UNIVERSIDADE DE MOGI DAS CRUZES
LETICIA BRUM CROFFI

VISÃO COMPUTACIONAL PARA DETECÇÃO DE CÂNCER EM IMAGENS DE CITOLOGIA CERVICAL

Mogi das Cruzes, SP
2025

UNIVERSIDADE DE MOGI DAS CRUZES

LETICIA BRUM CROFFI

VISÃO COMPUTACIONAL PARA DETECÇÃO DE CÂNCER EM IMAGENS DE CITOLOGIA CERVICAL

Dissertação apresentada ao Programa de Pós-Graduação da Universidade de Mogi das Cruzes como parte dos requisitos para obtenção do grau de Mestre em Ciência e Tecnologia em Saúde.
Área de concentração: Pesquisa, Desenvolvimento e Inovação em Saúde.

Orientadora: Dra. Silvia Cristina Martini
Coorientador: Dr. Wagner Alves de Souza Júdice

Mogi das Cruzes, SP

2025

Ficha catalográfica
Universidade de Mogi das Cruzes - Biblioteca Central

Croffi, Leticia Brum
C941v Visão computacional para detecção de câncer em imagens de citologia cervical / Leticia Brum Croffi -- Mogi das Cruzes, SP: Universidade de Mogi das Cruzes, 2025.
119 f. : il.

Dissertação (Mestrado em Ciência e Tecnologia em Saúde) -- Universidade de Mogi das Cruzes, 2025.
Orientação: Profa. Dra. Silvia Cristina Martini.
Inclui bibliografia.

1. Neoplasias/prevenção & controle 2. Redes neurais de Computação 3. Teste de Papanicolaou I. Martini, Silvia Cristina II. Universidade de Mogi das Cruzes III. Título.

CDD 616.07582

Elaborada por Maria Cristina Alves Silva Xavier – CRB-8/10077

UMC
UNIVERSIDADE

Tel.: (11) 4798-7000
0800 19 2001
www.umc.br

ATAS

ATA DA SESSÃO PÚBLICA DE APRESENTAÇÃO DE DISSERTAÇÃO DE MESTRADO PROFISSIONAL EM CIÊNCIA E TECNOLOGIA EM SAÚDE DA UNIVERSIDADE DE MOGI DAS CRUZES

Às quatorze horas do dia vinte e sete de janeiro de dois mil e vinte e cinco, pela Universidade de Mogi das Cruzes, realizou-se a defesa de dissertação intitulada: "VISÃO COMPUTACIONAL PARA DETECÇÃO DE CÂNCER EM IMAGENS DE CITOLOGIA CERVICAL" pelo(a) candidato(a) **Leticia Brum Croffi,** para obtenção do grau de mestre profissional em ciência e tecnologia em saúde. Tendo sido o número de créditos alcançados pelo(a) mesmo(a) no total de 44 (quarenta e quatro), a saber: 24 unidades de crédito em disciplinas de pós-graduação e 20 unidades de crédito no preparo da dissertação, o(a) aluno(a) perfaz assim os requisitos para obtenção do grau de mestre profissional. A comissão examinadora estava constituída dos senhores professores: Silvia Cristina Martini, e Marcia Aparecida Silva Bissaco, da Universidade de Mogi das Cruzes, e Marcelo Zanchetta do Nascimento, da Universidade Federal de Uberlândia, sob a presidência da primeira, como orientadora da dissertação. A sessão pública da defesa foi aberta pela senhora presidenta da comissão, tendo em seguida o(a) candidato(a) realizado a apresentação da dissertação. Concluída a apresentação, tiveram início as arguições pelos membros da comissão examinadora. A seguir a comissão, em sessão secreta, conforme julgamento discriminado por cada membro, considerou o(a) candidato(a)

Aprovada (aprovado(a)/reprovado(a)) por Unanimidade (unanimidade/maioria)

Mogi das Cruzes, 27 de janeiro de 2025.

Comissão Examinadora	Julgamento
Documento assinado digitalmente gov.br MARCELO ZANCHETTA DO NASCIMENTO Data: 27/01/2025 16:35:35-0300 Verifique em https://validar.iti.gov.br Prof. Dr. Marcelo Zanchetta do Nascimento	Aprovada (aprovado(a)/reprovado(a))
Profª Drª Marcia Aparecida Silva Bissaco	Aprovada (aprovado(a)/reprovado(a))
Profª Drª Silvia Cristina Martini	Aprovada (aprovado(a)/reprovado(a))

Scanned with CamScanner

AGRADECIMENTOS

Agradeço à professora Doutora Silvia Cristina Martini pela orientação do trabalho. Agradeço à professora Doutora Silvia Regina Matos da Silva Boschi por ter aceitado a minha inscrição e me encorajado a me matricular mesmo eu morando em outra cidade e sendo da área de Ciência da Computação e não da área médica.

Agradeço especialmente ao meu marido por todo apoio e incentivo, sempre. À minha mãe e ao meu padrasto, que são médicos, e ao meu pai, todos eles referências de vida para mim.

