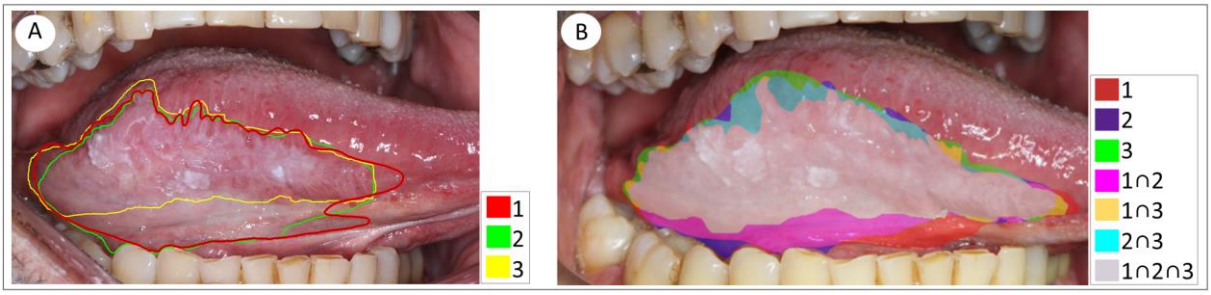


Figure 2. Annotation (A) and Intersection over Union for all labellers (B). Annotation. Note the disagreement regarding the inclusion of a white area in the base of the tongue, that can be either an ill-defined leukoplakia or ischemia due holding and pulling the tongue.

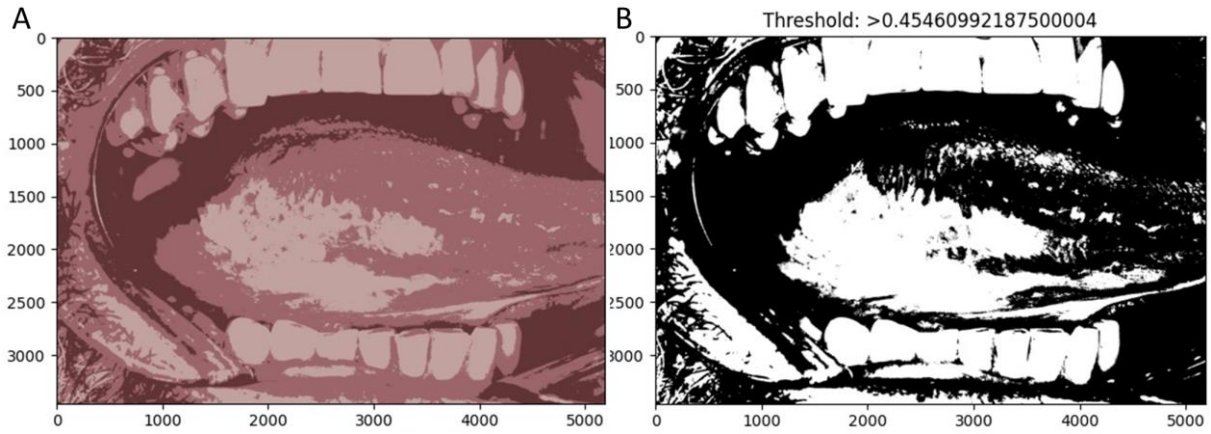


Figure 3: Unsupervised segmentation with K-means (A) and Otsu multi-tresholding algorithms.

Supervised segmentation approaches have been consistently applied in clinical images of oral cancer and OPMD. In these approaches, a CNN (e.g., U-Net) is trained with manually annotated images (ground-truth annotation) to learn how to proper segment valid image information from the background. These segmented images (area containing the lesion and/or normal areas) will be used to train a CNN for further classification. Ferrer-Sánchez [6] implemented U-Net for segmentation and a multi-task CNN for the classification of OPMD to assist in the prediction of malignant transformation. Semantic and instance segmentation were investigated by Tanriver [12], which further trained a CNN for benign, OPMD, and malignant classification. Semantic segmentation is considered pixel-wise, as it assigns each pixel to a class by grouping multiple objects in one category. In contrast, instance segmentation distinct each object within a category. Instance segmentation is considered an upgraded version of semantic segmentations, as it delineates object's boundaries by combining both object detection and semantic segmentation tasks. Warin [14] reported results of segmentation using Faster R-CNN (two-stage object detection algorithm) and YOLOv4 (one-stage object detection algorithm), and classification using DenseNet-121 and ResNet-50 for normal versus OPMD classification, in which the intersection of three labeller's annotations was used as the ground truth for CNN training. In a different approach, Warin [15] adopted Faster R-CNN, YOLOv5, RetinaNet and CenterNet2 to detect OSCC and OPMDs in clinical photographs. Alternatively, Thomas [13] applied active contour without edges, a semiautomatic approach that improves segmentation time.

The definitive diagnosis of oral leukoplakia should always enrol histopathological analysis to discard other known lesions, but there are reports of unexplained striking proportions (11 – 17%) of lesions diagnosed solely based on clinical skills [16,17]. Within the scope of OPMD, the clinician's perception variability is reported by some authors, with reports of 58.7% of practitioners facing difficulties in this diagnosis [18]. Over the past years,

clinicians were consistently prone to the improvement of leukoplakia identification from 46.15% in 1983 to 76% in 2015 [19].

Clinical criteria are widely variable among studies and the importance of establishing a good representation of the lesion conflates the concepts of non-homogeneous clinical appearance and a higher malignization risk [20]. In previous work, homogeneous leukoplakia classification allowed the inclusion of fissured lesions [21,22], while non-homogeneous leukoplakias are defined by a range of clinical presentations as erythematous, granular, nodular, polypoid or with excrescences lesions [22,23], with verrucous leukoplakia being classified separately as a non-homogeneous exophytic lesion with wrinkled or corrugated surface [24]. These wide range of clinical criteria suffered adjustments over the years [1,2] and may justify, along with the different educational backgrounds, the source of disagreement between the enrolled labellers. This is a variation that should be taken into account and handled.

Our results demonstrate an important interobserver disagreement while analysing OPMD using known clinical criteria evaluation. This known subjectivity is corroborated by the IoU quantification that reflect this diverse perception. From this analysis, the authors acknowledge there is a substantial probability of transferring the subjectivity of human analysis for CNNs during training. The use of large datasets holds the potential to overcome this limitation. However, despite being prone to interobserver variability, manual segmentation remains the best strategy to set the models with a close-to-real-life reference, providing individualized clinical data and contributing to build a robust model. Moreover, the authors understand the urge of research in oral medicine field and recommend managing this bias by including at least two experienced clinicians in the dataset construction and to use the image of the union as the ground truth reference for CNN training, since it is better to include false positive areas than to neglect true positive areas.

Conclusion

The present work aimed to assess clinical images annotation of OPMD and early OSCC to understand the sources of disagreements to aid in the understanding of interobserver variability during dataset construction.

The assessment of OPMD is subjective, with widely variable criteria that have changed over the years, and usually rely on eye-ball observation. Additionally, these clinical criteria are adopted in a heterogeneous way by clinicians and specialists who have different educational backgrounds and personal experience. These factors together explain the difference in the interpretation of these lesions.

Our results illustrated how the clinicians' perception can introduce bias in the ground truth annotations used to train DL models for object segmentation and further classification, especially in the domain of bland, white, striated lesions. The authors accept the risk of including false positive reference in the segmented image and recommend using the union of more than one annotation as ground truth for CNN training. Standardization in ML-based methodologies for OPMD and early HNC diagnosis is required for a better assessment and accurate diagnosis. Further work will encompass the classification of the segmented images.

Acknowledgements

The authors would like to gratefully acknowledge the Coordination for the Improvement of Higher Education Personnel (CAPES/PROEX, Brazil) process number 001, the National Council for Scientific and Technological Development (CNPq, Brazil), and the grants from São Paulo Research Foundation (FAPESP, Brazil) process number 2019/26676-7.

Conflict of interest

We declare that the authors have no financial relationship with any commercial associations, current and within the past five years, that might pose a potential, perceived or real conflict of interest. These include grants, patent-licensing arrangements, consultancies, stock or other equity ownership, advisory board memberships, or payments for conducting or publicizing our study.

Ethics Approval

This study is in accordance with the Declaration of Helsinki and was approved by the Piracicaba Dental Ethical Committee, Registration number CAAE: 42235421.9.0000.5418.

Author contribution

All authors have made substantial contributions to the conception (ALDA, ARSS), draft and design (ALDA, ESCS, ISPF, CSS, TBS), and review (MAL, PAV, SAK, ATP, LPK, ACPLFC, ARSS) of the paper's final version. The authors agree to be accountable for all aspects of the work and ensures that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Funding

This study was funded by the Coordination for the Improvement of Higher Education Personnel (CAPES/PROEX, Brazil) process number 001, the National Council for Scientific and Technological Development (CNPq, Brazil), and the grants from São Paulo Research Foundation (FAPESP, Brazil) process number: 2009/53839-2, which supported the acquisition of the equipment used.

References

1. Warnakulasuriya S, Kujan O, Aguirre-Urizar JM, Bagan JV, González-Moles MÁ, Kerr AR, Lodi G, Mello FW, Monteiro L, Ogden GR, Sloan P, Johnson NW. Oral potentially malignant disorders: A consensus report from an international seminar on nomenclature and classification, convened by the WHO Collaborating Centre for Oral Cancer. Oral Dis. 2021 Nov;27(8):1862-1880. doi: 10.1111/odi.13704.
2. El-Naggar AK, J. K. G. J. T. T. S. P. *WHO Classification of Head and Neck Tumours*. (Lyon: IARC, 2017).
3. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–74
4. IoU
5. Camalan S, Mahmood H, Binol H, Araújo ALD, Santos-Silva AR, Vargas PA, Lopes MA, Khurram SA, Gurcan MN. Convolutional Neural Network-Based Clinical Predictors of Oral Dysplasia: Class Activation Map Analysis of Deep Learning Results. Cancers (Basel). 2021 Mar 14;13(6):1291. doi: 10.3390/cancers13061291.
6. Ferrer-Sánchez A, Bagan J, Vila-Francés J, Magdalena-Benedito R, Bagan-Debon L. Prediction of the risk of cancer and the grade of dysplasia in leukoplakia lesions using deep learning. Oral Oncol. 2022 Sep;132:105967. doi: 10.1016/j.oraloncology.2022.105967.
7. Figueroa KC, Song B, Sunny S, Li S, Gurushanth K, Mendonca P, Mukhia N, Patrick S, Gurudath S, Raghavan S, Imchen T, Leivon ST, Kolur T, Shetty V, Bushan V, Ramesh R, Pillai V, Wilder-Smith P, Sigamani A, Suresh A, Kuriakose MA, Birur P, Liang R. Interpretable deep learning approach for oral cancer classification using guided attention inference network. J Biomed Opt. 2022 Jan;27(1):015001. doi: 10.1117/1.JBO.27.1.015001.

8. Fu Q, Chen Y, Li Z, Jing Q, Hu C, Liu H, Bao J, Hong Y, Shi T, Li K, Zou H, Song Y, Wang H, Wang X, Wang Y, Liu J, Liu H, Chen S, Chen R, Zhang M, Zhao J, Xiang J, Liu B, Jia J, Wu H, Zhao Y, Wan L, Xiong X. A deep learning algorithm for detection of oral cavity squamous cell carcinoma from photographic images: A retrospective study. EClinicalMedicine. 2020 Sep 23;27:100558. doi: 10.1016/j.eclinm.2020.100558.
9. Jubair F, Al-Karadsheh O, Malamos D, Al Mahdi S, Saad Y, Hassona Y. A novel lightweight deep convolutional neural network for early detection of oral cancer. Oral Dis. 2022 May;28(4):1123-1130. doi: 10.1111/odi.13825.
10. Jurczyszyn K, Gedrange T, Kozakiewicz M. Theoretical Background to Automated Diagnosing of Oral Leukoplakia: A Preliminary Report. J Healthc Eng. 2020 Sep 13;2020:8831161. doi: 10.1155/2020/8831161
11. Shamim M, Syed S, Shiblee M, Usman M, Ali S. Automated detection of oral pre-cancerous tongue lesions using deep learning for early diagnosis of oral cavity cancer. Oxford University Press: Computer Journal. 2019. doi:10.13140/RG.2.2.28808.16643.
12. Tanriver G, Soluk Tekkesin M, Ergen O. Automated Detection and Classification of Oral Lesions Using Deep Learning to Detect Oral Potentially Malignant Disorders. Cancers (Basel). 2021 Jun 2;13(11):2766. doi: 10.3390/cancers13112766.
13. Thomas B, Kumar V, Saini S. Texture analysis based segmentation and classification of oral cancer lesions in color images using ANN. 2013 IEEE International Conference on Signal Processing, Computing and Control (ISPCC), 2013, pp. 1-5, doi: 10.1109/ISPCC.2013.6663401.
14. Warin K, Limprasert W, Suebnukarn S, Jinaporntham S, Jantana P. Performance of deep convolutional neural network for classification and detection of oral potentially malignant disorders in photographic images. Int J Oral Maxillofac Surg. 2022 May;51(5):699-704. doi: 10.1016/j.ijom.2021.09.001.

15. Warin K, Limprasert W, Suebnukarn S, Jinaporntham S, Jantana P, Vicharueang S. AI-based analysis of oral lesions using novel deep convolutional neural networks for early detection of oral cancer. PLoS One. 2022 Aug 24;17(8):e0273508. doi: 10.1371/journal.pone.0273508.
16. Epstein, J. B., Gorsky, M., Fischer, D., Gupta, A., Epstein, M., & Elad, S. (2007). A survey of the current approaches to diagnosis and management of oral premalignant lesions. *Journal of the American Dental Association*, 138, 1555–1562 quiz 1614.
17. Pentenero M, Sutera S, Lodi G, Bagan JV, Farah CS. Oral leukoplakia diagnosis and treatment in Europe and Australia: Oral Medicine Practitioners' attitudes and practice. Oral Dis. 2022 Jul 6. doi: 10.1111/odi.14301. Epub ahead of print. PMID: 35792047.
18. Tarakji B. Dentists' Perception of Oral Potentially Malignant Disorders. Int Dent J. 2022 Jun;72(3):414-419. doi: 10.1016/j.identj.2022.01.004.
19. Nelson S, Heft-Allen M, Narayana N. Accuracy of leukoplakia diagnoses: a retrospective study. Gen Dent. 2022 Mar-Apr;70(2):14-17.
20. van der Waal I. Oral leukoplakia, the ongoing discussion on definition and terminology. Med Oral Patol Oral Cir Bucal. 2015 Nov 1;20(6):e685-92. doi: 10.4317/medoral.21007.
21. Stojanov IJ, Woo SB. AAOM clinical practice statement subject: leukoplakia. Oral Surg Oral Med Oral Pathol Oral Radiol. (2018) 126:331–4. doi: 10.1016/j.oooo.2018.06.006
22. Saldivia-Siracusa C, González-Arriagada WA. Difficulties in the Prognostic Study of Oral Leukoplakia: Standardisation Proposal of Follow-Up Parameters. Front Oral Health. 2021 Feb 5;2:614045. doi: 10.3389/froh.2021.614045. PMID: 35047990; PMCID: PMC8757698.
23. Dantas Da Silveira ÉJ, Lopes MFF, Madeira Silva LM, Fachetti Ribeiro B, Costa Lima K,Guedez Queiroz LM, et al. Potentially malignant oral lesions: clinical and

morphological analysis of 205 cases. J Bras Patol eMed Lab. (2009) 45:233–8. doi: 10.1590/S1676-24442009000300008

24. Warnakulasuriya S. Clinical features and presentation of oral potentially malignant disorders. Oral Surg Oral Med Oral Pathol Oral Radiol. 2018 Jun;125(6):582-590. doi: 10.1016/j.oooo.2018.03.011.

2.4 Artigo: Machine Learning for the prediction of toxicities from head and neck cancer treatment: a systematic review with meta-analysis

Anna Luíza Damaceno Araújo[a], Matheus Cardoso Moraes[b], Maria Eduarda Pérez-de-Oliveira[a] Viviane Mariano da Silva[b], Cristina Saldivia Siracusa[a], Caique Mariano Pedroso[a], Marcio Ajudarte Lopes[a], Pablo Agustin Vargas[a], Sara Kochanny[d], Alexander T. Pearson[c,d], Syed Ali Khurram[e], Luiz Paulo Kowalski[f,g], Mark S Chambers[h], Alan Roger Santos-Silva[a]

[a] Oral Diagnosis Department, Piracicaba Dental School, University of Campinas (UNICAMP), Piracicaba, São Paulo, Brazil.
[b] Institute of Science and Technology, Federal University of São Paulo (ICT-Unifesp), São José dos Campos, São Paulo, Brazil.
[c] Section of Hemathology/Oncology, Department of Medicine, University of Chicago, Chicago, Illinois, United States of America.
[d] University of Chicago Comprehensive Cancer Center, Chicago, IL, USA.
[e] Unit of Oral and Maxillofacial Pathology, School of Clinical Dentistry, University of Sheffield, S10 2TA, Sheffield, United Kingdom.
[f] Department of Head and Neck Surgery and Otorhinolaryngology, A.C. Camargo Cancer Center, São Paulo, Brazil.
[g] Head and Neck Surgery Department and LIM 28, University of São Paulo Medical School, São Paulo, Brazil.
[h] Department of Head and Neck Surgery, Division of Surgery, M.D Anderson Center, The University of Texas, Houston, Texas, United States of America.

Corresponding Author: Alan Roger Santos-Silva
Oral Diagnosis Department, Piracicaba Dental School, UNICAMP
Adress: Av. Limeira, no 901, Areião, Piracicaba, São Paulo, Brazil.
Postal code: 13414- 903
Phone number: +55 19 21065320
E-mail: alan@unicamp.br

Competing Interests statement: None to declare.

Abstract

Introduction: The aim of the present systematic review (SR) is to summarize Machine Learning (ML) models currently used to predict head and neck cancer (HNC) treatment-related toxicities, and to understand the impact of image biomarkers (IBMs) in prediction models (PMs). The present SR was conducted following the guidelines of the PRISMA 2022 and registered in PROSPERO database (CRD42020219304). **Methods:** The acronym PICOS was used to develop the focused review question (PMs can accurately predict HNC treatment toxicities?) and the eligibility criteria. The inclusion criteria enrolled Prediction Model Studies (PMSs) with patient cohorts that were treated for HNC and developed toxicities. Electronic database search encompassed PubMed, EMBASE, Scopus, Cochrane Library, Web of Science, and LILACS. Risk of Bias (RoB) was assessed through PROBAST and the results were synthesized based on the data format (with and without IBMs) to allow comparison. **Results:** A total of 28 studies and 4,713 patients were included. Xerostomia was the most frequently investigated toxicity (17; 60.71%). Sixteen (57.14%) studies reported using radiomics features in combination with clinical or dosimetrics/dosiomics for modelling. High RoB was identified in 23 studies.Meta-analysis (MA) showed an area under the receiver operating characteristics curve (AUROC) of 0.82 for models with IBMs and 0.81 for models without IBMs (p value <0.001), demonstrating no difference among IBM- and non-IBM-based models. The evidence was appraised as of low certainty. **Discussion:** The development of a PM based on sample-specific features represents patient selection bias and may affect a model's performance. Heterogeneity of the studies as well as non-standardized metrics prevent proper comparison of studies, and the absence of an independent/external test does not allow the evaluation of the model's generalization ability. **Conclusion:** IBM-featured PMs are not superior to PMs based on non-IBM predictors.

Key-words: Machine Learning; Prediction Model Studies; Convolutional Neural Network; Xerostomia; Mucosistis; Dysphagia; Osteoradionecrosis; Hypothyroidism; Hearing Loss;

Introduction

Prediction Models (PM) applied to oncologic patients enhance the identification of treatment endpoints (i.e., structural changes, toxicities, organs-at-risk dose, complications and treatment failure), as well as oncologic outcomes, and pathologic findings [1]. The use of PM to identify head and neck cancer (HNC) treatment toxicities can aid in the individualization of treatment plans and requires the implementation of multi-variable models able to process data from different sources, as charting data, health records, dosiomics/dosimetric parameters, genomics, pathological and a wide range of quantitative image biomarkers (IBMs) extracted from imaging data as computed tomography (CT) and magnetic resonance image (MRI) defined as radiomics features.

Conventional Normal Tissue Complication Probability (NTCP) models for treatment toxicity prediction are based on a mathematical function in which multiple input variables such as clinical, demographic and dosimetric parameters [2], as well as baseline complaints (e.g., patient-rated xerostomia and sticky saliva) are used to predict an outcome. These models evolved from the Lyman–Kutcher–Burman model [3] and are known to have limited learning capacity, since they require the user's feedback to establish the correct correlation of the prediction made with the inputs selected to compose the model, making it an almost empirical process. Additionally, there is still an unexplained variance in predicting such outcomes with NTCP models [4], which could be improved by dose-independent radiomic-based approaches, allowing use of models prior to treatment planning for fast identification of susceptible patients [5].

To overcome these drawbacks, novel ML models for toxicity prediction consider individual patient characteristics and allow association with IBMs automatically extracted from radiomic features which, in theory, perform at least as well as the conventional NTCP models or are likely to outperform them, since training models with image characteristics and personalized data can be more representative and lead to more realistic outputs. However, previously published articles demonstrate that performance improvement of models was minor [6] while some authors state that radiomic data does not show superiority over NTCP models [5,7].

Within this frame of reference, the aim of this systematic review (SR) is to summarize Machine Learning (ML) models currently used to predict HNC treatment-related toxicities, with no restrictions on treatment modalities or data type, and focusing on the performance of models used and their reliability for clinical decision support. The present SR is based on the

following review questions: i) PMs can accurately predict HNC treatment toxicities? Additionally, we aim to identify the toxicities currently under investigation, to identify the type of PMs used to predict the endpoints, and to understand the impact of IBMs in the performance of PMs.

Methods

Eligibility Criteria

The acronym PICOS was used to develop the focused review question and the eligibility criteria - which were framed based on the *Guidance for defining review question* (CHARMS Checklist) *and* the *Guide for SR and meta-analysis of PMSs* [8,9] - in which Participants/Populations are HNC patients, Intervention are the cancer treatments (any modality of radiotherapy and/or chemotherapy), the Outcome is the prediction of a given end point (toxicity), and the included Studies are prognostic PMSs that developed the model in a retrospective cohort of patients **(Appendix I)**. Since some studies report IBM-based models only without directly comparing their results with standard non-IBM-based models using the same dataset, the authors decided to include studies that only report non-IBM-based models as well, if they fit all the other criteria. Studies that aimed to develop a framework for data mining, and to predict OARs, disease-free survival/recurrence or response to treatment were not included. Performance metrics are widely variable within studies. To provide a better assessment of the results and allow comparison of performances, only studies reporting at least one of the following metrics were included: area under the Receive Operation Characteristics (ROC) curve (AUROC), accuracy, sensitivity, specificity, precision, and F1-score. To be included in the meta-analysis (MA) of the AUROC curve, studies should provide AUROC, confidence interval, and standard error values (or allow calculation). The full eligibility criteria and article selection flowchart are shown in **figure 1.**

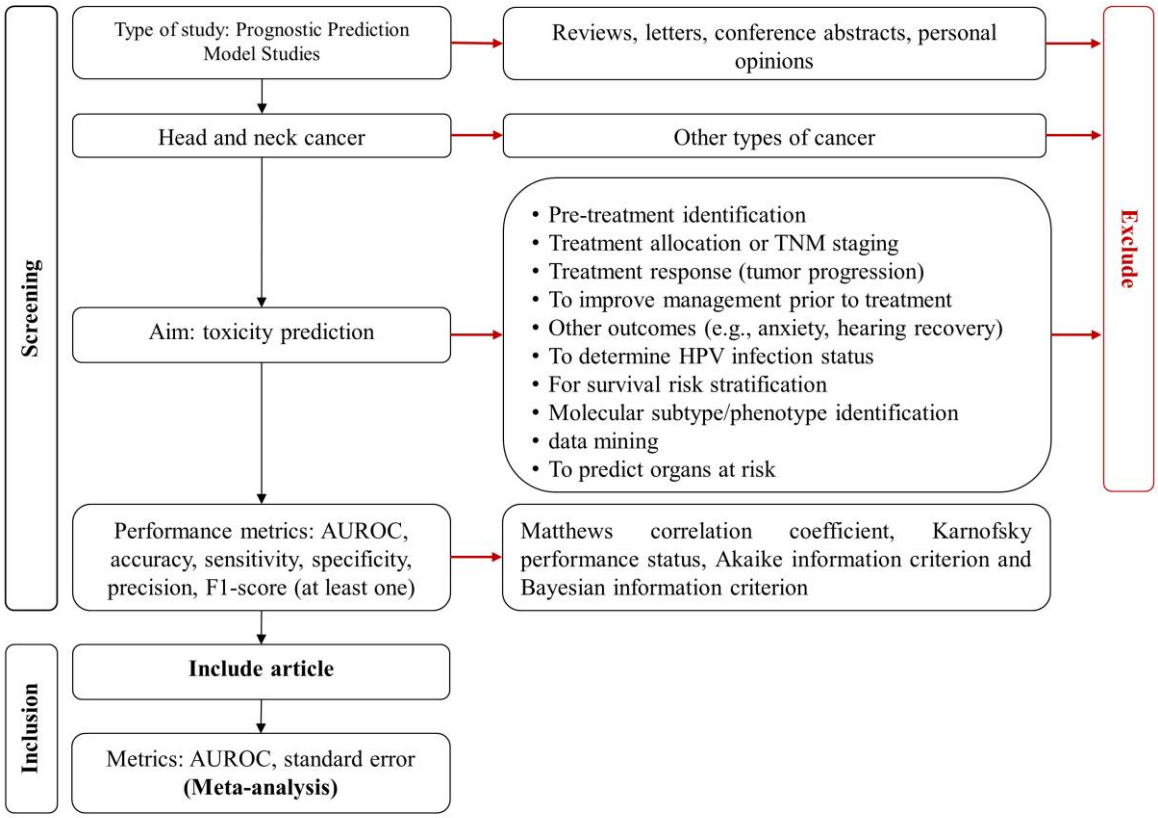


Figure 1. Eligibility criteria and article selection flow

Information sources and search strategy

Individualized search strategies were carried out on June 18th, 2022 for the following electronic databases: PubMed, EMBASE, Scopus, Cochrane Library, Web of Science, and LILACS. Additionally, the gray literature was retrieved on Google Scholar, ProQuest, and Open Grey. Furthermore, the reference lists of included studies were screened to identify any additional relevant reference that could be missed in search strategy. The complete search strategy is shown in **Appendix I.**

Selection process

The study selection was made in two phases and by two independent reviewers (ALDA and MCM). The first phase was performed by reading titles and abstracts on Rayyan [10], excluding papers non-related to the SR subject. The second phase was proceeded with full-text reading of articles, applying the same eligibility criteria previously established. Divergences in any phase were resolved by mutual agreement among authors.

Data Collection process and data items

The choice of data that should be extracted from each included article was made by discussion among authors. The data was collected by one reviewer (ALDA) and cross-checked by a second reviewer (MCM). The variables encompassed author, year, cancer type, treatment modality, toxicity, data modality, model, the most discriminative features (predictors), sampling methods, performance metrics, and conclusion.

Risk of Bias (RoB) *assessment*

RoB of each study was assessed independently by two authors (ALDA and MCM) was assessed through PROBAST [8,11], a tool designed for PMSs. Additionally, an analysis of the included studies based on the TRIPOD [12,13], a guideline to develop PMSs, is provided, enlightening the methodological flaws of primary studies. The PROBAST-AI and TRIPOD-AI [14], extensions tools specifically designed for ML-based PMSs, are currently being developed and further studies should apply them.

Effect measures

This SR aims to summarize the performance of ML-based PMs used to predict HNC treatment-related toxicities, with no restrictions on treatment modalities or data type. The effect measures considered were AUROC and confidence interval, since this metric allows to evaluate class separation and generalization of the PM.

Synthesis Methods

The learning modalities were classified as supervised and unsupervised, and the predictive models as classical/traditional and modern (i.e., representative learning models). When possible, missing metrics were calculated based on reported metrics.

Given the methodologies' heterogeneity, common in PMSs, data was organized in clusters for analysis. First, a descriptive analysis based on the given endpoint was conducted. Accordingly, since the goal of the present SR is to retrieve information of PM about toxicity, with and without IBMs, a comparative analysis was oriented by data format (i.e. clinical, dosimetrics, dosiomics and radiomics). Within this setting, when a study brings results comparing different associations of predictors for the same model, only the model with the best predictors combination was included, and when a study compares several AI models with

and without IBMs, the best performance for each model and each modality of data was included to allow comparison.

For the MA, an analysis of the (AUROC) was conducted in MedCalc by entering area under the curve (AUC) and standard error. For a better assessment, two MA were conducted among studies that presented external validation: one for IBM-based prognostic models (2 studies, 10 models) [5,15] and the other for non-IBM-based (2 studies, 3 models) [5,16]. Statistical heterogeneity was calculated using an inconsistency (I^2>50% = significant heterogeneity). Since a high heterogeneity was present, the random effect model was chosen to evaluate AUROC curves and compare the IBM-based models and the non-IBM-based models.

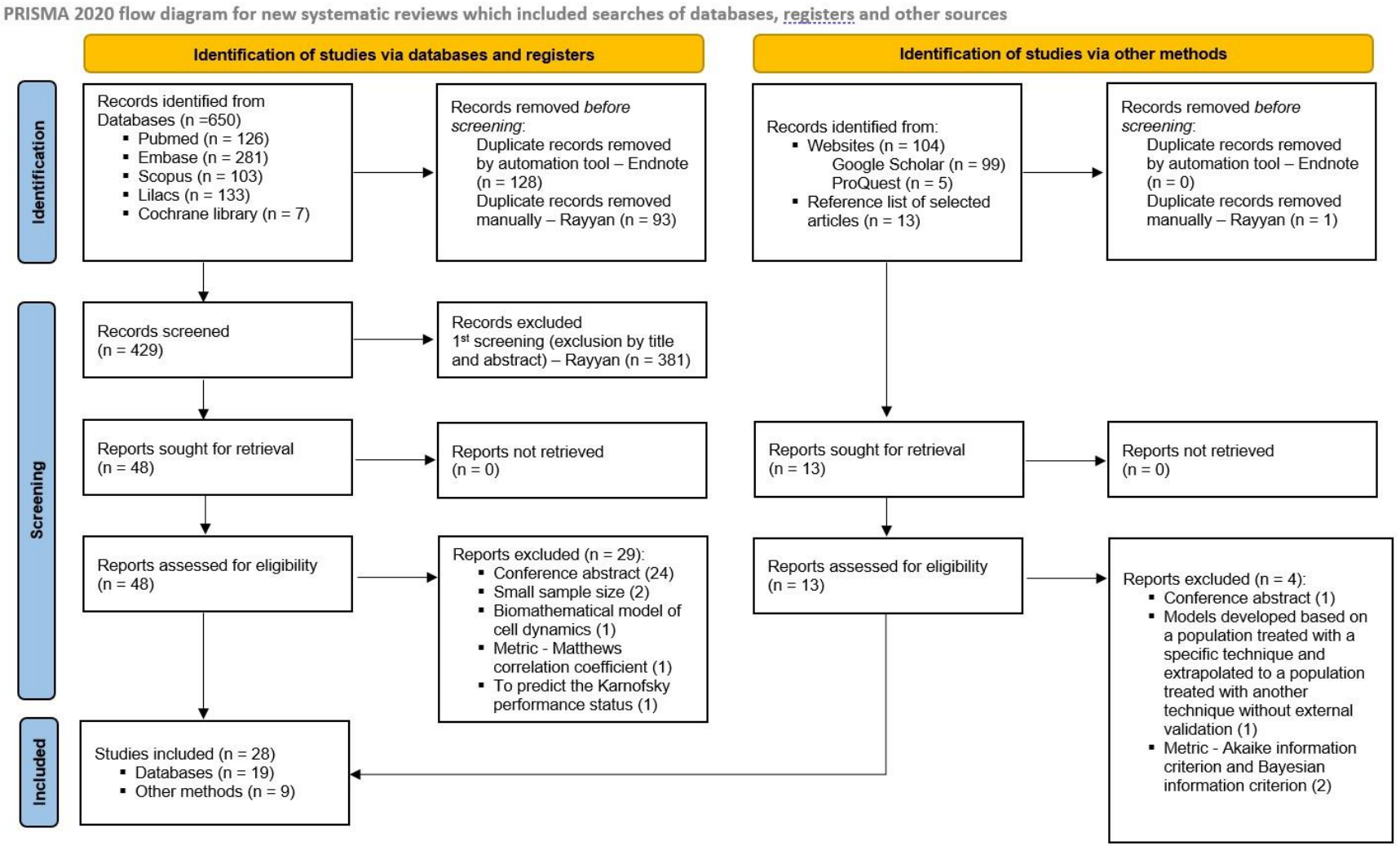
Certainty Assessment

The certainty of the evidence (high, moderate, low, or very low) was appraised using the Grading of Recommendation, Assessment, Development, and Evaluation (GRADE) instrument [17] as reference and the GRADEpro [18], which is based on study design, risk of bias, inconsistency, indirectness, imprecision, and other considerations, including publication bias and effect magnitude. The assessment was scored as high, moderate, low, or very low.

Results

Study selection

Amongst a total of 818 records identified through the search strategy, 28 articles [2,4–7,15,16,19–39] fulfilled eligibility criteria and were included in the qualitative synthesis. The study selection process is summarized in the PRISMA Flowchart (**figure 2**) and the reasons for exclusion of each article read in full text in the second phase are described in **Appendix II.**



From: 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:n71. doi: 10.1136/bmj.n71. For more information, visit: http://www.prisma-statement.org/

Figure 2. PRISMA Flowchart

Study characteristics

A total of 4,713 patients treated for HNC comprised the cohorts used for developing and validating all models. The tumor sites were cochlea, oral cavity, nasal cavity, paranasal sinus, nasopharynx, oropharynx, hypopharynx, larynx, parotid, and unknown primary site. All included studies reported radiotherapy (RT) modalities as treatments (e.g., IMRT, 3DCRT, ST-IMRT, SW-IMRT, VMAT, TomoTherapy) in combination or not with chemotherapy, cisplatin, or cetuximab.

Xerostomia was the end point most frequently investigated (17; 60.71%) [4,6,7,16,21,22,26,28,30–35,37–39]. Five studies (17.85%) enhanced the use of parotid-sparing RT in the cohorts [4,6,37–39]. Xerostomia PM were based on patient-rated xerostomia in 6 (35.3%) studies [4,6,21,37–39], physician-rated in 1 (6%) [28], while the remaining studies were based on different xerostomia criteria. Among these 17 studies, 3 (10.7%) also investigated sticky saliva prediction [6,7,21]. Dysphagia was the second most commonly investigated endpoint, with 3 (10.7%) studies [2,15,36], followed by weight loss/need for feeding tube (2; 7.1%) [23,25] mucositis (1; 3.5%) [24], saliva amount (1; 3.5%) [29], osteoradionecrosis (1; 3.5%) [27], sensorineural hearing loss (1; 3.5%) [19], and radiation-induced hypothyroidism (1; 3.5%) [5], and radiation-related caries (1; 3.5%) [20].

The majority of included studies (27; 96.4%) applied supervised learning modalities, with Logistic Regression (LR), Multivariable LR, and Penalized LR being the most frequent classifier (19; 67.8%) [2,4–7,15,16,21,23–25,27–30,34–39] in comparison or not with other traditional classifiers or modern models, while only 6 (21.4%) studies reported RL models (i.e., Multilayer Perceptron and Convolutional Neural Networks) [5,16,20,25,27,30] and the remaining had reported classical methods for toxicity prediction. Among studies using RL models, 1 (1; 3.5%) study also developed an ensemble model which associated a logistic regression model and a CNN for image processing [25]. One study (1; 3.5%) applied an unsupervised data mining approach (clustering) [2]. More details on feature (predictor) selection and classification algorithm are shown in **table 1**.