RESUMO

Este estudo investigou o uso de redes neurais convolucionais (CNNs) e Vision Transformers (ViTs) na classificação de imagens de citologia cervical, abrangendo células normais e anormais. O objetivo principal era desenvolver um método de processamento de imagens de citologia cervical para detecção de câncer cervical em imagens de citologia cervical. A citologia cervical apresenta desafios significativos devido à alta variabilidade das amostras e à necessidade de diagnósticos precisos para evitar falsos negativos e positivos. Ao revisar estudos e modelos já utilizados, foi proposto um experimento de combinar as capacidades das CNNs para capturar características locais e dos ViTs para identificar características, buscando melhorar o desempenho em cenários ainda não testados. Foram utilizados o modelo EfficientNet-b4 (CNN) e o modelo DeiT3-base (ViT), avaliados individualmente e combinados em paralelo com a soma de características extraídas. Os dados consistiram em imagens de conjuntos públicos e de fontes externas, simulando condições próximas à prática clínica. A análise de explicação foi realizada por meio do Grad-CAM, visando garantir maior transparência nos processos de classificação. A interpretabilidade dos modelos é fundamental para sua aplicação na área da saúde, pois permite avaliar se as decisões do modelo estão baseadas em características relevantes ou se são influenciadas por ruídos. Os resultados mostraram que o modelo DeiT3-base apresentou a melhor performance, com sensibilidade de 79,2%, especificidade de 74,7% e AUC de 0,77. O modelo EfficientNet-b4 alcançou sensibilidade de 76,6%, especificidade de 41,7% e AUC de 0,59, enquanto a combinação dos modelos não resultou em melhorias significativas, apresentando sensibilidade de 50,6%, especificidade de 73,0% e AUC de 0,62. Apesar do desempenho promissor do DeiT3, a elevada ocorrência de falsos negativos no conjunto de teste evidencia a necessidade de avanços na adaptação de modelos a imagens sem filtragem. Como sugestões futuras, destaca-se a exploração de outras arquiteturas CNN, técnicas mais eficazes de fusão e abordagens robustas para melhorar a qualidade dos dados de entrada. Este estudo contribui para o avanço do uso de inteligência artificial em aplicações clínicas, enfatizando a importância da explicação e da confiabilidade em sistemas de apoio ao diagnóstico.

Palavras-chave: câncer cervical; diagnóstico por imagem; aprendizado profundo; transformers; interdisciplinar.

ABSTRACT

This study investigated the use of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) in classifying cervical cytology images, covering both normal and abnormal cells. The main objective was to develop an image processing method for cervical cytology to detect cervical cancer in cytology images. Cervical cytology presents significant challenges due to the high variability of samples and the need for precise diagnoses to avoid false negatives and positives. By reviewing previously used studies and models, an experiment was proposed to combine the capabilities of CNNs to capture local features and ViTs to identify global features, aiming to improve performance in untested scenarios. The EfficientNet-b4 model (CNN) and the DeiT3-base model (ViT) were used, evaluated individually and combined in parallel with the sum of extracted features. The data consisted of images from public datasets and external sources, simulating conditions close to clinical practice. The explanation analysis was performed using Grad-CAM to ensure greater transparency in the classification processes. Model interpretability is essential for its application in healthcare, as it allows assessing whether the model's decisions are based on relevant features or influenced by noise. The results showed that the DeiT3-base model achieved the best performance, with a sensitivity of 79.2%, specificity of 74.7%, and an AUC of 0.77. The EfficientNet-b4 model reached a sensitivity of 76.6%, specificity of 41.7%, and an AUC of 0.59, while the combination of models did not result in significant improvements, presenting a sensitivity of 50.6%, specificity of 73.0%, and an AUC of 0.62. Despite the promising performance of DeiT3, the high occurrence of false negatives in the test set highlights the need for advancements in adapting models to unfiltered images. As future suggestions, the exploration of other CNN architectures, more effective fusion techniques, and robust approaches to improve input data quality are emphasized. This study contributes to the advancement of artificial intelligence in clinical applications, highlighting the importance of explainability and reliability in diagnostic support systems.

Keywords: cervical cancer; diagnostic image; deep learning; transformers; interdisciplinar.

LISTA DE ILUSTRAÇÕES

Figura 1 - A Evolução natural do câncer por HPV 19
Figura 2 - Evolução de Lesões Pré-Cancerosas a Carcinoma Escamoso Invasivo .. 20
Figura 3 - Evolução de Lesões Pré-cancerosas sob a perspectiva da Citologia 21
Figura 4 - Teste convencional de Papanicolau 22
Figura 5 - Citologia em Base Líquida 22
Figura 6 - Algoritmo Tradicional 28
Figura 7 - Algoritmo de Aprendizado de Máquina 28
Figura 8 - Arquitetura da Rede Neural Convolucional (CNN) 32
Figura 9 - CNNs x ViTs 35
Figura 10 - Imagens e seus respectivos mapas de atenção 37
Figura 11 - Arquitetura do modelo de Vision Transformer (ViT) 37
Figura 12 - ViT: Primeiro passo 38
Figura 13 - ViT: Segundo passo 39
Figura 14 - ViT: Último passo 42
Figura 15 - Fusão dos Modelos: CNN + ViT 50
Figura 16 - *Data Augmentation* 53
Figura 17 - CNN: Erro por época 54
Figura 18 - CNN: Acurácia por época 55
Figura 19 - CNN Validação: Matriz de Confusão 56
Figura 20 - CNN Validação: Área sob a curva ROC (AUC) 57
Figura 21 - CNN Teste: Matriz de Confusão 58
Figura 22 - CNN Teste: Área sob a curva ROC (AUC) 59
Figura 23 - Classificações de Validação incorretas: Falsos Negativos (Total: 54) 60
Figura 24 - Classificações de Validação incorretas: Falsos Positivos (Total: 175) 61
Figura 25 - Classificações de Validação corretas: Verdadeiros Negativos (Total: 125) 62
Figura 26 - Classificações de Validação corretas: Verdadeiros Positivos (Total: 177) 62
Figura 27 - Classificações de Teste incorretas: Falsos Negativos (Total: 139) 63
Figura 28 - Classificações de Teste incorretas: Falsos Positivos (Total: 61) 64
Figura 29 - Classificações de Teste corretas: Verdadeiros Negativos (Total: 62) 64