Table 1. Included studies (to be continued)

Author/ year	Toxicity	Feature selection	Classification model/ algorithm	IBM	Data	PMSt	Train/ valid set	Test set	AUC	CI	SE	Study conclusion and most important predictors
Abdollahi et al, 2018 [19]	Sensorineural hearing loss	CfsSubsetEval	Decision Stump	Y	D/C/R	1b	47	NA	NI	NI	NI	Prediction improvement when using gEUD effect. Needs validation in larger sample.
			Hoeffding	Y	D/C/R	1b			NI	NI	NI	
		LASSO penalized LR	all ML models without gEUD effect	Y	D/C/R	1b			NI	NI	NI	
			all ML models with gEUD effect	Y	D/C/R	1b			0.79	NI	NI	
			Elastic-Net Regularized GLM ("Glmnet")	Y	D/C/R	1b			0.88	NI	NI	
Araújo Faria et al, 2021 [20]	Radiation-related caries	KBestSelect	ANN	Y	R	2a	10*	5*	0.98	NI	NI	1. High accuracy to select the best radiomic features. 2. The selection approach for the best features is crucial. 3. Needs validation in larger sample.
Beetz et al, 2012a [21]	Xerostomia at 12 months after radiotherapy	Pearson correlation, FS and extended bootstraping	M-LR	N	D/C	1a	178	NA	0.68	(0.60-0.76)	NI	Indicate which OARs should be spared to optimise current treatment with IMRT. The two-factor model containing base line xerostomia and the mean dose to the contralateral gland performed significantly better in that respect.
	Sticky saliva until the end of treatment		M-LR	N	D/C	1a			0.68	(0.60-0.76)	NI	
Beetz et al, 2012b [7]	Xerostomia at 12 months after radiotherapy	Pearson correlation, FS and extended bootstraping	M-LR	N	D/C	1a	165	NA	0.82	(0.76-0.89)	NI	Dose distributions in the minor salivary glands have limited significance for the develop ment of patient-rated symptoms related to salivary dysfunction among patients treated with 3D-CRT. Patient-rated xerostomia and sticky saliva cannot be predicted with one simple relationship between the dose distribution in an OAR and an endpoint.
	Sticky saliva until the end of treatment		M-LR	N	D/C	1a	167		0.84	(0.78-0.90	NI	

Table 1. Included studies (continuation)

Author/ year	Toxicity	Feature selection	Classification model/ algorithm	IBM	Data	PMSt	Train/ valid set	Test set	AUC	CI	SE	Study conclusion and most important predictors
Buettner et al., 2012 [22]	Xerostomia at 12 months after radiotherapy	Bayesian model-selection algorithm	Bayesian M-LR (mean dose model)	N	D/C	3	63	48	0.73	NI	NI	Dose–response models based on morphological descriptors of the dose distribution are more accurate than standard mean-dose models. When generating IMRT treatment plans, spatial information should be taken into account and could result in lower complication rates.
			Bayesian M-LR (morphological model 6)	N	D/C	3			0.8	NI	NI	
Cheng et al., 2019 [23]	Weight loss	LASSO	LR	Y	D/C/R	2a	63	100	NI	NI	NI	The predictions may be optimized by our inclusion of radiomics features into the model. Features included volume, shape, first-order statistics for the distribution of intensities, and texture20 for the parotid and submandibular glands, larynx, and superior constrictor muscles.
Dean et al, 2017 [24]	Mucositis grade (2 or lower = severe)	LASSO	Penalised LR	Y	D/C/R	2a	179	NA	0.71	NI	0.10	Using a novel mucosal surface contour organ at risk did not improve the predictive performance of severe acute mucositis NTCP models.
			RF	Y	D/C/R	2a			0.69	NI	0.09	
Dean et al, 2018 [15]	Dysphagia [severe (grade 3 or worse) and non-severe (less than grade 3)]	LASSO	Penalized LR (standard)	Y	D/C/R	3	173	90	0.82	NI	0.004*	1. The authors recommend the model in xx for clinical decision-support, due its superior performance when compared to the model developed by the authors in terms of the probability calibration. 2. Doses of approximately 1 Gy/fraction were most strongly associated with severe dysphagia.

Table 1. Included studies (continuation)

Author/ year	Toxicity	Feature selection	Classification model/ algorithm	IBM	Data	PMSt	Train/ valid set	Test set	AUC	CI	SE	Study conclusion and most important predictors
Dohopolski et al., 2022 [25]	Need for a feeding tube (G-tube or NG-tube) or 10% or more weight loss during RT	RFE	LR	N	C	2*	162*	55	0.69	(0.63-0.76)	NI	1. An ensemble model combining clinical parameters and 3D imaging was statistically superior to the clinical model alone, highlighting how DL may augment clinical care.2. External test is required to incite more immediate clinical changes.
			SVM	N	C	1b			0.68	(0.62-0.75)	NI	
			MLP	N	C	1b			0.66	(0.60-0.73	NI	
		NA	ResNet-50	Y	R	1b			0.63	(0.56-0.70)	NI	
			MedicalNet (transfer learning with a ResNet-based model)	Y	R	2*			0.73	(0.67-0.79)	NI	
			Ensemble clinical LR model and MedicalNet imaging model	Y	C/R	2*			0.75	(0.69-0.81)	NI	
Gabry´s et al, 2018 [26]	Xerostomia (early)	UFS-F	k-nearest neighbors	Y	D/R	1b	153	NA	0.65	(0.62–0.68)	NI	1. Due to strong dependence on patient-specific factors, there is a need for personalized data-driven risk profiles in future development of NTCP. 2. Feature selection allowed for a reduction of model complexity. 3. In small clinical datasets, simple LR can perform as well as top-ranking ML algorithms.
	Xerostomia (late)	MB-LR	Gradient tree boosting	Y	D/R	1b			0.65	(0.59–0.70)	NI	
	Xerostomia (long term)	MB-LR	Extra trees	Y	D/R	1b			0.88	(0.84–0.91)	NI	
	Xerostomia (longitudinal)	RFE-LR	Gradient tree boosting	Y	D/R	1b			0.63	(0.52–0.71)	NI	
Humbert-Vidan et al, 2021 [27]	Osteoradionecrosis	Univariate analysis, $\chi 2$ and Mann–Whitney U Non-parametric test.	M-LR	N	D/C	2*	96	20	NI	NI	NI	Successfully prediction of ORN using ML methods.
			SVM	N	D/C	2*			NI	NI	NI	
			RF	N	D/C	2*			NI	NI	NI	
			AdaBoost	N	D/C	2*			NI	NI	NI	
			ANN	N	D/C	2*			NI	NI	NI	
Jiang et al, 2019 [28]	Xerostomia at 3 months of follow up	U-LR	Ridge LR	N	D/C	2a	427	NA	0.69	NI	0.08	1. The AUC performance using dosimetric features is not significantly different from voxel-dose features. 2. The specific subvolume in the PG and SMG of the contralateral PG are the most predictive features.
		U-LR	LASSO LR	N	D/C	2a			0.67	NI	0.06	
		U-LR	RF	N	D/C	2a			0.69	NI	0.07	

Table 1. Included studies (continuation)

Author/year	Toxicity	Feature selection	Classification model/algorithm	IBM	Data	PMSt	Train/ Valid Set	Test set	AUC	CI	SE	Study conclusion and most important predictors
Liu et al, 2019 [29]	Saliva amount	RFE	Linear regression	Y	C/R	3	35	4	NI	NI	NI	1. The proposed method was able to accurately predict patients' saliva amount at early stage and prevent the xerostomia symptom in advance. 2. 10 of 14 selected features were radiomics, which can effectively represent features' changes during RT.
Men et al, 2019 [30]	Xerostomia	NA	3D $_r$CNN	Y	D/R	2a	706*	78*	0.84	(0.74-0.91)	NI	1. The present automatic extraction of both radiomics and dosiomics features is a great advantage over the traditional LR model. 2. The inclusion of CT image could improve the performance. 3. Even without the input contour, CNN could achieve slightly inferior/ comparable performance. 4. More studies are warranted.
			3D $_r$CNN	Y	D/R (no contour)	2a			0.82	(0.72-0.90)	NI	
			3D $_r$CNN	Y	D/R (no CT)	2a			0.78	(0.67-0.88)	NI	
			3D $_r$CNN	Y	R	2a			0.7	(0.58-0.80)	NI	
		FS	LR	N	D	2a			0.68	(0.56-0.80)	NI	
			LR	N	D/C	2a			0.74	(0.64-0.84)	NI	
Nakatsugawa et al, 2019 [31]	grade >2 xerostomia at 3 to 6 months of follow up	FS	Bivariate LR	N	D/C	NA	297	NA	0.6164	NI	NI	Updating prediction models with prospective data collection is effective for maintaining the performance of our xerostomia prediction.
Nardone et al, 2018 [32]	Xerostomia at 12 months after radiotherapy	Correlation and FS	Bivariate LR	N	D/C	2a	78	NI	0.77	(0.65-0.99)	NI	1. Texture analysis seems to improve the knowledge of the predictive factors of this kind of radiation therapy's toxicity. 2. Textural features (RLNU and GLCM) could be associated with radiosensitivity of the PG (lower number of acinar cells, vascularization or greater ratio of adipose tissue. 3. Further studies on a large population are needed to better estimate the actual preliminary data.
				Y	D/C/R	2a			0.91	(0.75-0.98)	NI	

Table 1. Included studies (continuation)

Author/ year	Toxicity	Feature selection	Classification model/ algorithm	IBM	Data	PMSt	Train/ valid set	Test set	AUC	CI	SE	Study conclusion and most important predictors
Pota et al, 2017 [33]	Xerostomia at 12 months after radiotherapy	FS	Likelihood-Fuzzy Analysis	N	C	1b	19	NA	NI	NI	NI	1. Models reached high accuracy thanks to the employment of radiomics-based features. 2. This work can address future studies in considering radiomics features besides parotid shrinkage in the construction of normal tissue complication probability (NTCP) models. 3. The performance of both single variable and multiple variables models, obtained by different methods are very similar. 4. Different methods identify the same best predictors for both endpoints.
				N	D	1b			NI	NI	NI	
				Y	R	1b			NI	NI	NI	
				Y	R	1b			NI	NI	NI	
				Y	R	1b			NI	NI	NI	
			Naïve Bayes	N	C	1b			NI	NI	NI	
				N	D	1b			NI	NI	NI	
				Y	R	1b			NI	NI	NI	
				Y	R	1b			NI	NI	NI	
				Y	R	1b			NI	NI	NI	
Rosen et al, 2018 [34]	Xerostomia (grade ≥1)	LASSO	Penalized LR	N	D/C	1b	105	NA	0.71	(0.60-0.81)	NI	Early treatment CBCT-measured PG density changes were shown to be associated with long-term xerostomia and improved predictions over PG dose alone.
				Y	D/C/R	1b			0.72	(0.60-0.83)	NI	
	Xerostomia (grade ≥2)		Penalized LR	N	D/C	1b			0.69	(0.62-0.77)	NI	
				Y	D/C/R	1b			0.78	(0.64-0.91)	NI	
Sheikh et al, 2019 [35]	Xerostomia	Spearman correlation, LASSO, and internal LOO-cross-validation	GLM	N	D	3	216	50	0.63	(0.51–0.81)	NI	1. Image features from salivary glands significantly contributed to xerostomia prediction. 2. Higher order texture features for both ipsi- and contralateral salivary glands were important predictors. 3. combining multimodal image features with dosimetry features improved xerostomia prediction. 4. The model's performance improved with DVH+CT+MR features compared to DVH or CT+MR.
				Y	R	3			0.57	(0.45–0.71)	NI	
				Y	R	3			0.66	(0.54–0.82)	NI	
				Y	R	3			0.7	(0.57–0.82)	NI	
				Y	D/R	3			0.7	(0.57–0.82)	NI	
				Y	D/R	3			0.56	(0.40–0.68)	NI	
				Y	D/R	3			0.6	(0.50–0.73)	NI	
				Y	C/R	3			0.73	(0.62-0.86)	NI	
				Y	D/C/R	3			0.68	(0.52–0.80)	NI	

Table 1. Included studies (continuation)

Author/ year	Toxicity	Feature selection	Classification model/ algorithm	IBM	Data	PMSt	Train/ valid set	Test set	AUC	CI	SE	Study conclusion and most important predictors
Smyczynska et al., 2021 [5]	Radiation-induced hypothyroidism	t-tests, hierarchical clustering and FS	GP (variant Ia)	N	D	3	98	60	0.9	NI	0.07	Our models tend to be slightly less sensitive, but more specific and accurate. Radiomic-based models are dose-independent (can be used prior to treatment planning allowing faster selection of susceptible population) but did not outperform state-of-art NTCP models. Radiomic features came from original, logarithm, exponential and wavelet (LLL, HHH) images (measured nonuniformity of the thyroid region as coarseness, zone percentage) and were higher in patients who developed the toxicity. The only feature lower in these patients was least axis length of thyroid.
			GP (variant Ib)	N	D	3			0.9	NI	0.07	
			GP (variant II)	N	D	2a			0.95	NI	0.05	
			LR (variant Ia)	Y	R	3			0.89	NI	0.07	
			MLP_4 (variant Ib)	Y	R	3			0.94	NI	0.05	
			MLP_4 (variant II)	Y	R	2a			0.91	NI	0.07	
			MLP_2 (variant Ia)	Y	C/R	3			0.95	NI	0.05	
			MLP_4 (variant Ib)	Y	C/R	3			0.94	NI	0.05	
			MLP_2 (variant II)	Y	C/R	2a			0.92	NI	0.06	
Soares et al, 2018 [16]	Xerostomia at 12 months	Clinical knowledge***	RF	N	D/C	3	114	24	0.69	NI	0.018371	RF presented high performance and good discriminative ability. The role of age, gender, severity of xerostomia prior to radiation therapy and planned mean physical dose in the contralateral and ipsilateral parotids appeared to be of main importance.
			Stochastic Boosting	N	D/C	2a			0.65	NI	NI	
			SVM	N	D/C	2a			0.66	NI	NI	
			LR	N	D/C	2a			0.47	NI	NI	
			Clustering	N	D/C	2a			0.43	NI	NI	
		NA	ANN	N	D/C	2a			0.61	NI	NI	
Ursino et al, 2021 [36]	Dysphagia (disturbed swallowing (penetration/ aspiration) at 12 months)	a pipeline specifically designed using CERR	Linear SVM	Y	D	1b	38	NA	0.85	NI	NI	Swallowing organs at risk have been poorly considered until recently and are worth further investigating in clinical research.
			LR	Y	D	1b			0.82	NI	NI	
			RF	Y	D	1b			0.94	NI	NI	

Table 1. Included studies (continuation)

Author/ year	Toxicity	Feature selection	Classification model/ algorithm	IBM	Data	PMSt	Train/ valid set	Test set	AUC	CI	SE	Study conclusion and most important predictors
van Dijk et al., 2017a [6]	Xerostomia	Pearson correlation and LASSO	M-LR	N	D/C	1b	249	NA	0.75	(0.69–0.81)	NI	Prediction significantly improved by including CT biomarker ‘‘Short Run Emphasis” (might be a measure of non-functional fatty parotid tissue). The maximum CT intensity was associated with sticky saliva (probably related with vascularization). These IBM are a first step to identifying patient characteristics that explain the patient-specific response of healthy tissue to dose.
	Xerostomia		M-LR	Y	D/C/R	1b			0.77	(0.71–0.82)	NI	
	Sticky saliva		M-LR	N	D/C	1b			0.74	(0.67–0.80)	NI	
	Sticky saliva		M-LR	Y	D/C/R	1b			0.77	(0.71–0.83)	NI	
van Dijk et al., 2017b [37]	Xerostomia at 12 months	NA	M-LR (Ref. model[re])	N	D/C	4	107	107	0.76	(0.67–0.86)		Mean PG dose significantly correlated with ΔPG-surface and did not add information to the ΔPG-surface model in predicting late xerostomia in this cohort. The model with DPG surface and acute xerostomia early after radiation therapy significantly improved model performance to predict late xerostomia.
		Pearson correlation, stepwise FS repeated in 1000 bootstrapped samples	LR	Y	R	1b			0.76	(0.66–0.85)		
			LR	Y	C/R	1b			0.82	(0.72-0.91)		
			LR	Y	D/C/R	1b			0.82	(0.73-0.91)		
			LR	N	C	1b			0.85	(0.77-0.93)		
			LR	Y	C/R	1b			0.9	(0.84-0.96)		
van Dijk et al., 2018a [4]	Xerostomia	NA	M-LR (Ref. model[re])	N	D/C	4	161	161	0.73	(0.65-0.81)		The addition of the predictive intensity PET-IBM (90th percentile of SUV) to a model with parotid gland dose and baseline xerostomia improved the prediction performance. Resulting from both the Lasso regularisation and forward selection, the 90th percentile of SUVs (P90) was the most predictive of all intensity PET-IBMs. The most predictive textural PET-IBM was the Long Run High Grey-level Emphasis 3 (LRHG3E).The SRE neither significantly improved the reference model nor did it add to the PET-IBM models with P90 and LRHG3E in this cohort subset.
		LASSO	M-LR	Y	D/C/R	1b			0.77	(0.69–0.84)		
			M-LR	Y	D/C/R	1b			0.77	(0.70–0.84)		
			M-LR	Y	D/C/R	1b			0.77	(0.69–0.84)		

Table 1. Included studies (continuation)

Author/ year	Toxicity	Feature selection	Classification model/ algorithm	IBM	Data	PMSt	Train/ valid set	Test set	AUC	CI	SE	Study conclusion and most important predictors
van Dijk et al., 2018b [38]	Xerostomia at 12 months after radiotherapy	NA	M-LR (Ref. model[re])	N	D/C	4	43	25	0.65	(0.41-0.88)	NI	The prediction performance of xerostomia based on parotid dose and baseline xerostomia only was improved by the addition of the predictive intensity MR-IBM P90 (high fat concentration is related to a higher risk of developing xerostomia). More research is needed.
		Step-wise FS	M-LR	Y	D/C/R	3			0.83	(0.66-0.99)	NI	
			M-LR	Y	D/C/R	3			0.83	(0.67-0.99)	NI	
van Dijk et al., 2019 [39]	Xerostomia at 12 months after radiotherapy	Not performed FS	M-LR	N	D/C	2*	56	14	0.80	NI	NI	Mid-treatment parotid gland changes substantially improve the prediction of late radiation induced xerostomia.
			M-LR	Y	C/R	2*			0.85	NI	NI	
			M-LR	Y	D/C/R	2*			0.93	NI	NI	
Wentzel et al., 2020 [2]	Dysphagia	hierarchical agglomerative clustering (data mining)	LR	Y	C/R	NA	200	NA	0.84	NI	NI	The proposed methodology of automatically generating a simple stratified risk score for dysphagia could be applied to identifying high-risk groups of other negative patient outcomes and better guide future treatment recommendations.
				Y	R	NA			0.68	NI	NI	
				N	C	NA			0.7	NI	NI	
				Y	C/R	NA			0.82	NI	NI	
				N	C	NA			0.79	NI	NI	
				Y	R	NA			0.77	NI	NI	
				Y	C/R	NA			0.8	NI	NI	The combination of T-stage and spatial clusters notably improves performance. Spatial clusters and clinical features combined reached the best performance.
	Feeding tube			Y	C/R	NA			0.71	NI	NI	
				Y	R	NA			0.64	NI	NI	
				N	C	NA			0.6	NI	NI	
				Y	C/R	NA			0.76	NI	NI	
				N	C	NA			0.64	NI	NI	
				Y	R	NA			0.72	NI	NI	
				Y	C/R	NA			0.67	NI	NI	

* calculated by the authors

** 2 (did not mention if the split was random);

***In this article, no statistics were used to select the predictors (the study has less bias than if univariate analysis was done);

ANN: Artificial Neural Network;

AUC: area under the curve;

C: clinical;

CI: Confidence interval;

CNN: Convolutional Neural Network;

D: dosimetrics;

DVH: dose–volume histogram;

FS: forward selection approach/forward method of Sequential Feature Selector;

gEUD: generalized equivalent uniform dose

GLM: Generalized linear models (multiple LR);

GP: Gaussian Process;

IBM: image biomarkers;

LASSO: least absolute shrinkage and selection operator;

LOO: leave-one-out;

LR: logistic regression;

MB-LR: model based feature selection by logistic regression;

$MLP_{2:}$ Multilayer Perceptron with 2 neurons in single hidden layer;

$MLP_{4:}$ Multilayer Perceptron with 4 neurons in single hidden layer;

M-LR: multivariate logistic regression;

NA: not applicable;

NI: not informed;

NN: Neural Network;

OARs: organs at risk;

PG: parotid gland;

PMSt: Prediction Model Studies type;

R: radiomics;

$_{r}$CNN: residual Convolutional Neural Network;

Ref. model^{re} : reference model retested (model proposed by Beetz 2012; Howeling, 2010);

RFE: recursive feature elimination;

RFE-LR: recursive feature elimination by logistic regression;

SE: standard error;

SVM: Support Vector Machine;

UFS-F: univariate feature selection by f-score;

U-LR: univariate logistic regression;

Among the 28 included studies, 11 (%) reported NTCP models [4–7,15,21,22,24,26,37,38], 1 (3.5%) reported models with only dosimetrics [36] and 1 (3.5%) reported a model using only radiomics [20], and 4 (14.3%) reported multivariable models trained with dosimetrics, clinical and radiomics features combined [15,19,23,24]. The remaining included articles reported models trained with different combinations of data (e.g., dose and clinical, clinical and radiomics, dose and radiomics, etc). In total, 16 (57.14%) studies reported using radiomics features in combination with clinical or dosimetrics/dosiomics for modelling, in which 12 (42.85%) compare different models with and without IBMs using the same dataset (per study) [2,4–6,25,30,32,34,35,37–39] **(table 1)**. For the purposes of the present systematic review, data modalities frequencies and performances of models using IBMs are displayed in comparison with those not using IBMs for a comparative visual perspective in **figure 3**, as well as performances according to the Learning Modalities.

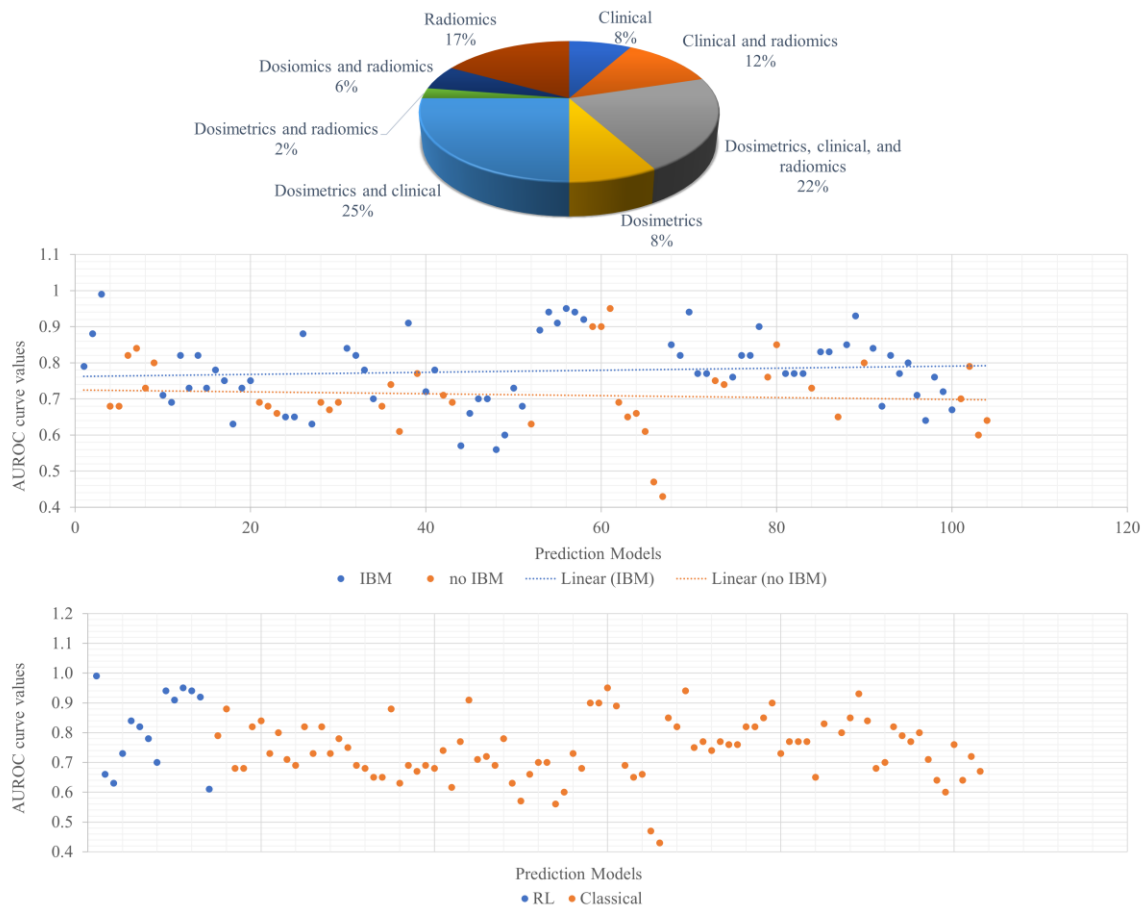


Figure 3. Data modalities frequencies and visual comparison of models performances

RoB in studies

To provide an idea of how consistent methodologies are and to enlighten methodological flaws in primary studies, an evaluation based on the TRIPOD Checklist [12,13] was conducted. Among the included studies, only 3 (10.7%) mentioned relying on TRIPOD to develop their PMs. Overall quantification of items identified that source of data is unclear in 6 (21%) of studies, and in 10 (36%) the number/ location of centers, the eligibility criteria and/or treatments were not mentioned. Only 12 (43%) of the studies reported performances with confidence interval.

An important delimitation the TRIPOD provides is regarding the type of PMSs according to the data sampling. Among the included studies, 2 studies reported 4 models classified as type 1 [7,21] (development only), 8 studies reported 23 models classified as type 2a [5,16,20,23,24,28,30,32] (development and validation with random sampling), seven studies reported 27 models classified as type 3 [5,15,16,22,29,35,38] (development and validation with non-random sampling), and three studies reported 3 models classified as type 4 [4,37,38] (validation only). Three studies reported 11 models in which it was not possible to identify if the split was random or not [25,27,39]. This classification did not apply to 1 study that aimed to evaluate the impact of continuous model update [31] and 1 study that applied unsupervised clustering classification [2], both methodologies not driven by sampling methods. Only models reporting external validation metrics were considered for meta-analysis, encompassing 3 studies and 13 models (10 with IBMs and 3 without IBMs). Only 4 (14%) describe the medical context and specify the studies' objectives. The majority of studies do not define if the study describes the development of the model, validation, or both.

Seven (25%) studies randomly split their dataset to develop and validate the proposed model internally [20,23,25,27,30,32,39]. Among these, two studies [25,30] evenly split their data set into a training, validation, and test set, but these were not considered external validation because data set splitting fit the study as type 2a [30] or did not fit in any type because splitting modality was unclear [25]. For those performing external validation, a total of 1010 patients were included in the development of the model, and 569 patients were included in the external test set [4,5,15,16,22,29,35,37,38].

PROBAST [8,11] was checked for RoB assessment and applicability of primary PMSs.

Table 2. Risk of Bias in included studies

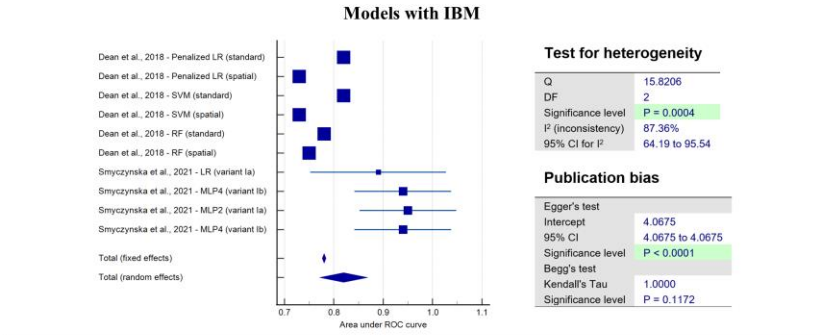
		ROB				Applicability				
	Author/ year [ref]	**Participants**	**Predictors**	**Outcomes**	**Analysis**	**Participants**	**Predictors**	**Outcomes**	**ROB**	**Applicability**
1	Abdollahi et al, 2018 [19]	?	+	+	-	+	+	+	-	+
2	Araújo Faria et al, 2021 [20]	+	+	+	-	+	+	+	-	+
3	Beetz et al, 2012a [21]	+	-	-	+	+	+	+	-	+
4	Beetz et al, 2012b [7]	+	+	-	+	+	+	+	-	+
5	Buettner et al., 2012 [22]	+	+	+	-	+	+	+	-	+
6	Cheng et al., 2019 [23]	?	+	+	-	+	+	+	-	+
7	Dean et al, 2017 [24]	+	+	+	-	+	+	+	+	+
8	Dean et al, 2018 [15]	+	+	+	-	+	+	+	-	+
9	Dohopolski et al., 2022 [25]	+	-	-	+	+	+	+	-	+
10	Gabry´s et al, 2018 [26]	+	+	+	+	+	+	+	-	+
11	Humbert-Vidan et al, 2021 [27]	+	+	+	-	+	+	+	-	+
12	Jiang et al, 2019 [28]	+	+	+	-	+	+	+	-	+
13	Liu et al, 2019 [29]	?	+	+	-	+	+	+	-	+
14	Men et al, 2019 [30]	+	+	+	+	+	+	+	+	+
15	Nakatsugawa et al, 2019 [31]	+	+	+	-	+	+	+	-	+
16	Nardone et al, 2018 [32]	+	+	+	+	+	+	+	+	+
17	Pota et al, 2017 [33]	?	?	?	-	+	+	+	-	+
18	Rosen et al, 2018 [34]	+	+	+	+	+	+	+	+	+
19	Sheikh et al, 2019 [35]	+	+	+	+	+	+	+	+	+
20	Smyczynska et al., 2021 [5]	+	?	?	-	+	+	+	-	+
21	Soares et al, 2018 [16]	?	+	-	-	+	+	+	-	+
22	Ursino et al, 2021 [36]	+	+	+	-	+	+	+	-	+
23	van Dijk et al., 2017a [6]	+	-	-	+	+	+	+	-	+
24	van Dijk et al., 2017b [37]	+	-	-	+	+	+	+	-	+
25	van Dijk et al., 2018a [4]	+	-	-	+	+	+	+	-	+
26	van Dijk et al., 2018b [38]	+	-	-	+	+	+	+	-	+
27	van Dijk et al., 2019 [39]	+	-	+	-	+	+	+	-	+
28	Wentzel et al., 2020 [2]	+	-	-	-	+	+	+	-	+

PROBAST = Prediction model Risk Of Bias ASsessment Tool; ROB = risk of bias.
* + indicates low ROB/low concern regarding applicability; − indicates high ROB/high concern regarding applicability; and ? indicates unclear
ROB/unclear concern regarding applicability.

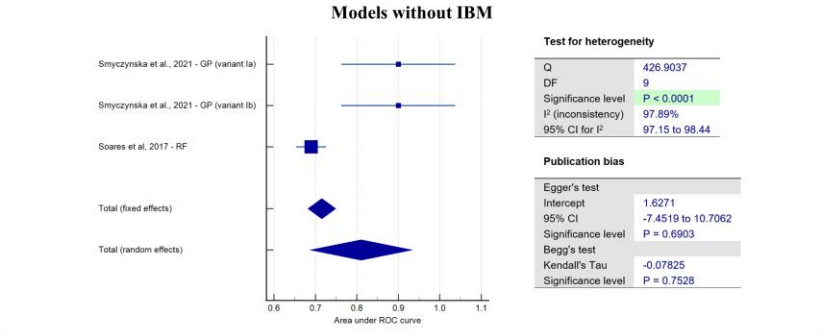
In the Patient domain, 3 studies [16,29,33] presented unclear RoB regarding the omissions of the reasons for patient exclusions. Studies presenting high RoB for the Predictors domain were those using subjective measurements of the endpoint (e.g., patient-rated xerostomia) [4,6,21,25,37–39] when it is expected to use a guideline/criteria or when different endpoint measurements (inconsistent predictors) were made within the same cohort [2,25]. For the Outcome domain, as all outcomes were determined appropriately in a standard way, a high RoB was identified when the outcome was not defined and determined in a similar way for all participants [2,25]. In some studies, predictors were not excluded from the outcome definition (i.e., to predict xerostomia, baseline xerostomia was used as a predictor). In the present analysis, the authors considered the use of baseline xerostomia as a source of bias [4,6,7,16,21,37,38]. Blinding of predictor (i.e., to assess predictor without the knowledge of the outcome status) does not apply to ML-based PM, since we need the conclusion for each patient to properly train a model with retrospective data. Therefore, this signalling question does not apply to the present studies, as well as the time interval between assessments. In the Analysis domain, studies considered as having high RoB were those in which the authors used univariable analysis to select the predictors [27,28], those in which the model performance metrics were not properly evaluated (i.e., studies not presenting AUROC – the only metric that properly evaluates the separation of the patients who developed the endpoint from those who did not – or those presenting only AUROC without confidence interval) [2,5,15,16,19,20,22–24,27–29,31,33,36,39]. Overfitting and optimistic performance were accountable in all models reporting model development with proper internal validation (cross-validation, bootstrapping) with only one study presenting an unclear status for this domain [5]. Sample size considerations are also important, but criteria are not clear for development studies [8] Therefore, only model validation studies were evaluated for this signalling question, with 5 from a total of 9 studies having an appropriate sample size (up to 100 participants) [4,15,16,35,37]. A high RoB was identified in 23 studies (**table 2**). No applicability concerns were raised.

Results of syntheses

The MA of the AUROC of studies that presented external validation [5,15,16] is shown in (**figure 4**). The RoB of the studies included in MA were considered low [15], unclear [5], and high [16]. Additionally, since the studies are highly heterogeneous for both IBM-based models [5,15] (I^2=87,36%), and for non-IBM-based models [5,16] (I^2=97.89%), the authors considered the random effect model to evaluate AUROC curves and compare the IBM-based models and the non-IBM-based models, avoiding relying on overestimated results. This analysis shows an AUROC area of 0.82 for models with IBMs and 0.81 for models without IBMs (p value <0.001), meaning there is no difference between the prediction performance of IBM- and non-IBM-based models, with IBM-based models performing only slightly better.



Study	ROC Area	Standard Error	95% CI	z	P	Weight (%)	
						Fixed	Random
Dean et al., 2018 - Penalized LR (standard)	0.820	0.00422	0.812 to 0.828			21.81	11.70
Dean et al., 2018 - Penalized LR (spatial)	0.730	0.00527	0.720 to 0.740			13.96	11.67
Dean et al., 2018 - SVM (standard)	0.820	0.00422	0.812 to 0.828			21.81	11.70
Dean et al., 2018 - SVM (spatial)	0.730	0.00527	0.720 to 0.740			13.96	11.67
Dean et al., 2018 - RF (standard)	0.780	0.00527	0.770 to 0.790			13.96	11.67
Dean et al., 2018 - RF (spatial)	0.750	0.00527	0.740 to 0.760			13.96	11.67
Smyczynska et al., 2021 - LR (variant Ia)	0.890	0.0700	0.753 to 1.000			0.079	6.07
Smyczynska et al., 2021 - MLP4 (variant Ib)	0.940	0.0500	0.842 to 1.000			0.16	7.95
Smyczynska et al., 2021 - MLP2 (variant Ia)	0.950	0.0500	0.852 to 1.000			0.16	7.95
Smyczynska et al., 2021 - MLP4 (variant Ib)	0.940	0.0500	0.842 to 1.000			0.16	7.95
Total (fixed effects)	0.780	0.00197	0.776 to 0.784	396.193	<0.001	100.00	100.00
Total (random effects)	0.820	0.0248	0.771 to 0.868	33.043	<0.001	100.00	100.00



Study	ROC Area	Standard Error	95% CI	z	P	Weight (%)	
						Fixed	Random
Smyczynska et al., 2021 - GP (variant Ia)	0.900	0.0700	0.763 to 1.000			6.05	28.64
Smyczynska et al., 2021 - GP (variant Ib)	0.900	0.0700	0.763 to 1.000			6.05	28.64
Soares et al, 2017 - RF	0.690	0.0184	0.654 to 0.726			87.89	42.73
Total (fixed effects)	0.715	0.0172	0.682 to 0.749	41.539	<0.001	100.00	100.00
Total (random effects)	0.810	0.0629	0.687 to 0.934	12.874	<0.001	100.00	100.00

MedCalc® Statistical Software version 20.113 (MedCalc Software Ltd, Ostend, Belgium; https://www.medcalc.org; 2022)

Figure 4. Meta-analysis

Certainty of evidence

The certainty of the body of evidence from studies included in MA were appraised as low certainty of evidence downgraded for bias and inconsistency (**Appendix III**). This result was mainly due the high RoB and great inconsistency (I^2) of studies that contribute to MA. It is worthy to mention that, since these models are developed and tested based on the real outcome (i.e., toxicity development) of a cohort of patients, the impact of the outcome and their importance rating does not fit in the present analysis. A narrative GRADEpro table is presented, without the magnitude of the effect estimation, since confidence intervals were not reported by the studies included in the MA.