Figura 30 - Classificações de Teste corretas: Verdadeiros Positivos (Total: 126)..... 65
Figura 31 - ViT: Erro por época 66
Figura 32 - ViT: Acurácia por época 66
Figura 33 - ViT Validação: Matriz de Confusão 67
Figura 34 - ViT Validação: Área sob a curva ROC (AUC) 69
Figura 35 - ViT Teste: Matriz de Confusão 69
Figura 36 - ViT Teste: Área sob a curva ROC (AUC) 71
Figura 37 - Classificações de Validação incorretas: Falsos Negativos (Total: 48)..... 72
Figura 38 - Classificações de Validação incorretas: Falsos Positivos (Total: 76) 72
Figura 39 - Classificações de Validação corretas: Verdadeiros Negativos (Total: 224) 73
Figura 40 - Classificações de Validação corretas: Verdadeiros Positivos (Total: 183) 73
Figura 41 - Classificações de Teste incorretas: Falsos Negativos (Total: 61)..... 74
Figura 42 - Classificações de Teste incorretas: Falsos Positivos (Total: 67) 75
Figura 43 - Classificações de Teste corretas: Verdadeiros Negativos (Total: 56)..... 75
Figura 44 - Classificações de Teste corretas: Verdadeiros Positivos (Total: 204)..... 76
Figura 45 - Primeira configuração testada do modelo de fusão 77
Figura 46 - Modelo de Fusão 1: Erro por época 77
Figura 47 - Modelo de Fusão 1: Acurácia por época 78
Figura 48 - Segunda configuração testada do modelo de fusão 79
Figura 49 - Modelo de Fusão 2: Acurácia por época 79
Figura 50 - Modelo de Fusão 2: Erro por época 80
Figura 51 - Fusão Validação: Matriz de Confusão 81
Figura 52 - Fusão Validação: Área sob a curva ROC (AUC) 82
Figura 53 - Fusão Teste: Matriz de Confusão 83
Figura 54 - Fusão Teste: Área sob a curva ROC (AUC) 84
Figura 55 - Classificações de Validação incorretas: Falsos Negativos (Total: 114) ... 85
Figura 56 - Classificações de Validação incorretas: Falsos Positivos (Total: 81) 86
Figura 57 - Classificações de Validação corretas: Verdadeiros Negativos (Total: 219) 87
Figura 58 - Classificações de Validação corretas: Verdadeiros Positivos (Total: 117) 87

Figura 59 - Classificações de Teste incorretas: Falsos Negativos (Total: 106)..... 88
Figura 60 - Classificações de Teste incorretas: Falsos Positivos (Total: 49) 88
Figura 61 - Classificações de Teste corretas: Verdadeiros Negativos (Total: 74)..... 89
Figura 62 - Classificações de Teste corretas: Verdadeiros Positivos (Total: 159)..... 90
Figura 63 - Grad-CAM: CNN, ViT e Fusão 93

LISTA DE TABELAS

Tabela 1 - CNN Validação: Métricas por classe 56
Tabela 2 - CNN Validação: Métricas 57
Tabela 3 - CNN Teste: Métricas por classe 58
Tabela 4 - CNN Teste: Métricas 59
Tabela 5 - ViT Validação: Métricas por classe 68
Tabela 6 - ViT Validação: Métricas 68
Tabela 7 - ViT Teste: Métricas por classe 70
Tabela 8 - ViT Teste: Métricas 70
Tabela 9 - Fusão Validação: Métricas por classe 81
Tabela 10 - Fusão Validação: Métricas 82
Tabela 11 - Fusão Teste: Métricas por classe 83
Tabela 12 - Fusão Teste: Métricas 84
Tabela 13 - Comparação dos 3 modelos no conjunto de Validação 90
Tabela 14 - Comparação dos 3 modelos no conjunto de Teste 91

LISTA DE ABREVIATURAS E SIGLAS

FCN	Fully Convolutional Network
CRF	Conditional Random Fields
CT	Computed Tomography
MRI	Magnetic Resonance Imaging
PET	Positron Emission Tomography
SPECT	Single-Photon Emission Computed Tomography
fMRI	Functional Magnetic Resonance Imaging
CTC	Circulating Tumor Cells
CNN	Convolutional Neural Network
DNN	Deep Neural Network
RGB	Red Green Blue
R-CNN	Region Based Convolutional Neural Network
INCA	Instituto Nacional de Câncer
NIC	Neoplasia Intraepitelial Cervical
HPV	Human Papiloma Virus (Papiloma Virus Humano)
CPU	Central Process Unit (Unidade Central de Processamento)
GPU	Graphics Processing Unit (Unidade de Processamento Gráfico)
NiN	Network in Network
VGG	Visual Geometry Group
MLP	Multilayer Perceptron
MHA	Multi-head Attention
SVM	Support Vector Machine
ReLU	Rectified Linear Unit
ViT	Vision Transformer
NLP	Natural Language Processing
AUC	Area Under the Curve
ROC	Receiver Operating Characteristic
GRAD-CAM	Gradient-weighted Class Activation Mapping