Discussion

The present SR evaluated PMSs based on multi-variable models to predict a wide range of toxicities from HNC treatment. Multi-variable models based on LR are on the frontier of statistics and ML-based modelling. The primary researcher (ALDA) conducted a previous literature review to identify the state-of-the-art for toxicity prediction and retrieved two relevant Systematic Reviews (SR) regarding the use of radiomics and ML for radiotherapy in HNC, which included several outcomes (such as HPV status, distant metastasis, etc.) [40], and addressed the radiation-induced toxicity prediction in HNC through radiomics [41]. Despite these findings, the authors proceed with the present SR because it is suitable to identify PMSs that address a wide range of HNC treatments, including not only toxicities but also common complications of HNC treatment, and compares models developed using IBMs against not using IBMs.

Some authors [26] advocate that there are no published studies systematically evaluating how distinct RL-based models are from NTCP models. Essentially, this distinction relies mostly upon Deep Learning models being essentially different from classical ML models like LR, and therefore, data modality and selection of predictors are also distinct. There is an urge to understand if there is a gain in prediction power by using these models, since NTCP models present a series of limitations as they discard organ-specific spatial information, and are usually based on a single CT. Moreover, data format, treatment modalities/planning, patient population, cancer type, and predictors enrolled in developing a PM can affect the interpretability of the results.

Traditional models are dependent on the sample size and need more examples to better separate classes. A methodological intrinsic bias of such studies may be modelling according

to sample-specific features, as when, if all patients receive similar radiation dose, these metrics will not be statistically significant to be discriminative. This patient selection bias affects the model's performance and jeopardizes consistent conclusions from primary studies, whose rationale is the prediction of toxicities to facilitate clinical conduct towards a personalized treatment plan. Additionally, it reinforces the need for more personalized treatments (how can we train a model to correctly address and personalize treatment if the model was trained with a standard and generic profile?). According to the present SR, the distinction among RL models performing better than classical models was not clear.

Models based only on dosimetric features use mean dose and partial information from the dose-volume histogram (DVH), while conventional NTCP models utilize all information of the DVH curve by concatenating all metrics to a single factor with dose-response functions (DRFs). Radiomics models, on the other hand, extract features directly from medical images (known as textural or intensity features). Similarly, Dosiomics attempt to extract 3D spatial features from radiation therapy dose distribution. For this reason, to allow proper comparison and analyses of the reported PMs, the present SR takes into account PMs based on imaging data PMs based on non-imaging data (clinical and dosimetric).

According to some authors, radiomic data does not show superiority over NTCP models [5,7]. According to the present SR, no inference can be made regarding the use of IBM-based models performing better than non-IBM-based models. The authors expected to visualize differences between studies not using imaging data and those using it by plotting the AUROC curves, but no sufficient distinction among the values was seen to stand for one modality. MA is highly indicated to properly assess all available well-reported evidence.

The TRIPOD statement is a checklist for reporting multivariable PM for individual prognosis or diagnosis. This tool is designed to aid in the transparent reporting of PMSs and the authors suggest that a copy of this checklist must be provided along with the primary prognostic study in submission. When a single dataset is available, all data can be used to develop the model (types 1a) or to develop and "internally validate" the model through resampling strategies (type 1b) as cross-validation, with both fitting in the model development category. When a single dataset is split for development and validation purposes, this split can be random (type 2a) or non-random (type 2b), with type 2a being an inefficient form of internal validation [11] and type 2b being considered an intermediary between "internal" and "external validation". If separate data sets are available for development and validation, it characterizes an "external validation study" (type 3). Type 4 studies are those validating the

performance of previously published studies on separate data, also fitting in external validation. According to the retrieved papers, frequently the same study conducts different splitting regimens for each model, thus the reviewers conducted this analysis for each model according to the sampling strategy [12]. The type of study is a great source of bias according to the PROBAST.

PROBAST [8,11] is a tool for RoB assessment and applicability of PMSs. The first step of PROBAST focuses on specifying the review question [11]. For the present SR, the intended use of the model is for toxicity prediction, the targeted participants are HNC patients, the predictors used in the modelling are clinical, dosiomic and radiomic features, and the predicted outcomes are HNC treatment toxicities. The second step addresses the classification of PMSs, which can be defined as: i) a development study, which focuses on developing the model without external validation, frequently including resampling strategies or "internal validation" (e.g., cross-validation, bootstrapping) to assess the predictive performance without bias; ii) a validation study, which quantifies the predictive performance in an "external dataset"; and iii) both. This assessment applies to each different model reported in studies. The third next step is to assess RoB and applicability, with all domains being evaluated separately for each distinct model in each study. RoB can occur if the design, conduction, or analysis are not ideal, which could lead to systematic error occurence and distortion of a model's performance estimation. Applicability concerns refers to a situation when the study population, predictors or outcomes differ from those specified in the review question, which can easily occur since SR questions are usually broad. A high RoB was also identified in studies that quantifies the predictive performance of the model in the same dataset, which tends to provide an over-optimistic performance (even more if the univariable analysis is used to select the predictors, or if forward stepwise selection takes place in multivariable analysis). In such case, an internal validation step with bootstrapping or cross-validation is required [8].

Since xerostomia is the most investigated endpoint, the authors made a specific assessment of the design of these PMSs. The impact of using this predictor for training ML classifiers is uncertain, as frequently, a pre-selection step is conducted to identify which predictors are most likely to be associated with the outcome; additionally, some studies consistently apply xerostomia baseline to train the models. In the present analysis, the authors considered the use of baseline xerostomia as a source of bias [4,6,7,16,21,37,38]. Xerostomia is a complex endpoint due to the subjectivity and the influence of associated factors such as

age, dosimetric parameters and baseline xerostomia. According to [21], the response curve for baseline xerostomia patients is different from patients without baseline xerostomia and therefore, this methodological choice inserts bias in PMSs. This multifactorial influence also applies to baseline xerostomia being used as a predictor to train the model, meaning the patients may present none or some level of baseline xerostomia prior to treatment, which is usually referred to by some authors as the models having the best prediction power, which could be simply an over-optimistic result. As demonstrated by [4,39], IBMs predictors are not associated with xerostomia baseline and, in fact, IBMs are responsible for improving the results. Additionally, the association of baseline xerostomia with the outcome may be only a coincidence, since the dose usually correlates with xerostomia[39]. Therefore, if a predictor defines or composes the outcome, it is most likely to lead to over-optimistic results [8].

There is a trend where using imaging biomarkers has been shown to improve toxicity prediction [4,6,34,37–39] but according to the present meta-analysis, multimodality image-featured models have a similar performance when compared to those without imaging data. Regarding the studies included in the present meta-analysis, there is only a marginal improvement seen in IBM-based PMs. It is worth considering that, even though these are based on types 3 and 4 PMSs, and being corrected based on heterogeneity for MA, only three studies attend the criteria to be included in the meta-analysis. The high heterogeneity of primary studies is explained by the tumor's site and treatment modalities being highly variable, as well as the selected predictors, and the models to predict the endpoints [9]. The authors advise the readers to have a critical interpretation of the results. Additionally, in medical research, the amount of data is crucial and, frequently, not easy to retrieve in terms of complete medical documentation and follow-up. Ideally, external validation or independent test should be performed for better assessment of models' generalization ability. This final step is important to provide a notion of how well the model can perform when assessing unseen data.

Finally, the scientific literature on AI for medical imaging is vast and diverse, with results being reported in metrics and graphical illustrations, sometimes indicated for punctual purposes. The *Pattern Recognition Community* [42] adopts true positives (TP), false negatives (FN), classification accuracy (TP+TN)/n and F1-score [2TP/(2TP+FP+FN)] to evaluate models' performances. However, to perform a meta-analysis, an evaluation of the pooled AUROC curve requires values of confidence interval and standard error (calculated from the square root of the test sample and standard deviation) [43]. The area under the receiver

operating characteristic (AUROC) curve is considered a more intuitive, discriminatory, and consistent measure than accuracy and shows the true positive rate (TPR) against the false positive rate (FPR) [FP/(FP+TN] to further evaluate the model. This metric provides a quantitative notion of class separation, recommended when there is a binary classification problem [44].

Main conclusions

The development of a PM according to sample-specific features represents patient selection bias and may affect a model's performance. This is the major drawback that impairs good performance models to be developed and implemented, preventing the dissemination of these support systems. Feature selection is a great challenge in radiomic studies and special attention should be given to retrieval of reliable information on several biomarkers and their correlated clinical outcomes. For future studies and based on the present SR, we recommend the readers to check a summary of data and a specific set of features that has proven to be more representative and has reached better performances

An important methodological choice that impairs proper evaluation is the heterogeneity of the studies and the absence of an independent/external test to assess the generalization ability of the models. Moreover, performance metrics are widely variable within studies. The authors suggest that the following metrics should be reported to allow fair comparison and future meta-analysis: sensitivity (recall), precision (positive predictive value), specificity, AUROC, confidence interval, and standard error. The authors also highly recommend that future studies should always report the confusion matrix, since it allows the calculation of several metrics. The absence of metrics impaired studies to be included in meta-analysis, which limits further conclusions. To conclude, IBM-featured PMs are not superior to PMs based on non-IBM predictors. However, it is important to state that the present MA was conducted among models from only three studies that present high inconsistency, high RoB and low certainty of evidence.

Other Information

Protocol and registration

The present SR was conducted following the guidelines of Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) [45,46] and the PRISMA-P [47,48]

checklist, which is registered at the International Prospective Register of Systematic Reviews (PROSPERO) database under protocol number CRD42020219304.

Acknowledgments

The authors would like to gratefully acknowledge the Coordination for the Improvement of Higher Education Personnel (CAPES/PROEX, Brazil) process number 001, the National Council for Scientific and Technological Development (CNPq, Brazil), and the grants from São Paulo Research Foundation (FAPESP, Brazil) process number 2019/26676-7.

Ethics Approval

Not applicable.

Author contribution

All authors made substantial contributions to the conception, draft, and review of the data for the work.

Funding

None.

References

[1] Chinnery T, Arifin A, Tay KY, Leung A, Nichols AC, Palma DA, et al. Utilizing Artificial Intelligence for Head and Neck Cancer Outcomes Prediction From Imaging. Canadian Association of Radiologists Journal 2021;72:73–85. https://doi.org/10.1177/0846537120942134.

[2] Wentzel A, Hanula P, van Dijk L v., Elgohari B, Mohamed ASR, Cardenas CE, et al. Precision toxicity correlates of tumor spatial proximity to organs at risk in cancer patients receiving intensity-modulated radiotherapy. Radiotherapy and Oncology 2020;148:245–51. https://doi.org/10.1016/j.radonc.2020.05.023.

[3] Kutcher GJ, Burman C. Calculation of complication probability factors for non-uniform normal tissue irradiation: The effective volume method gerald. International Journal of Radiation Oncology*Biology*Physics 1989;16:1623–30. https://doi.org/10.1016/0360-3016(89)90972-3.

[4] van Dijk L v., Noordzij W, Brouwer CL, Boellaard R, Burgerhof JGM, Langendijk JA, et al. 18F-FDG PET image biomarkers improve prediction of late radiation-induced xerostomia. Radiotherapy and Oncology 2018;126:89–95. https://doi.org/10.1016/j.radonc.2017.08.024.

[5] Smyczynska U, Grabia S, Nowicka Z, Papis-Ubych A, Bibik R, Latusek T, et al. Prediction of Radiation-Induced Hypothyroidism Using Radiomic Data Analysis Does Not Show Superiority over Standard Normal Tissue Complication Models. Cancers (Basel) 2021;13:5584. https://doi.org/10.3390/cancers13215584.

[6] van Dijk L v., Brouwer CL, van der Schaaf A, Burgerhof JGM, Beukinga RJ, Langendijk JA, et al. CT image biomarkers to improve patient-specific prediction of radiation-induced xerostomia and sticky saliva. Radiotherapy and Oncology 2017;122:185–91. https://doi.org/10.1016/j.radonc.2016.07.007.

[7] Beetz I, Schilstra C, Burlage FR, Koken PW, Doornaert P, Bijl HP, et al. Development of NTCP models for head and neck cancer patients treated with three-dimensional conformal radiotherapy for xerostomia and sticky saliva: The role of dosimetric and clinical factors. Radiotherapy and Oncology 2012;105:86–93. https://doi.org/10.1016/j.radonc.2011.05.010.

[8] 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:W1. https://doi.org/10.7326/M18-1377.

[9] Debray TPA, Damen JAAG, Snell KIE, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ 2017;356:i6460. https://doi.org/10.1136/bmj.i6460.

[10] Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Syst Rev 2016;5:210. https://doi.org/10.1186/s13643-016-0384-4.

[11] 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;170:51. https://doi.org/10.7326/M18-1376.

[12] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. Ann Intern Med 2015;162:55–63. https://doi.org/10.7326/M14-0697.

[13] Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, 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:W1–73. https://doi.org/10.7326/M14-0698.

[14] 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:e048008. https://doi.org/10.1136/bmjopen-2020-048008.

[15] Dean J, Wong K, Gay H, Welsh L, Jones A-B, Schick U, et al. Incorporating spatial dose metrics in machine learning-based normal tissue complication probability (NTCP) models of severe acute dysphagia resulting from head and neck radiotherapy. Clin Transl Radiat Oncol 2018;8:27–39. https://doi.org/10.1016/j.ctro.2017.11.009.

[16] Soares I, Dias J, Rocha H, Khouri L, do Carmo Lopes M, Ferreira B. Predicting xerostomia after IMRT treatments: a data mining approach. Health Technol (Berl) 2018;8:159–68. https://doi.org/10.1007/s12553-017-0204-4.

[17] Schünemann H, Brożek J, Guyatt G, Oxman A, editors. GRADE handbook for grading quality of evidence and strength of recommendations. The GRADE Working Group, 2013. Available from guidelinedevelopment.org/handbook; n.d.

[18] GRADEpro GDT: GRADEpro Guideline Development Tool [Software] McMaster University and Evidence Prime, 2022. Available from gradepro.org. n.d.

[19] Abdollahi H, Mostafaei S, Cheraghi S, Shiri I, Rabi Mahdavi S, Kazemnejad A. Cochlea CT radiomics predicts chemoradiotherapy induced sensorineural hearing loss in head

and neck cancer patients: A machine learning and multi-variable modelling study. Physica Medica 2018;45:192–7. https://doi.org/10.1016/j.ejmp.2017.10.008.

[20] de Araujo Faria V, Azimbagirad M, Viani Arruda G, Fernandes Pavoni J, Cezar Felipe J, dos Santos EMCMF, et al. Prediction of Radiation-Related Dental Caries Through PyRadiomics Features and Artificial Neural Network on Panoramic Radiography. J Digit Imaging 2021;34:1237–48. https://doi.org/10.1007/s10278-021-00487-6.

[21] Beetz I, Schilstra C, van der Schaaf A, van den Heuvel ER, Doornaert P, van Luijk P, et al. NTCP models for patient-rated xerostomia and sticky saliva after treatment with intensity modulated radiotherapy for head and neck cancer: The role of dosimetric and clinical factors. Radiotherapy and Oncology 2012;105:101–6. https://doi.org/10.1016/j.radonc.2012.03.004.

[22] Buettner F, Miah AB, Gulliford SL, Hall E, Harrington KJ, Webb S, et al. Novel approaches to improve the therapeutic index of head and neck radiotherapy: An analysis of data from the PARSPORT randomised phase III trial. Radiotherapy and Oncology 2012;103:82–7. https://doi.org/10.1016/j.radonc.2012.02.006.

[23] Cheng Z, Nakatsugawa M, Zhou XC, Hu C, Greco S, Kiess A, et al. Utility of a Clinical Decision Support System in Weight Loss Prediction After Head and Neck Cancer Radiotherapy. JCO Clin Cancer Inform 2019:1–11. https://doi.org/10.1200/CCI.18.00058.

[24] Dean JA, Welsh LC, Wong KH, Aleksic A, Dunne E, Islam MR, et al. Normal Tissue Complication Probability (NTCP) Modelling of Severe Acute Mucositis using a Novel Oral Mucosal Surface Organ at Risk. Clin Oncol 2017;29:263–73. https://doi.org/10.1016/j.clon.2016.12.001.

[25] Dohopolski M, Wang K, Morgan H, Sher D, Wang J. Use of deep learning to predict the need for aggressive nutritional supplementation during head and neck radiotherapy. Radiotherapy and Oncology 2022;171:129–38. https://doi.org/10.1016/j.radonc.2022.04.016.

[26] Gabryś HS, Buettner F, Sterzing F, Hauswald H, Bangert M. Design and Selection of Machine Learning Methods Using Radiomics and Dosiomics for Normal Tissue Complication Probability Modeling of Xerostomia. Front Oncol 2018;8. https://doi.org/10.3389/fonc.2018.00035.

[27] Humbert-Vidan L, Patel V, Oksuz I, King AP, Guerrero Urbano T. Comparison of machine learning methods for prediction of osteoradionecrosis incidence in patients with head and neck cancer. Br J Radiol 2021;94:20200026. https://doi.org/10.1259/bjr.20200026.

[28] Jiang W, Lakshminarayanan P, Hui X, Han P, Cheng Z, Bowers M, et al. Machine Learning Methods Uncover Radiomorphologic Dose Patterns in Salivary Glands that Predict Xerostomia in Patients with Head and Neck Cancer. Adv Radiat Oncol 2019;4:401–12. https://doi.org/10.1016/j.adro.2018.11.008.

[29] Liu Y, Shi H, Huang S, Chen X, Zhou H, Chang H, et al. Early prediction of acute xerostomia during radiation therapy for nasopharyngeal cancer based on delta radiomics from CT images. Quant Imaging Med Surg 2019;9:1288–302. https://doi.org/10.21037/qims.2019.07.08.

[30] Men K, Geng H, Zhong H, Fan Y, Lin A, Xiao Y. A Deep Learning Model for Predicting Xerostomia Due to Radiation Therapy for Head and Neck Squamous Cell Carcinoma in the RTOG 0522 Clinical Trial. International Journal of Radiation Oncology*Biology*Physics 2019;105:440–7. https://doi.org/10.1016/j.ijrobp.2019.06.009.

[31] Nakatsugawa M, Cheng Z, Kiess A, Choflet A, Bowers M, Utsunomiya K, et al. The Needs and Benefits of Continuous Model Updates on the Accuracy of RT-Induced

Toxicity Prediction Models Within a Learning Health System. International Journal of Radiation Oncology*Biology*Physics 2019;103:460–7. https://doi.org/10.1016/j.ijrobp.2018.09.038.

[32] Nardone V, Tini P, Nioche C, Mazzei MA, Carfagno T, Battaglia G, et al. Texture analysis as a predictor of radiation-induced xerostomia in head and neck patients undergoing IMRT. Radiol Med 2018;123:415–23. https://doi.org/10.1007/s11547-017-0850-7.

[33] Pota M, Scalco E, Sanguineti G, Farneti A, Cattaneo GM, Rizzo G, et al. Early prediction of radiotherapy-induced parotid shrinkage and toxicity based on CT radiomics and fuzzy classification. Artif Intell Med 2017;81:41–53. https://doi.org/10.1016/j.artmed.2017.03.004.

[34] Rosen BS, Hawkins PG, Polan DF, Balter JM, Brock KK, Kamp JD, et al. Early Changes in Serial CBCT-Measured Parotid Gland Biomarkers Predict Chronic Xerostomia After Head and Neck Radiation Therapy. International Journal of Radiation Oncology*Biology*Physics 2018;102:1319–29. https://doi.org/10.1016/j.ijrobp.2018.06.048.

[35] Sheikh K, Lee SH, Cheng Z, Lakshminarayanan P, Peng L, Han P, et al. Predicting acute radiation induced xerostomia in head and neck Cancer using MR and CT Radiomics of parotid and submandibular glands. Radiation Oncology 2019;14:131. https://doi.org/10.1186/s13014-019-1339-4.

[36] Ursino S, Giuliano A, Martino F di, Cocuzza P, Molinari A, Stefanelli A, et al. Incorporating dose–volume histogram parameters of swallowing organs at risk in a videofluoroscopy-based predictive model of radiation-induced dysphagia after head and neck cancer intensity-modulated radiation therapy. Strahlentherapie Und Onkologie 2021;197:209–18. https://doi.org/10.1007/s00066-020-01697-7.

[37] van Dijk L v., Brouwer CL, van der Laan HP, Burgerhof JGM, Langendijk JA, Steenbakkers RJHM, et al. Geometric Image Biomarker Changes of the Parotid Gland Are Associated With Late Xerostomia. International Journal of Radiation Oncology*Biology*Physics 2017;99:1101–10. https://doi.org/10.1016/j.ijrobp.2017.08.003.

[38] van Dijk L v., Thor M, Steenbakkers RJHM, Apte A, Zhai T-T, Borra R, et al. Parotid gland fat related Magnetic Resonance image biomarkers improve prediction of late radiation-induced xerostomia. Radiotherapy and Oncology 2018;128:459–66. https://doi.org/10.1016/j.radonc.2018.06.012.

[39] van Dijk L v., Langendijk JA, Zhai T-T, Vedelaar TA, Noordzij W, Steenbakkers RJHM, et al. Delta-radiomics features during radiotherapy improve the prediction of late xerostomia. Sci Rep 2019;9:12483. https://doi.org/10.1038/s41598-019-48184-3.

[40] Giraud P, Giraud P, Gasnier A, el Ayachy R, Kreps S, Foy J-P, et al. Radiomics and Machine Learning for Radiotherapy in Head and Neck Cancers. Front Oncol 2019;9. https://doi.org/10.3389/fonc.2019.00174.

[41] Carbonara R, Bonomo P, di Rito A, Didonna V, Gregucci F, Ciliberti MP, et al. Investigation of Radiation-Induced Toxicity in Head and Neck Cancer Patients through Radiomics and Machine Learning: A Systematic Review. J Oncol 2021;2021:1–9. https://doi.org/10.1155/2021/5566508.

[42] Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit 1997;30:1145–59. https://doi.org/10.1016/S0031-3203(96)00142-2.

[43] Borenstein M, Hedges L, Higgins J, Rothstein H. Introduction to meta-analysis. Chichester, UK: Wiley; 2009.

[44] Jin Huang, Ling CX. Using AUC and accuracy in evaluating learning algorithms. IEEE Trans Knowl Data Eng 2005;17:299–310. https://doi.org/10.1109/TKDE.2005.50.
[45] 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:n71. https://doi.org/10.1136/bmj.n71.
[46] Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 2021;372:n160. https://doi.org/10.1136/bmj.n160.
[47] Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev 2015;4:1. https://doi.org/10.1186/2046-4053-4-1.
[48] Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. BMJ 2015;349:g7647–g7647. https://doi.org/10.1136/bmj.g7647.

APPENDIX I - PROTOCOL

Machine Learning for the prediction of toxicities from head and neck cancer treatment: a systematic review with meta-analysis (PROSPERO protocol number: CRD42020219304)

Review question: Prediction Models can accurately predict head and neck cancer treatment toxicities?

II) What are the most addressed toxicities in prediction studies?
III) Which machine learning models were reported for toxicity prediction?
IV) Image-based models have better performance when compared to non-image-based models?

PICOS strategy
P - Head and neck cancer patients
I – Cancer treatment
C – not applicable
O - Toxicity prediction
S - Prediction Model Studies

Searching keywords	
#1	"head and neck" 'head and neck' (Embase variation) "head and neck" OR "cabeça e pescoço" OR "cabeza y cuello" (LILACS variation)
#2	"cancer treatment" OR radiotherapy OR "radiation therapy" OR "radiation oncology" OR "intensity modulated radiation therapy" OR chemotherapy OR "concomitant chemotherapy" 'cancer treatment' OR radiotherapy OR 'radiation therapy' OR 'radiation oncology' OR 'intensity modulated radiation therapy' OR chemotherapy OR 'concomitant chemotherapy' (Embase variation)
#3a	"side effects" OR "adverse effects" OR "adverse events" OR outcome OR toxicit* OR "radiation toxicit*" OR "radiotherapy-induced toxicity" 'side effects' OR 'adverse effects' OR 'adverse events' OR outcome OR toxicit* OR 'radiation toxicit*' OR 'radiotherapy-induced toxicity' (Embase variation)
#3b	"side effects" OR "adverse effects" OR "adverse events" OR outcome OR toxicit* OR "chemotherapy toxicit*" OR "chemotherapy-induced toxicity" OR "drug-induced toxicity" 'side effects' OR 'adverse effects' OR 'adverse events' OR outcome OR toxicit* OR 'chemotherapy toxicit*' OR 'chemotherapy-induced toxicity' OR 'drug-induced toxicity' (Embase variation)
#4	prediction
#5	"artificial intelligence" OR "machine learning" OR "deep learning" OR "convolutional neural network" OR "artificial neural network" 'artificial intelligence' OR 'machine learning' OR 'deep learning' OR 'convolutional neural network' OR 'artificial neural network' (Embase variation) "artificial intelligence" OR "inteligência artificial" OR "inteligencia artificial" (LILACS variation)
#6	radiomic* radiomic* OR radiômica* OR "radiómico" (LILACS variation)

Strategy I) #1 #2 #3a #4 #5
Strategy II) #1 #2 #3b #4 #5
Strategy III) #1 #2 #5 #6

Database	**#**	**Query** Search date: 18th june, 2022	**Results**
PubMed	I	((((("head and neck") AND ("cancer treatment" OR radiotherapy OR "radiation therapy" OR "radiation oncology" OR "intensity modulated radiation therapy" OR chemotherapy OR "concomitant chemotherapy")) AND ("side effects" OR "adverse effects" OR "adverse events" OR outcome OR toxicit* OR "radiation toxicit*" OR "radiotherapy-induced toxicity")) AND (prediction)) AND ("artificial intelligence" OR "machine learning" OR "deep learning" OR "convolutional neural network" OR "artificial neural network")	126
	II	((((“head and neck”) AND (“cancer treatment” OR radiotherapy OR “radiation therapy” OR “radiation oncology” OR “intensity modulated radiation therapy” OR chemotherapy OR “concomitant chemotherapy”)) AND (“side effects” OR “adverse effects” OR “adverse events” OR outcome OR toxicit* OR “chemotherapy toxicit*” OR “chemotherapy-induced toxicity” OR “drug-induced toxicity”)) AND (prediction)) AND (“artificial intelligence” OR “machine learning” OR “deep learning” OR “convolutional neural network” OR “artificial neural network”)	
	III	(((“head and neck”) AND (“cancer treatment” OR radiotherapy OR “radiation therapy” OR “radiation oncology” OR “intensity modulated radiation therapy” OR chemotherapy OR “concomitant chemotherapy”)) AND (“artificial intelligence” OR “machine learning” OR “deep learning” OR “convolutional neural network” OR “artificial neural network”)) AND (radiomic*)	
Embase	I	'head and neck' AND ('cancer treatment' OR radiotherapy OR 'radiation therapy' OR 'radiation oncology' OR 'intensity modulated radiation therapy' OR chemotherapy OR 'concomitant chemotherapy') AND ('side effects' OR 'adverse effects' OR 'adverse events' OR outcome OR toxicit* OR 'radiation toxicit*' OR 'radiotherapy-induced toxicity') AND prediction AND ('artificial intelligence' OR 'machine learning' OR 'deep learning' OR 'convolutional neural network' OR 'artificial neural network')	281
	II	'head and neck' AND ('cancer treatment' OR radiotherapy OR 'radiation therapy' OR 'radiation oncology' OR 'intensity modulated radiation therapy' OR chemotherapy OR 'concomitant chemotherapy') AND ('side effects' OR 'adverse effects' OR 'adverse events' OR outcome OR toxicit* OR 'chemotherapy toxicit*' OR 'chemotherapy-induced toxicity' OR 'drug-induced toxicity') AND	

		prediction AND ('artificial intelligence' OR 'machine learning' OR 'deep learning' OR 'convolutional neural network' OR 'artificial neural network')	
	III	'head and neck' AND ('cancer treatment' OR radiotherapy OR 'radiation therapy' OR 'radiation oncology' OR 'intensity modulated radiation therapy' OR chemotherapy OR 'concomitant chemotherapy') AND ('artificial intelligence' OR 'machine learning' OR 'deep learning' OR 'convolutional neural network' OR 'artificial neural network') AND radiomic*	
Scopus	I	(TITLE-ABS-KEY ("head and neck") AND TITLE-ABS-KEY ("cancer treatment" OR radiotherapy OR "radiation therapy" OR "radiation oncology" OR "intensity modulated radiation therapy" OR chemotherapy OR "concomitant chemotherapy") AND TITLE-ABS-KEY ("side effects" OR "adverse effects" OR "adverse events" OR outcome OR toxicit* OR "radiation toxicit*" OR "radiotherapy-induced toxicity") AND TITLE-ABS-KEY (prediction) AND TITLE-ABS-KEY ("artificial intelligence" OR "machine learning" OR "deep learning" OR "convolution neural network" OR "artificial neural network"))	103
	II	(TITLE-ABS-KEY ("head and neck") AND TITLE-ABS-KEY ("cancer treatment" OR radiotherapy OR "radiation therapy" OR "radiation oncology" OR "intensity modulated radiation therapy" OR chemotherapy OR "concomitant chemotherapy") AND TITLE-ABS-KEY ("side effects" OR "adverse effects" OR "adverse events" OR outcome OR toxicit* OR "chemotherapy toxicit*" OR "chemotherapy-induced toxicity" OR "drug-induced toxicity") AND TITLE-ABS-KEY (prediction) AND TITLE-ABS-KEY ("artificial intelligence" OR "machine learning" OR "deep learning" OR "convolutional neural network" OR "artificial neural network"))	
	III	(TITLE-ABS-KEY ("head and neck") AND TITLE-ABS-KEY ("cancer treatment" OR radiotherapy OR "radiation therapy" OR "radiation oncology" OR "intensity modulated radiation therapy" OR chemotherapy OR "concomitant chemotherapy") AND TITLE-ABS-KEY ("artificial intelligence" OR "machine learning" OR "deep learning" OR "convolutional neural network" OR "artificial neural network") AND TITLE-ABS-KEY (radiomic*))	
Cochrane	I	“head and neck” in Title Abstract Keyword AND “cancer treatment” OR radiotherapy OR “radiation therapy” OR “radiation oncology” OR “intensity modulated radiation therapy” OR chemotherapy OR “concomitant chemotherapy” in Title Abstract Keyword AND “side effects” OR “adverse effects” OR “adverse events” OR	7

		outcome OR toxicit* OR “radiation toxicit*” OR “radiotherapy-induced toxicity” in Title Abstract Keyword AND prediction in Title Abstract Keyword AND “artificial intelligence” OR “machine learning” OR “deep learning” OR “convolutional neural network” OR “artificial neural network” in Title Abstract Keyword	
	II	“head and neck” in Title Abstract Keyword AND “cancer treatment” OR radiotherapy OR “radiation therapy” OR “radiation oncology” OR “intensity modulated radiation therapy” OR chemotherapy OR “concomitant chemotherapy” in Title Abstract Keyword AND “side effects” OR “adverse effects” OR “adverse events” OR outcome OR toxicit* OR “chemotherapy toxicit*” OR “chemotherapy-induced toxicity” OR “drug-induced toxicity” in Title Abstract Keyword AND prediction in Title Abstract Keyword AND “artificial intelligence” OR “machine learning” OR “deep learning” OR “convolutional neural network” OR “artificial neural network” in Title Abstract Keyword	
	III	“head and neck” in Title Abstract Keyword AND “cancer treatment” OR radiotherapy OR “radiation therapy” OR “radiation oncology” OR “intensity modulated radiation therapy” OR chemotherapy OR “concomitant chemotherapy” in Title Abstract Keyword AND “artificial intelligence” OR “machine learning” OR “deep learning” OR “convolutional neural network” OR “artificial neural network” in Title Abstract Keyword AND radiomic* in Title Abstract Keyword	
Web of Science	I	“head and neck” (Topic) and “cancer treatment” OR radiotherapy OR “radiation therapy” OR “radiation oncology” OR “intensity modulated radiation therapy” OR chemotherapy OR “concomitant chemotherapy” (Topic) and “side effects” OR “adverse effects” OR “adverse events” OR outcome OR toxicit* OR “radiation toxicit*” OR “radiotherapy-induced toxicity” (Topic) and prediction (Topic) and “artificial intelligence” OR “machine learning” OR “deep learning” OR “convolutional neural network” OR “artificial neural network” (Topic) https://www.webofscience.com/wos/woscc/summary/adbd661d-58af-4bdc-86e6-427076980465-3ebf771e/relevance/1	52
	II	“head and neck” (Topic) and “cancer treatment” OR radiotherapy OR “radiation therapy” OR “radiation oncology” OR “intensity modulated radiation therapy” OR chemotherapy OR “concomitant chemotherapy” (Topic) and “side effects” OR “adverse effects” OR “adverse events” OR outcome OR toxicit* OR “chemotherapy toxicit*” OR “chemotherapy-induced	

		toxicity" OR "drug-induced toxicity" (Topic) and prediction (Topic) and "artificial intelligence" OR "machine learning" OR "deep learning" OR "convolutional neural network" OR "artificial neural network" (Topic) https://www.webofscience.com/wos/woscc/summary/938fc58b-2d49-4c59-8c1c-e176d16e864b-3ebfc592/relevance/1	
	III	"head and neck" (Topic) and "cancer treatment" OR radiotherapy OR "radiation therapy" OR "radiation oncology" OR "intensity modulated radiation therapy" OR chemotherapy OR "concomitant chemotherapy" (Topic) and "artificial intelligence" OR "machine learning" OR "deep learning" OR "convolutional neural network" OR "artificial neural network" (Topic) and radiomic* (Topic) https://www.webofscience.com/wos/woscc/summary/6f0c2a6f-3ccd-4b6e-9d07-8c3eb8b53336-3ebfb6bc/relevance/1	
Lilacs		("head AND neck" OR "cabeça e pescoço" OR "cabeza y cuello") AND (radiomic* OR radiômica* OR "radiómica")	133
Google Scholar		"head and neck" AND toxicity AND prediction AND radiomics	99**
ProQuest		TI,AB("head and neck") AND TI,AB(toxicity) AND TI,AB(prediction) AND TI,AB(radiomics)	5
OpenGrey		"head and neck" AND toxicity AND prediction AND radiomics	0

This search was conducted in June 18th, 2022.