SUMÁRIO

1 INTRODUÇÃO 13
1.1 Justificativa 15
1.2 Motivação 16
1.3 Objetivos 17
1.3.1 Objetivo Geral 17
1.3.2 Objetivos Específicos 17
2 REVISÃO BIBLIOGRÁFICA 18
2.1 Câncer Cervical 18
2.1.1 Teste de Papanicolau 21
2.1.2 Imagem Médica 24
2.2 Inteligência Artificial 27
2.2.1 Visão Computacional 29
2.2.2 Aprendizado Profundo 30
2.2.3 Redes Neurais Convolucionais 31
2.2.4 Vision Transformer 34
2.2.5 Explicabilidade para modelos de IA 42
3 MATERIAL E MÉTODOS 44
3.1 Ferramentas e Tecnologias 44
3.2 Conjuntos de Imagens 45
3.3 Etapa 1 – Definição de projeto 47
3.4 Etapa 2 – Classificação de imagens 47
3.5 Etapa 3 – CNN 48
3.6 Etapa 4 – ViT 49
3.6 Etapa 5 – CNN + ViT 49
3.7 Etapa 6 – Métricas 51

3.8 Etapa 7 – Explicabilidade 52
4 RESULTADOS 53
4.1 Modelo CNN 53
4.1.1 Modelo CNN: Métricas de Avaliação 55
4.1.2 Modelo CNN: Explicabilidade 60
4.2 Modelo ViT 65
4.2.1 Modelo ViT: Métricas de Avaliação 67
4.2.2 Modelo ViT: Explicabilidade 71
4.3 Modelo de Fusão 76
4.3.1 Modelo de Fusão: Métricas de Avaliação 80
4.3.2 Modelo de Fusão: Explicabilidade 85
4.4 Comparação dos Modelos 90
4.5 Artigos 94
5 DISCUSSÃO 95
6 CONCLUSÃO 98
REFERÊNCIAS 99
APÊNDICE A - Publicações da Dissertação 106
ANEXO A - Registro de software 117

1 INTRODUÇÃO

Nos últimos 60 anos, melhores condições de saneamento e o desenvolvimento de vacinas e antibióticos trouxeram grande declínio na mortalidade por doenças infecciosas. Com a melhoria na prevenção primária e secundária de doenças cardiovasculares, mudanças demográficas e fatores de risco levaram à observação atual de que o câncer é a primeira ou segunda principal causa de morte prematura (ou seja, em pessoas entre 30 a 69 anos de idade) em 134 de 183 países, e ocupa o terceiro ou quarto lugar em outros 45 países. No Brasil, o câncer é a segunda principal causa de morte prematura, depois das doenças cardiovasculares (WILD; WEIDERPASS; STEWART, 2020).

O câncer cervical é o quarto câncer mais comum em termos de incidência e mortalidade em mulheres, com uma estimativa de 660.000 novos casos e 350.000 mortes em todo o mundo em 2022. A doença é o tipo de câncer mais comum em 25 países e a principal causa de morte por câncer em 37 países, principalmente na África subsaariana, bem como na América do Sul e no Sudeste Asiático (BRAY *et al.*, 2022).

A incidência e as taxas de mortalidade por câncer cervical diminuíram na maioria dos países nas últimas décadas, como resultado da detecção de lesões pré-cancerosas por meio de exames de citologia cervical, o que mostra a importância desse exame, além da vacina para HPV que também diminuiria consideravelmente a incidência de novos casos (WILD; WEIDERPASS; STEWART, 2020).

O exame de citologia cervical, ou teste de Papanicolau, é um método muito utilizado para o diagnóstico de malignidades cervicais (KESSLER, 2017), onde especialistas coletam células do colo do útero de uma paciente e as examinam em lâminas de vidro para avaliar, através de um microscópio (GENÇTAV; AKSOY; ÖNDER, 2012) se há alguma malignidade. A triagem por Papanicolau é significativa para ajudar as mulheres a prevenir o câncer de colo de útero. No entanto, tal triagem manual é trabalhosa e subjetiva para os citologistas, considerando que há milhares de células em cada lâmina. Como resultado, é necessário aplicar sistemas de computador para auxiliar o diagnóstico de câncer cervical. O uso de sistemas de computador pode tornar a inspeção citológica cervical mais rápida e precisa.

Tradicionalmente, métodos de Aprendizado de Máquina (*Machine Learning*) foram aplicados na detecção precoce do câncer cervical, mas a extração manual de características e a segmentação inadequada podem causar resultados insatisfatórios de classificação.

A tecnologia de Visão Computacional de Aprendizado Profundo (*Deep Learning*) foi muito utilizada na classificação de imagens de células cervicais alcançado grande sucesso, mas extrair características discriminativas ainda era um desafio para tarefas de classificação. Embora tanto as características locais quanto as globais sejam importantes para a classificação de imagens, muitos métodos existentes prestam atenção ou nas características locais ou nas globais. A proposta desse experimento leva em consideração a sinergia das informações locais e globais das imagens citológicas, para obter características discriminativas em tarefas de classificação.

As Redes Neurais Convolucionais (*Convolutional Neural Networks* - CNNs) se tornaram uma ferramenta importante nas aplicações de Aprendizado Profundo para auxiliar na detecção de malignidades em imagens médicas de citologia cervical (WILLIAM *et al.*, 2018). Os *Vision Transformers* (ViT) então propostos recentemente alcançaram resultados comparáveis aos das CNNs em classificação de imagens, detecção de objetos, segmentação de imagens e processamento de imagens (DOSOVITSKIY *et al.*, 2021). Além disso, o ViT pode aprender dependências de escopo global, o que torna os *Transformers* uma ferramenta alternativa atraente às CNNs. Com o sucesso significativo dos *Vision Transformers* em várias tarefas de Visão Computacional (HAN *et al.*, 2023), o ViT foi considerado como objeto de estudo nesse projeto com várias técnicas de *data augmentation* para melhorar o desempenho da classificação de câncer cervical e aliviar problemas de baixa qualidade nas imagens. A aprendizagem de dependência de escopo global com *Transformers* combinadas a rede neural convolucional poderia ajudar a focar na região do colo do útero em específico sem a necessidade de localização utilizando uma rede de detecção e recorte de imagem. Assim resolvendo a questão da falta de sinergia das informações locais e globais que as redes neurais convolucionais aplicadas sozinhas apresentavam.

1.1 Justificativa

O exame de citologia cervical resultou em uma redução substancial tanto na morbidade quanto na mortalidade do câncer cervical. No entanto, nos últimos anos, a sensibilidade e especificidade deste teste têm sido questionadas, especialmente no que diz respeito aos resultados falso-negativos, nos quais a doença não foi detectada, bem como as taxas de falso-positivo que resultam da interpretação equivocada de condições benignas. Tanto os testes falso-negativos quanto os falso-positivos têm implicações médicas, financeiras, psicológicas e ocasionalmente legais. Uma baixa sensibilidade de exames de citologia cervical colocaria as mulheres em risco de desenvolver uma doença invasiva. Há também preocupações sobre a reprodutibilidade do teste, pois há uma variação considerável na interpretação de esfregaços entre citopatologistas. Os resultados experimentais obtidos nos estudos demonstram variações notáveis em sensibilidade e especificidade (KOLIOPOULOS *et al.*, 2007).

O câncer cervical pode ser prevenido se exames de citologia cervical forem realizados rotineiramente, podendo levar à redução da morbidade e mortalidade. O sucesso do rastreamento tem sido relatado como dependente de vários fatores, incluindo o acesso às instalações, a qualidade dos testes de rastreamento, a adequação do acompanhamento, diagnóstico e tratamento das lesões detectadas. Os exames de citologia cervical são mais ineficazes em países de baixa e média renda devido à falta de trabalhadores da saúde treinados e qualificados, e à falta de recursos de saúde para apoiar programas de rastreamento. A incidência de câncer cervical pode ser reduzida por meio de exames regulares com a aplicação do teste de Papanicolau. No entanto, a análise manual das imagens do teste de Papanicolau é demorada, trabalhosa e sujeita a erros, pois centenas de células dentro de um único slide têm que ser examinadas sob um microscópio por um citopatologista treinado para cada paciente durante o rastreamento. A classificação visual humana de imagens microscópicas tende a ser subjetiva e inconsistente. Por conta disso, houve inúmeras tentativas de automatizar a análise do teste de Papanicolau desde sua introdução há mais de 70 anos (WILLIAM *et al.*, 2019).

Além disso, a implementação de modelos de Aprendizado Profundo para a análise de imagens médicas exige transparência e interpretabilidade, especialmente na área da saúde. Para isso, o uso de técnicas de explicabilidade, nos auxiliam a visualizar as regiões das imagens que influenciaram a decisão do modelo, contribuindo para um diagnóstico mais confiável e auxiliando na aceitação da tecnologia por profissionais da saúde (ADADI; BERRADA, 2018).

1.2 Motivação

Tradicionalmente, o exame de citologia cervical requer a identificação manual de células anormais sob um microscópio, o que é demorado, tedioso e sujeito a erros. Com taxas de prevalência de anormalidades de células escamosas tão baixas quanto 5% a 7% por exemplo, a grande maioria dos exames de Papanicolau que um citopatologista analisa são negativos. Encontrar campos de lâminas com células atípicas é incomum, por consequência, manter a alta vigilância é difícil, e a fadiga pode contribuir para erros de diagnóstico (BISCOTTI *et al.*, 2005). Neste contexto, um número crescente de sistemas automáticos tem sido proposto para reduzir o fardo sobre os citopatologistas e melhorar a eficiência do diagnóstico. Com o avanço da Inteligência Artificial e do processamento da imagem médica digital, tecnologias como Aprendizado de Máquina (*Machine Learning*) e Aprendizado Profundo (*Deep Learning*) foram diversas vezes aplicadas aos exames de citologia cervical para desenvolver sistemas para analisar as imagens médicas, devido ao seu alto desempenho e performance (MARINAKIS; DOUNIAS; JANTZEN, 2009; CHEN *et al.*, 2014; WILLIAM *et al.*, 2018).