PPENDIX II - EXCLUDED ARTICLES AND REASONS FOR EXCLUSION

Supplementary Table 1: Excluded articles and reasons for exclusion (to be continued)

	Author	Title	Year	Country	Journal/Conference name	Exclusion Criteria
1	Abe K., et al	The feasibility of MVCT-based radiomics for delta-radiomics in head and neck cancer	2019	Japan	Medical Physics	Conference abstract
2	Abusaif A., et al	Radiomic Correlates of Mandibular Osteoradionecrosis After Radiation Treatment of Head and Neck Cancer Patients	2021	US	International Journal of Radiation Oncology Biology Physics	Conference abstract
3	Berger T., et al	Predicting xerostomia in head and neck cancer using imaging biomarkers from daily tomotherapy MVCTs	2020	UK	Radiotherapy and Oncology	Conference abstract
4	Elgohari B., et al	Mid-Treatment Apparent Diffusion Coefficient Predicts Late Xerostomia following Head and Neck Cancer Radiotherapy	2020	Egypt	International Journal of Radiation Oncology Biology Physics	Conference abstract
5	Elhalawani H., et al	Exploration of an Early Imaging Biomarker of Osteoradionecrosis in Oropharyngeal Cancer Patients: Case-Control Study of the Temporal Changes of Mandibular Radiomics Features	2018	US	International Journal of Radiation Oncology Biology Physics	Conference abstract
6	Hernandez A.M., et al	The role of ensemble machine learning algorithms to predict weight loss following head and neck radiation therapy	2016	US	International Journal of Radiation Oncology	Conference abstract
7	Hui X., et al	A risk prediction model for head and neck radiation toxicities: Novel insights to reduce the risk of head and neck radiation-induced xerostomia	2016	US	International Journal of Radiation Oncology	Conference abstract

Supplementary Table 1: Excluded articles and reasons for exclusion (continuation)

8	Humbert-Vidan L., et al	Prediction of voxelwise mandibular osteoradionecrosis maps in HNC patients using deep learning	2019	UK	Radiotherapy and Oncology	Conference abstract
9	Lakshminarayanan P., et al	A shape-based dose model for the prediction of high grade radiation induced xerostomia for head and neck cancer patients	2017	US	International Journal of Radiation Oncology Biology Physics	Conference abstract
10	Maffei N., et al	A Neural Network predictions and follow-up toxicity correlation to validate re-planning during RT	2016	Italy	Radiotherapy and Oncology	Conference abstract
11	Men K., et al	A deep learning method for xerostomia prediction in head-and-neck radiotherapy	2019	US	Medical Physics	Conference abstract
12	Nakatsugawa M., et al	The value of continuous toxicity updates on the accuracy of prediction models within a learning health system	2017	US	International Journal of Radiation Oncology Biology Physics	Conference abstract
13	Nakatsugawa M., et al	Prediction of toxicity in irradiated head and neck cancer patients based on the geometry of high/middle/low ptvs to surrounding oars	2015	US	International Journal of Radiation Oncology Biology Physics	Conference abstract
14	Neves L.V.F., et al	Feasibility Of Prediction Of Radiation-Related Caries In Head-Neck Cancer Patients Using Machine Learning And Radiomics Features	2020	Brazil	International Journal of Radiation Oncology Biology Physics	Conference abstract
15	Noble D., et al	Does delivered OAR dose improve prediction of late toxicity in head & neck cancer patients?	2020	UK	Radiotherapy and Oncology	Conference abstract
16	Pilz K., et al	Prediction of dysphagia and xerostomia based on CT imaging features of HNSCC patients	2017	Germany	Radiotherapy and Oncology	Conference abstract

Supplementary Table 1: Excluded articles and reasons for exclusion (continuation)

17	Reddy J.P., et al	Applying a Machine Learning Approach to Predict Acute Radiation Toxicities for Head and Neck Cancer Patients	2019	US	International Journal of Radiation Oncology Biology Physics	Conference abstract
18	Reiazi R., et al	The prediction of mandibular osteoradionecrosis in head and neck cancer patients using CT-derived radiomics features	2021	US	Clinical Cancer Research	Conference abstract
19	Reiazi R., et al	The Prediction of Mandibular Osteoradionecrosis (ORN) in Head and Neck Radiotherapy Using CT-Derived Radiomic Features	2021	US	International Journal of Radiation Oncology Biology Physics	Conference abstract
20	Sharma D., et al	Predicting Radiotherapy Response in Head and Neck Patients Using Quantitative Ultrasound	2018	US	IEEE Computer Society	Conference abstract
21	Tseng H.H., et al	A recurrent neural network for xerostomia prediction in head and neck cancer from daily CBCT images	2018	US	Medical Physics	Conference abstract
22	Van Dijk L.V., et al	Prediction of late xerostomia with clinical, atlas based and deep learning contours	2020	US	Radiotherapy and Oncology	Conference abstract
23	Yaohua W., et al	Predicting late symptoms of head and neck cancer treatment using LSTM and patient reported outcomes	2021	US	Proc Int Database Eng Appl Symp	Conference abstract
24	Wojcieszynski A.P., et al Moore J.H., Metz J.M.	Machine Learning to Predict Toxicity in Head and Neck Cancer Patients Treated with Definitive Chemoradiation	2019	US	International Journal of Radiation Oncology Biology Physics	Conference abstract
25	Chinnery T., et al	A CT-based radiomics model for predicting feeding tube insertion in oropharyngeal cancer	2022	US	Proceedings Volume 12033 Spie Medical Imaging	Conference abstract

Supplementary Table 1: Excluded articles and reasons for exclusion (continuation)

26	Zhang H.H., et al	Modeling plan-related clinical complications using machine learning tools in a multiplan IMRT framework	2009	USA	Int J Radiat Oncol Biol Phys	Small sample size
27	Zhang H.H., et al	The minimum knowledge base for predicting organ-at-risk dose-volume levels and plan-related complications in IMRT planning	2010	USA	Phys Med Biol	Small sample size
28	Pardo-Montero J., et al	Classification of tolerable/intolerable mucosal toxicity of head-and-neck radiotherapy schedules with a biomathematical model of cell dynamics.	2021	Spain, UK	Med Phys	Biomathematical model of cell dynamics
29	El Naqa I., et al	Predicting radiotherapy outcomes using statistical learning techniques.	2009	USA	Phys Med Biol	Metric - Matthews correlation coefficient
30	Drago GP., et al	Forecasting the performance status of head and neck cancer patient treatment by an interval arithmetic pruned perceptron.	2002	Italy	IEEE Transactions on Bio-medical Engineering.	To predict the Karnofsky performance status
31	Beetz I., et al	External validation of three-dimensional conformal radiotherapy based NTCP models for patient-rated xerostomia and sticky saliva among patients treated with intensity modulated radiotherapy	2012	Netherlands	Radiotherapy and Oncology	Models developed based on a population treated with a specific technique and extrapolated to a population treated with another technique without external validation
32	Blanco AI., et al	Dose-volume modeling of salivary function in patients with head-and-neck cancer receiving radiotherapy.	2005	USA	Int J Radiat Oncol Biol Phys.	Metric - Akaike information criteria and Bayesian information criterion
33	El Naqa I., et al	Multivariable modeling of radiotherapy outcomes, including dose-volume and clinical factors	2006	USA	Int J Radiat Oncol Biol Phys.	Metric - Akaike information criteria and Bayesian information criterion

USA: United states of America; UK: United Kingdom

APPENDIX III - GRADEpro

Question: Prediction Models can accurately predict head and neck cancer treatment toxicities?

№ of studies	Certainty assessment						impact	Certainty	Importance
	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations			
Dysphagia [severe (grade 3 or worse) and non-severe (less than grade 3)] (assessed with: AUROC curve)									
1	observational studies	serious[a]	serious[b]	not serious	not serious	none	Not applicable	⨁⨁◯◯ Low	Not applicable
Radiation-induced hypothyroidism (assessed with: AUROC curve)									
1	observational studies	serious[c]	serious[b]	not serious	not serious	none	Not applicable	⨁⨁◯◯ Low	Not applicable
Xerostomia at 12 months (assessed with: AUROC curve)									
1	observational studies	serious[d]	serious[b]	not serious	not serious	none	Not applicable	⨁⨁◯◯ Low	Not applicable

Explanations

a. Dean et al, 2018 [15] have high RoB for Analysis domain.

b. Regarding Inconsistency: MA demonstrated a high heterogeneity of models with IBM (87.36%) and without IBM (97.89%), which is a reflex of confidence intervals of individual studies varying more than just by chance. This can be explained by expressive variation in: i) the predicted toxicity; ii) the data format (even among those using IBM, the modality of images and features selected are different for each study); iii) the feature selector; and iv) the classification model.

c. Smyczynska et al., 2021 [5] have unclear RoB in two domains (Predictors and Outcomes) and high RoB for Analysis domain.

d. Soares et al, 2017 [16] have unclear RoB in Participants domain and high RoB for Outcomes and Analysis domains.

3 DISCUSSÃO

De modo a ampliar a compreensão de patologistas orais, médicos orais e cirurgiões de cabeça e pescoço sobre as abordagens diagnósticas baseadas em IA, com foco especial nas CNNs, fundamentos teóricos e conceituais foram sumarizados e adaptados para a compreensão do público-alvo. Este artigo conceitual é um marco especial na literatura relacionada à CCP e vem de encontro a necessidade urgente de aproximar profissionais clínicos da área da saúde de conceitos importantes da engenharia biomédica, a fim de facilitar a compreensão das etapas envolvidas no processamento de imagens para fins de desenvolvimento de sistemas de apoio ao diagnóstico e prognóstico clínico e histopatológico. A implementação de sete arquiteturas de Deep Learning (DL) de última geração para gradação de displasia epitelial oral em imagens histopatológicas de lâminas digitalizadas, por sua vez, é um trabalho de originalidade ímpar, sem precedentes na literatura, e vem de encontro a necessidade de expandir abordagens focadas em DL para o diagnóstico de DOPM. Adicionalmente, a condução de uma RS sobre modelos de ML atualmente usados para prever toxicidades relacionadas ao tratamento de CCP de modo a avaliar as evidências sobre o impacto de BMIs em PMs, permite a compreensão de quais metodologias mais apropriadas para treinar PM, bem como a caracterização de banco de dados e personalização individual de tratamento em *point-of-care* que forneça condições reais que culminem em predições realistas, e a compreensão de como a falta d epadronização tanto na execução dos estudos primários quanto o reporte de resultados influencia profundamente na interpretação dos estudos.

A avaliação adequada dos critérios citológicos e arquiteturais visa objetificar a classificação de displasia epitelial oral (DEO). No entanto, mesmo fazendo uso desse sistema, a concordância interobservador pode ficar limitada entre 62% e 90% (Speight et al., 2015; Ranganathan et al., 2019). Estudos assistidos por sistemas digitais demonstram um aumento estatisticamente significativo na densidade de volume nuclear, razão célula-núcleo, área nuclear e perímetro de células displásicas quando comparadas em diferentes graus de displasia, com valores maiores a medida que o grau de displasia aumenta (Prema et al., 2020). Essa premissa corrobora que a forma e o tamanho dos núcleos é um importante atributo para a diferenciação de diferentes graus de displasia. Essas evidências sugerem que há potencial em métodos que apliquem visão computacional para diferenciar graus de displasia de modo a superar a subjetividade envolvida nessas análises. No entanto, anotação das imagens associada à fragmentação em patches pode inserir viés ao fornecer uma grande interseção de características que descrevem ambas as classes abordadas no presente estudo. Nesse cenário,

fragmentos de imagens (*patches)* originário de áreas displásicas na camada basal serão corretamente rotuladas como possuindo alto risco de malignização (do inglês *high risk* (HR)], enquanto *patches* da área mais superior do epitélio podem apresentar alterações insuficientes para se enquadrar na classificação de HR. Em resumo, uma área anotada pode gerar *patches* muito diferentes de acordo com o nível do epitélio e esses patches estariam anotados sob um mesmo rótulo, o que pode gerar confusão do "padrão ouro", confundindo as CNNs, prejudicando o aprendizado e favorecendo a "memorização" e o overfitting.

Os critérios clínicos para descrever e classificar DOPM são amplamente variáveis entre os estudos e a importância de estabelecer uma boa representação da lesão confunde os conceitos de aspecto clínico não homogêneo e maior risco de malignização (van der Waal, 2015). Essa ampla gama de critérios clínicos sofreu ajustes ao longo dos anos e pode justificar, juntamente com as diferentes formações educacionais, fonte de discordância entre os observadores. Esta é uma variação que deve ser levada em consideração ao desenvolver modelos de DL para o diagnóstico de DOPM. Uma limitação identificada no presente estudo foi a dificuldade em delimitar as bordas das lesões, principalmente em áreas desfocadas e locais onde a má vascularização confere um aspecto mais pálido à mucosa (por exemplo, borda lateral da língua quando o clínicos está tracionando para a tomada da foto), o que pode ser um fator de confusão quando a classificação de pixels está em andamento (figura 1). Abordagens baseadas na faixa de valores de pixel geralmente são associadas a descritores de textura para determinar se uma área deve ser segmentada e, dependendo da faixa de valores de pixel pré-determinada, essa abordagem pode excluir áreas importantes da imagem segmentada, exigindo importantes etapas de calibração, que representa um desafio quando se trata de lesões não homogêneas onde a faixa de valores representativos dos pixels deve incluir áreas avermelhadas e esbranquiçadas (figura 2) .

Para a predição de toxicidade em pacientes submetidos à radioterapia e quimioterapia para tratamento de cancer de cabeça e pescoço (CCP), modelos baseados apenas em características dosimétricas usam dose média e informações parciais do histograma dose-volume (DVH), enquanto os modelos convencionais (do inglês *Normal Tissue Complication Probability* (NTCP)] utilizam todas as informações da curva DVH concatenando todas as métricas em um único fator com funções dose-resposta (DRFs). Os modelos de Radiômica, por outro lado, extraem características diretamente de imagens médicas (conhecidas como características de textura ou intensidade). Da mesma forma, a Dosiômica tenta extrair características espaciais 3D da distribuição de doses de radioterapia. Por esta razão, para permitir a devida comparação e análise dos MPs relatados, a presente RS leva em

consideração modelos preditivos (MPs) baseados em biomarcadores de imagem (BMIs) e em dados não-imaginológicos (clínicos e dosimétricos).

Há uma tendência em afirmar que o uso de biomarcadores de imagem (BMIs) melhora o poder de predizer a ocorrência de toxicidades (van Dijk et al., 2017a; van Dijk et al., 2017b; van Dijk et al., 2018a; van Dijk et al., 2018b; Rosen et al., 2018; van Dijk et al., 2019) mas, segundo alguns autores, modelos baseados em BMIs não apresentam superioridade sobre os modelos NTCP (Beetz et al., 2012; Smyczynska et al., 2021). De acordo com a presente RS, nenhuma inferência pode ser feita em relação ao uso de modelos baseados em BMIs terem desempenho melhor do que modelos não baseados em BMIs. Os autores esperavam visualizar as diferenças entre os estudos que não usam dados de imagem e aqueles que os usam, traçando as curvas AUROC, mas não foi observada distinção suficiente entre os valores para nenhuma das modalidades. Em relação aos estudos incluídos na presente meta-análise, há apenas uma melhora marginal observada em PMs baseados em BMIs. Vale ressaltar que apenas três estudos atendem aos critérios para serem incluídos na MA. A alta heterogeneidade dos estudos primários é explicada pelo local do tumor e as modalidades de tratamento serem altamente variáveis, bem como os preditores selecionados e os modelos para prever os desfechos (Debray et al., 2017). Os autores aconselham os leitores a ter uma interpretação crítica dos resultados. Além disso, na pesquisa médica, a quantidade de dados é crucial e, muitas vezes, não é fácil de recuperar em termos de documentação médica completa e acompanhamento. Idealmente, a validação externa ou teste independente deve ser realizado para melhor avaliação da capacidade de generalização dos modelos. Esta etapa final é importante para fornecer uma noção de quão bem o modelo pode executar ao avaliar dados não vistos.

Por fim, a literatura científica sobre IA para imagens médicas é vasta e diversificada, sendo os resultados relatados em métricas e ilustrações gráficas, algumas vezes indicadas para fins pontuais. A *Pattern Recognition Community* (Bradley, 1997) adota verdadeiros positivos (TP), falsos negativos (FN), precisão de classificação (TP+TN)/n e F1-score [2TP/(2TP+FP+FN)] para avaliar o desempenho dos modelos. A área sob a curva da característica de operação do receptor (AUROC) é considerada uma medida mais intuitiva, discriminatória e consistente do que a precisão e mostra a taxa de verdadeiros positivos (TPR) contra a taxa de falsos positivos (FPR) [FP/(FP+TN] para analisar a capacidade de generalização do modelo. Esta métrica fornece uma noção quantitativa de separação de classes, recomendada quando há um problema de classificação binária (Jin et al., 2005). Para realizar uma meta-análise de curva AUROC agrupada, é necessário que os estudos reportem

valores de intervalo de confiança e erro padrão (calculado a partir da raiz quadrada da amostra de teste e desvio padrão) (Borenstein et al., 2009). Foi constatado que a maioria dos estudos não possuem todas as métrica snecessárias para a correta interpretação dos dados dos estudos prímários, inviabilizando a inclusão de estudos importantes na MA.

4 CONCLUSÃO

De acordo com cada estudo desenvolvido, o presente trabalho pode concluir que:

- A compreensão sistêmica dos conceitos de IA e Medicina/Patologia é indispensável para promover comunicação eficiente entre os times multidisciplinares envolvidos.
- Foi possível observar um grande potencial de aprendizado das redes estudadas. No entanto, com base na presente metodologia de anotação das imagens e nos hiperparâmetros utilizados, não foi possível atingir uma boa capacidade de generalização dos modelos para serem aplicados em conjuntos de dados da vida real.
- A percepção dos clínicos pode introduzir viés nas anotações usadas para treinar modelos de DL para detecção e classificação de objetos, especialmente no domínio de lesões brancas e estriadas.
- MPs que utilizam BMIs não são superiores aos que não usam BMAs. No entanto, é importante ressaltar que a presente MA foi realizada entre modelos de apenas três estudos que apresentam alta inconsistência, alto risco de viés e baixa certeza de evidência.

REFERÊNCIAS

Beetz I, Schilstra C, Burlage FR, Koken PW, Doornaert P, Bijl HP, et al. Development of NTCP models for head and neck cancer patients treated with three-dimensional conformal radiotherapy for xerostomia and sticky saliva: The role of dosimetric and clinical factors. Radiotherapy and Oncology 2012;105:86–93. https://doi.org/10.1016/j.radonc.2011.05.010.

Borenstein M, Hedges L, Higgins J, Rothstein H. Introduction to meta-analysis. Chichester, UK: Wiley; 2009.

Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit 1997;30:1145–59. https://doi.org/10.1016/S0031-3203(96)00142-2.

Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. Ann Intern Med 2015;162:55–63. https://doi.org/10.7326/M14-0697.

Debray TPA, Damen JAAG, Snell KIE, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ 2017;356:i6460. https://doi.org/10.1136/bmj.i6460.

Jin Huang, Ling CX. Using AUC and accuracy in evaluating learning algorithms. IEEE Trans Knowl Data Eng 2005;17:299–310. https://doi.org/10.1109/TKDE.2005.50.

Krohn J, Beyleveld G, Bassens A. Deep Learning Illustrated: A Visual, Interactive Guide to Artificial Intelligence. 2019. ISBN 10: 0135121728; 13: 9780135121726.

Marée R. The Need for Careful Data Collection for Pattern Recognition in Digital Pathology. J Pathol Inform. 2017 Apr 10;8:19. doi: 10.4103/jpi.jpi_94_16.

McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. 1943. Bull Math Biol. 1990;52(1-2):99-115; discussion 73-97. PMID: 2185863.

Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev 2015;4:1. https://doi.org/10.1186/2046-4053-4-1.

Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, 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:W1–73. https://doi.org/10.7326/M14-0698.

* De acordo com as normas da UNICAMP/FOP, baseadas na padronização do International Committee of Medical Journal Editors - Vancouver Group. Abreviatura dos periódicos em conformidade com o PubMed.