A revisão bibliográfica possibilitou o mapeamento de diversos estudos para automação e desenvolvimento de ferramentas para o auxílio no diagnóstico de imagens de citologia cervical para detecção de malignidades. A sua maioria com a implementação de redes neurais convolucionais conseguindo métricas muito boas nos conjuntos de imagens públicas disponíveis. A disponibilidade de conjuntos de imagens médicas para exames de citologia cervical também aumentou consideravelmente nos últimos anos. No entanto, a metodologia continuou exatamente a mesma, dependendo apenas de um citopatologista que precisa analisar todas as células em

um esfregaço, sendo que este pode conter de 20 a 50 mil células. É irrealista pensar que um humano não estaria sujeito a erros nesse processo manual. Bengtsson e Malm (2014) apontaram que a chance de perder o diagnóstico de câncer em um único exame era bem alta, entre 20 e 30%. Isso significa que mais estudos são necessários com adição de novas tecnologias e explicabilidade dos modelos para que a comunidade médica comece a confiar nos resultados e entender sua importância.

O uso de modelos de explicabilidade surge como uma abordagem essencial nesse cenário, pois permite avaliar visualmente se o modelo está baseando suas decisões em características relevantes, possibilitando um monitoramento mais eficaz das previsões e fornecendo aos especialistas maior confiança na aplicação dessas tecnologias na prática clínica (ADADI; BERRADA, 2018).

1.3 Objetivos

1.3.1 Objetivo Geral

Desenvolver um método de processamento de imagens de citologia cervical baseado em Aprendizado Profundo para detecção de câncer cervical em imagens de citologia cervical.

1.3.2 Objetivos Específicos

Para atingir o objetivo geral, foram definidos os seguintes objetivos específicos:

- Revisar estudos e modelos utilizados para detecção de câncer em imagens de citologia cervical em portal de artigos científicos qualificados.
- Implementar modelo híbrido de Rede Neural Convolucional (CNN) combinado a um modelo de *Vision Transformers* (ViT) para detecção de câncer em imagens de citologia cervical.
- Calcular métricas para monitorar qualidade do modelo.
- Implementar modelo de explicabilidade visual para monitorar decisões do modelo.

2 REVISÃO BIBLIOGRÁFICA

Neste capítulo são explicadas várias metodologias e tecnologias a serem implementadas no desenvolvimento desse projeto, desde patologias e tipos de exames médicos realizados para o diagnóstico de câncer cervical, até conceitos técnicos e fórmulas matemáticas descrevendo arquiteturas específicas dos modelos a serem utilizados.

Foram observados trabalhos relacionados a aplicação de inteligência artificial para auxiliar diagnósticos por meio de exames de imagem, estudos e comparações realizados, tecnologias, material utilizado, dados públicos disponíveis, com foco principal em câncer cervical e exames de citologia. Para buscar as referências foram utilizados os termos e suas respectivas traduções para inglês: câncer cervical, citologia cervical, citopatologia, teste de Papanicolau, HPV, citologia em base líquida, rede neural convolucional, aprendizado profundo, *transformer*, *vision transformer*, Grad-CAM, diagnóstico por imagem médica, explicabilidade de modelos de visão computacional. A partir dos termos definidos, as consultas foram realizadas nas bases de dados científicas: IEEE Xplore, PubMed, arXiv, ScienceDirect, Web of Science e Google Scholar. Foram selecionados os artigos mais relevantes que abordavam esses assuntos.

Foram considerados artigos científicos, revisões sistemáticas e meta-análises publicados entre 2013 e 2024 que abordavam o uso de IA em diagnósticos de câncer cervical por meio de imagens médicas, e artigos mais antigos para fundamentação teórica em citopatologia e diagnóstico clínico médico. Inclusive revisões sistemáticas especificamente sobe a aplicação de algoritmos de visão computacional para diagnóstico de imagens de citologia cervical (MATIAS *et al.*, 2021; JIANG *et al.,* 2023). Para maiores detalhes da revisão bibliográfica vide o anexo A ao final da dissertação com o artigo publicado.

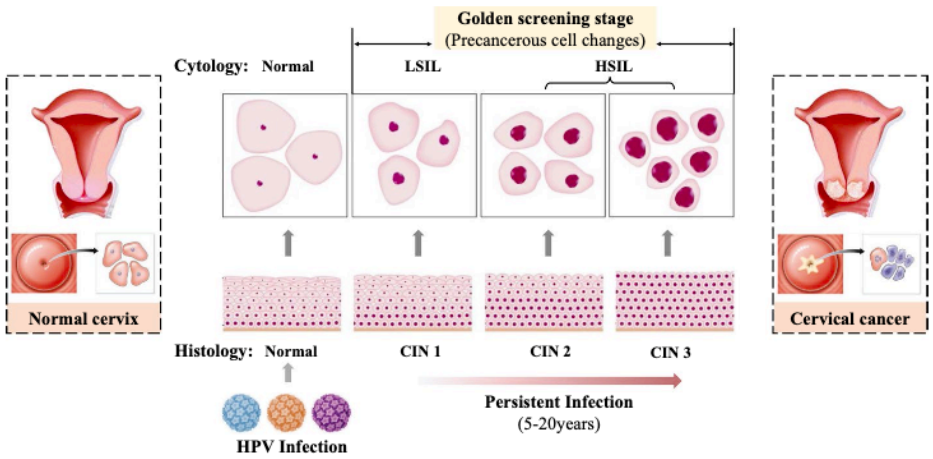
2.1 Câncer Cervical

O câncer cervical é um tipo de tumor maligno que surge do colo do útero e ameaça a vida e a saúde das mulheres. Existem dois tipos principais de câncer

cervical: o carcinoma de células escamosas; e o adenocarcinoma. O câncer cervical é de longe a doença mais comum relacionada ao Papilomavírus Humano (HPV), e quase todos os cânceres cervicais (mais de 95%) são causados por infecção persistente com alguns tipos de HPV. O HPV é a infecção sexualmente transmissível mais comum em todo o mundo e é a causa de quase todos os casos de câncer de colo do útero. HPV de alto risco pode causar câncer. Quinze tipos de HPV podem causar câncer cervical: 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 68, 73 e 82. As lesões cervicais pré-cancerosas mais frequentes são as neoplasias intraepiteliais cervicais; essas lesões podem se tornar invasivas. Já os tipos 16 e 18 são as cepas mais oncogênicas do vírus e são responsáveis por causar mais de 75% dos casos de câncer cervical e a maioria dos outros cânceres genitais (KESSLER, 2017; WAGGONER, 2003).

O câncer cervical tem um longo período de estágio pré-canceroso, e seu desenvolvimento é contínuo, conforme mostrado na Figura 1. As principais características das células pré-cancerosas concentram-se em alterações no núcleo. Por exemplo, o aumento do núcleo resulta em uma maior relação núcleo/citoplasma (relação N/C). É comum ver binucleação e multinucleação. Além disso, os nucléolos geralmente estão ausentes ou são pouco visíveis, se presentes. O contorno da membrana nuclear é bastante irregular. O diagnóstico precoce é crucial para o tratamento e prognóstico do câncer cervical (JIANG *et al.,* 2023).

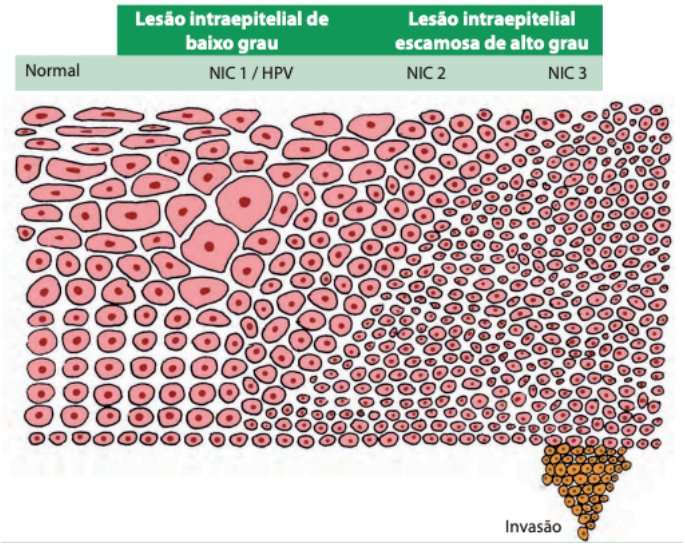
Figura 1 - A Evolução natural do câncer por HPV



Fonte: (JIANG *et al.*, 2023)

Richart em 1967, estudando a história natural do câncer do colo uterino, estabeleceu o conceito de neoplasia intraepitelial cervical (NIC) para as lesões precursoras do carcinoma escamoso invasor do colo uterino, considerando-as como um fenômeno único, contínuo e progressivo. Assim, na Figura 2, as neoplasias intraepiteliais cervicais foram divididas, histologicamente, em três graus, NIC 1, NIC 2 e NIC 3, que correspondem respectivamente à displasia leve, displasia moderada e displasia acentuada/carcinoma in situ (LIMA *et al.*, 2012).

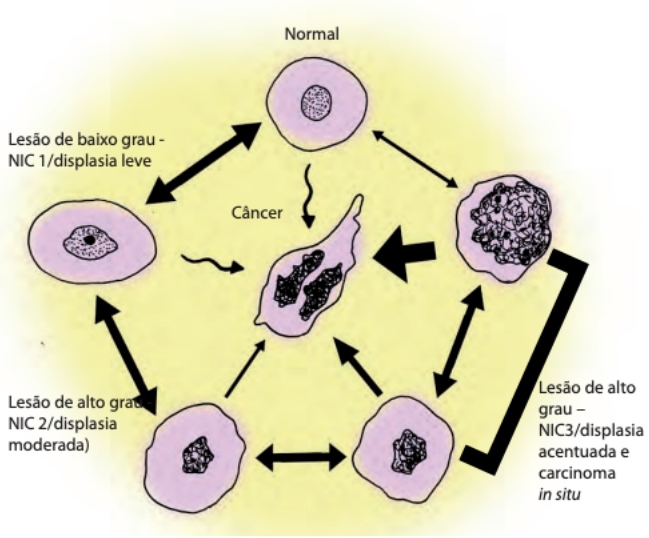
Figura 2 - Evolução de Lesões Pré-Cancerosas a Carcinoma Escamoso Invasivo



Fonte: (LIMA, 2012).