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:W1. https://doi.org/10.7326/M18-1377.

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:n71. https://doi.org/10.1136/bmj.n71.

Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 2021;372:n160. https://doi.org/10.1136/bmj.n160.

Prema, V., Thomas, T., Harikrishnan, P., Viswanathan, M., Srichinthu, K. & Rajkumar, K. Morphometric analysis of suprabasal cell layer in oral epithelial dysplasia: A computer-assisted microscopic study. J Pharm Bioallied Sci 12, 204 (2020).

Ranganathan, K. & Kavitha, L. Oral epithelial dysplasia: Classifications and clinical relevance in risk assessment of oral potentially malignant disorders. J Oral Maxillofac Pathol 23, 19–27 (2019).

Rosen BS, Hawkins PG, Polan DF, Balter JM, Brock KK, Kamp JD, et al. Early Changes in Serial CBCT-Measured Parotid Gland Biomarkers Predict Chronic Xerostomia After Head and Neck Radiation Therapy. International Journal of Radiation Oncology*Biology*Physics 2018;102:1319–29. https://doi.org/10.1016/j.ijrobp.2018.06.048.

Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev. 1958 Nov;65(6):386-408. doi: 10.1037/h0042519.

Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. BMJ 2015;349:g7647–g7647. https://doi.org/10.1136/bmj.g7647.

Smyczynska U, Grabia S, Nowicka Z, Papis-Ubych A, Bibik R, Latusek T, et al. Prediction of Radiation-Induced Hypothyroidism Using Radiomic Data Analysis Does Not Show Superiority over Standard Normal Tissue Complication Models. Cancers (Basel) 2021;13:5584. https://doi.org/10.3390/cancers13215584.

Speight, P. M., Abram, T. J., Floriano, P. N., James, R., Vick, J., Thornhill, M. H. et al. Interobserver agreement in dysplasia grading: toward an enhanced gold standard for clinical pathology trials. Oral Surg Oral Med Oral Pathol Oral Radiol 120, 474-482.e2 (2015).

van der Waal I. Oral leukoplakia, the ongoing discussion on definition and terminology. Med Oral Patol Oral Cir Bucal. 2015 Nov 1;20(6):e685-92. doi: 10.4317/medoral.21007.

van Dijk L v., Brouwer CL, van der Laan HP, Burgerhof JGM, Langendijk JA, Steenbakkers RJHM, et al. Geometric Image Biomarker Changes of the Parotid Gland Are Associated With Late Xerostomia. International Journal of Radiation Oncology*Biology*Physics 2017;99:1101–10. https://doi.org/10.1016/j.ijrobp.2017.08.003.

van Dijk L v., Brouwer CL, van der Schaaf A, Burgerhof JGM, Beukinga RJ, Langendijk JA, et al. CT image biomarkers to improve patient-specific prediction of radiation-induced xerostomia and sticky saliva. Radiotherapy and Oncology 2017;122:185–91. https://doi.org/10.1016/j.radonc.2016.07.007.

van Dijk L v., Langendijk JA, Zhai T-T, Vedelaar TA, Noordzij W, Steenbakkers RJHM, et al. Delta-radiomics features during radiotherapy improve the prediction of late xerostomia. Sci Rep 2019;9:12483. https://doi.org/10.1038/s41598-019-48184-3.

van Dijk L v., Noordzij W, Brouwer CL, Boellaard R, Burgerhof JGM, Langendijk JA, et al. 18F-FDG PET image biomarkers improve prediction of late radiation-induced xerostomia. Radiotherapy and Oncology 2018;126:89–95. https://doi.org/10.1016/j.radonc.2017.08.024.

van Dijk L v., Thor M, Steenbakkers RJHM, Apte A, Zhai T-T, Borra R, et al. Parotid gland fat related Magnetic Resonance image biomarkers improve prediction of late radiation-induced xerostomia. Radiotherapy and Oncology 2018;128:459–66. https://doi.org/10.1016/j.radonc.2018.06.012.

Zhang A, Lipton ZC, Li M, Smola AJ. Dive into Deep Learning. 2021 doi: doi.org/10.48550/arXiv.2106.11342

ANEXOS

Anexo 1 - Comitê de Ética em Pesquisa

UNICAMP - FACULDADE DE ODONTOLOGIA DE PIRACICABA DA UNIVERSIDADE DE CAMPINAS - FOP/UNICAMP

PARECER CONSUBSTANCIADO DO CEP

DADOS DA EMENDA

Título da Pesquisa: INTELIGÊNCIA ARTIFICIAL NO DIAGNÓSTICO E CLASSIFICAÇÃO DE DISPLASIAS EPITELIAIS ORAIS

Pesquisador: ANNA LUÍZA DAMACENO ARAÚJO

Área Temática:

Versão: 7

CAAE: 42235421.9.0000.5418

Instituição Proponente: Faculdade de Odontologia de Piracicaba - Unicamp

Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 5.603.372

Apresentação do Projeto:

O parecer inicial é elaborado com base na transcrição editada do conteúdo do registro do protocolo na Plataforma Brasil e dos arquivos anexados à Plataforma Brasil. Os pareceres de retorno, emendas e notificações são elaborados a partir dos dados e arquivos da última versão apresentada.

Trata-se de SOLICITAÇÃO DE EMENDA (E1) AO PROTOCOLO originalmente aprovado em 25/09/2021 para alterações nos objetivos, na amostra, na metodologia e no cronograma de execução da pesquisa. O parecer foi atualizado de acordo com a documentação apresentada. A solicitação está detalhadamente descrita ao final do parecer.

A EQUIPE DE PESQUISA citada na capa do projeto de pesquisa inclui ANNA LUÍZA DAMACENO ARAÚJO (Cirurgiã-dentista, Doutoranda no PPG em Estomatopatologia da FOP-UNICAMP, pesquisadora responsável), PABLO AGUSTIN VARGAS (Cirurgião-dentista, Docente da área de Patologia da FOP-UNICAMP) e ALAN ROGER DOS SANTOS SILVA (Cirurgião-dentista, Docente da área de Semiologia da FOP-UNICAMP), o que é confirmado na declaração dos pesquisadores e na PB.

Delineamento da pesquisa: Trata-se de estudo clínico-laboratorial, parcialmente in silico,

Endereço: Av.Limeira 901 Caixa Postal 52
Bairro: Areião **CEP:** 13.414-903
UF: SP **Município:** PIRACICABA
Telefone: (19)2106-5349 **Fax:** (19)2106-5349 **E-mail:** cep@fop.unicamp.br

Declaração de Manuseio Material Biológico / Biorepositório / Biobanco	Biorrep.pdf	19/02/2021 16:15:48	ANNA LUÍZA DAMACENO ARAÚJO	Aceito
Declaração de Instituição e Infraestrutura	DecInst.pdf	19/02/2021 16:15:21	ANNA LUÍZA DAMACENO ARAÚJO	Aceito
Declaração de Pesquisadores	DecPesq.pdf	19/02/2021 16:15:06	ANNA LUÍZA DAMACENO ARAÚJO	Aceito

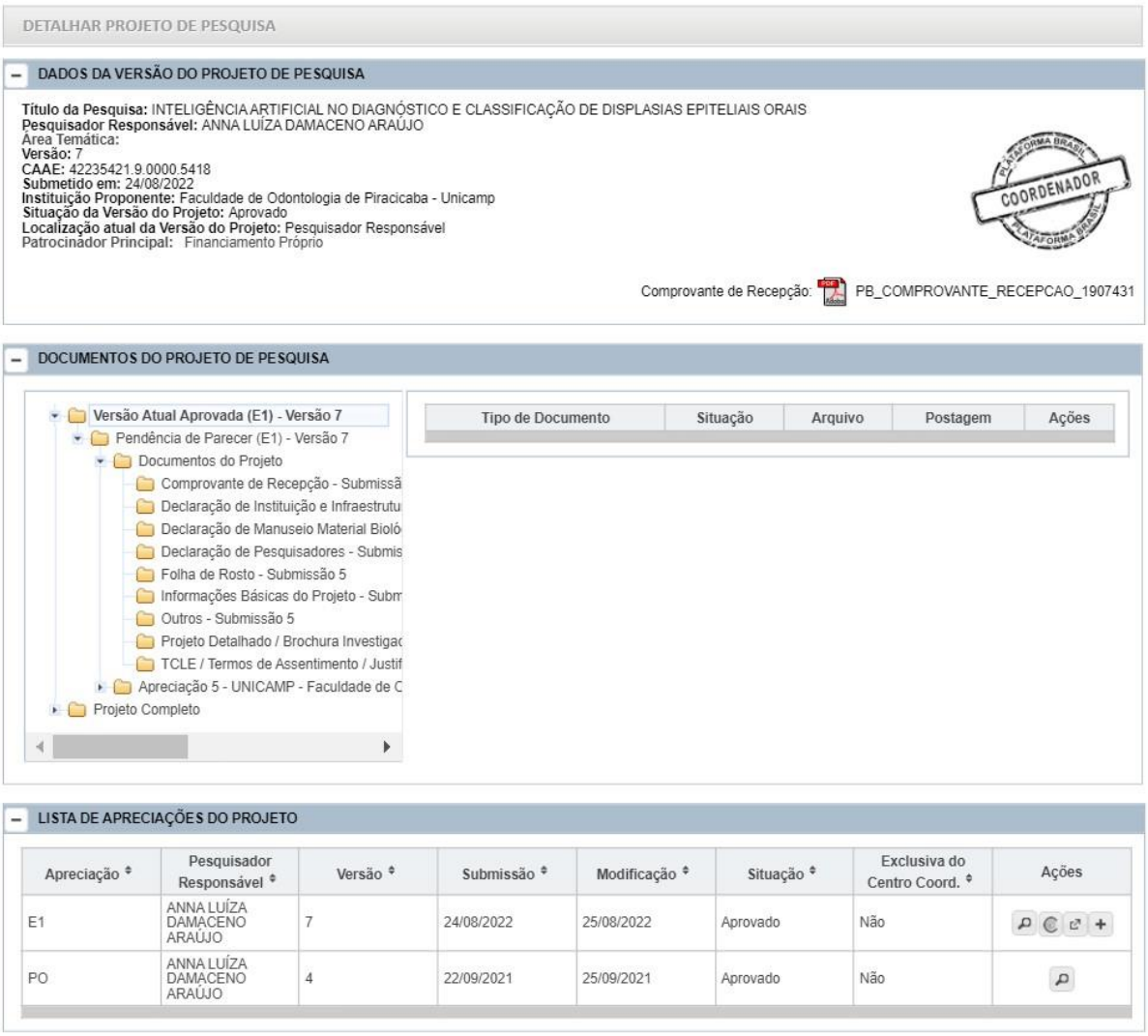
Situação do Parecer:
Aprovado

Necessita Apreciação da CONEP:
Não

PIRACICABA, 25 de Agosto de 2022

Assinado por:
jacks jorge junior
(Coordenador(a))

Anexo 2 - Situação do Projeto na Plataforma Brasil (print)



DETALHAR PROJETO DE PESQUISA

DADOS DA VERSÃO DO PROJETO DE PESQUISA

Título da Pesquisa: INTELIGÊNCIA ARTIFICIAL NO DIAGNÓSTICO E CLASSIFICAÇÃO DE DISPLASIAS EPITELIAIS ORAIS
Pesquisador Responsável: ANNA LUÍZA DAMACENO ARAÚJO
Área Temática:
Versão: 7
CAAE: 42235421.9.0000.5418
Submetido em: 24/08/2022
Instituição Proponente: Faculdade de Odontologia de Piracicaba - Unicamp
Situação da Versão do Projeto: Aprovado
Localização atual da Versão do Projeto: Pesquisador Responsável
Patrocinador Principal: Financiamento Próprio

COORDENADOR

Comprovante de Recepção: PB_COMPROVANTE_RECEPCAO_1907431

DOCUMENTOS DO PROJETO DE PESQUISA

- Versão Atual Aprovada (E1) - Versão 7
  - Pendência de Parecer (E1) - Versão 7
    - Documentos do Projeto
      - Comprovante de Recepção - Submissã
      - Declaração de Instituição e Infraestrutu
      - Declaração de Manuseio Material Bioló
      - Declaração de Pesquisadores - Submis
      - Folha de Rosto - Submissão 5
      - Informações Básicas do Projeto - Subm
      - Outros - Submissão 5
      - Projeto Detalhado / Brochura Investigac
      - TCLE / Termos de Assentimento / Justif
    - Apreciação 5 - UNICAMP - Faculdade de C
  - Projeto Completo

Tipo de Documento	Situação	Arquivo	Postagem	Ações

LISTA DE APRECIAÇÕES DO PROJETO

Apreciação	Pesquisador Responsável	Versão	Submissão	Modificação	Situação	Exclusiva do Centro Coord.	Ações
E1	ANNA LUÍZA DAMACENO ARAÚJO	7	24/08/2022	25/08/2022	Aprovado	Não	
PO	ANNA LUÍZA DAMACENO ARAÚJO	4	22/09/2021	25/09/2021	Aprovado	Não	

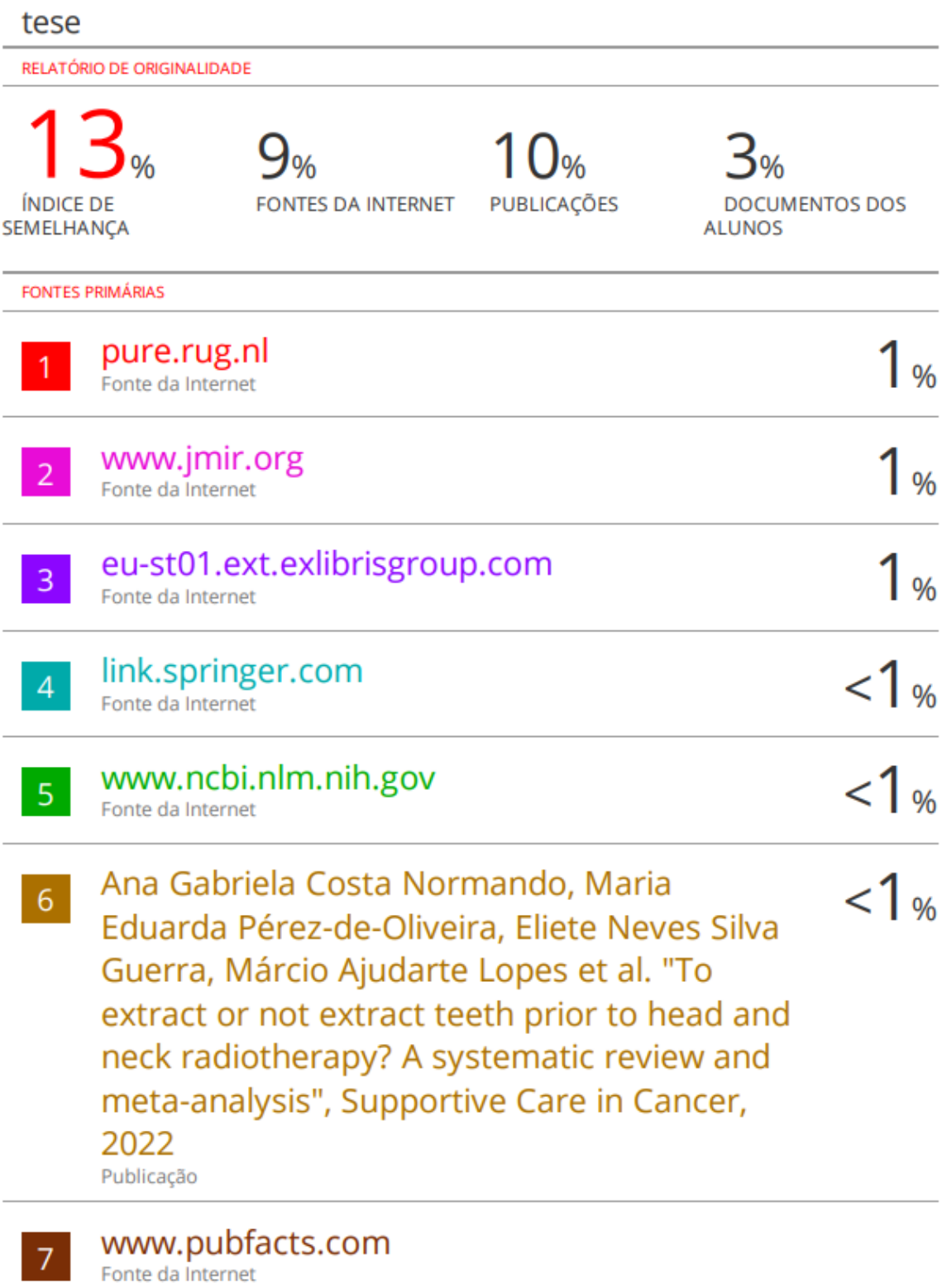
Anexo 3 - Documento de aceite do artigo (print do sistema online de submissão)

Journal of Oral Pathology & Medicine
Review

Machine Learning Concepts applied to Oral Pathology and Oral Medicine: A Convolutional Neural Networks Approach

Submission Status	Submitted	This submission has been sent to the editorial office and cannot be edited. Further instructions will be emailed to you from Manuscript Central.
Submitted On	8 September 2022 by Alan Santos-Silva	
Submission Started	7 September 2022 by Alan Santos-Silva	View Submission Overview

Anexo 4 - Relatório de similaridade da Plataforma Turnitin



tese

RELATÓRIO DE ORIGINALIDADE

13%	9%	10%	3%
ÍNDICE DE SEMELHANÇA	FONTES DA INTERNET	PUBLICAÇÕES	DOCUMENTOS DOS ALUNOS

FONTES PRIMÁRIAS

1	pure.rug.nl Fonte da Internet	1%
2	www.jmir.org Fonte da Internet	1%
3	eu-st01.ext.exlibrisgroup.com Fonte da Internet	1%
4	link.springer.com Fonte da Internet	<1%
5	www.ncbi.nlm.nih.gov Fonte da Internet	<1%
6	Ana Gabriela Costa Normando, Maria Eduarda Pérez-de-Oliveira, Eliete Neves Silva Guerra, Márcio Ajudarte Lopes et al. "To extract or not extract teeth prior to head and neck radiotherapy? A systematic review and meta-analysis", Supportive Care in Cancer, 2022 Publicação	<1%
7	www.pubfacts.com Fonte da Internet	