A Figura 3 ilustra o caminho habitual na progressão das lesões pré-cancerosas a partir de NIC 1, passando pelas etapas de NIC 2 e NIC 3, até a instalação do câncer invasivo. Eventualmente a NIC 2 pode evoluir diretamente para câncer. Lesões pré-cancerosas de alto grau em alguns casos podem surgir sem ser precedidas por lesões de baixo grau. E mesmo no estágio de NIC 3, a lesão pode regredir espontaneamente (LIMA, 2012).

Figura 3 - Evolução de Lesões Pré-cancerosas sob a perspectiva da Citologia



Fonte: (LIMA, 2012).

Em 1988, foi introduzido o Sistema Bethesda de terminologia e classificação diagnóstica para amostras citológicas cervicais. Esse sistema contou com duas atualizações posteriores, em 1991 e 2001. No Sistema Bethesda, as lesões pré-cancerosas do colo uterino são classificadas como lesões intraepiteliais escamosas de baixo grau (compreendendo as alterações citopáticas pelo HPV e a displasia leve / NIC 1) e lesões intraepiteliais escamosas de alto grau (correspondendo à displasia moderada/ NIC 2 e carcinoma in situ/NIC 3) (BETHESDA, 2015).

2.1.1 Teste de Papanicolau

Muitas instituições médicas recomendam a realização de exames de citologia cervical de rotina a cada alguns anos. Atualmente, o teste convencional de Papanicolau e a citologia em meio líquido (*Liquid Based Cytology* - LBC) são realizados para exames de citologia cervical no mundo todo (JIANG *et al.*, 2023). O teste de Papanicolau é um procedimento no qual as células cervicais são raspadas e observadas sob um microscópio. A Figura 4 ilustra todo o processo do teste de Papanicolau. Sob a orientação de um espéculo, uma escova macia será inserida na vagina para coletar células do colo do útero. Então, um esfregaço de Papanicolau

pode ser obtido espalhando uniformemente as células da escova sobre a lâmina de vidro. Após a coloração, os citologistas podem observar a amostra sob um microscópio e fazer um diagnóstico.

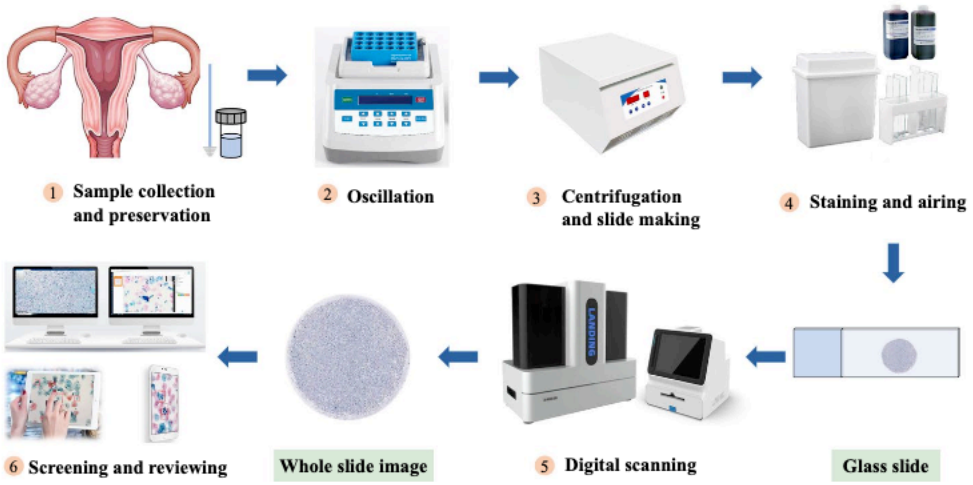
Figura 4 - Teste convencional de Papanicolau



Fonte: (JIANG *et al.*, 2023)

Devido à influência de sangue, muco, inflamação e outros fatores, o teste de Papanicolau muitas vezes adquire amostras de má qualidade, resultando em erros de detecção. Nos últimos anos, com a melhoria do nível de preparação de amostras, o LBC melhorou significativamente a qualidade de imagem das amostras de células cervicais e, assim, tornou-se gradualmente um dos principais exames de citologia cervical. Como ilustrado na Figura 5, as células coletadas são colocadas em uma solução de preservação para processamento adicional. Após oscilação e centrifugação, uma lâmina de vidro em meio líquido pode ser obtida por sedimentação natural. Então, a preparação da amostra em meio líquido é finalizada através de coloração e secagem ao ar.

Figura 5 - Citologia em Base Líquida



Fonte: (JIANG *et al.*, 2023)

Hoje em dia, com o desenvolvimento de equipamentos de imagem e técnicas de processamento digital, amostras de teste de Papanicolau e LBC geralmente são transformadas em lâminas digitais por meio de scanners patológicos para facilitar o exame retrospectivo. A patologia digital traz um impacto positivo e profundo no diagnóstico patológico tradicional, que digitaliza lâminas de vidro em imagens de lâminas inteiras para reduzir a carga de trabalho dos patologistas e melhorar a eficiência do diagnóstico em comparação com a observação visual baseada em microscópio (NIAZI; PARWANI; GURCAN, 2019). A preparação em meio líquido junto com lâminas digitais é uma alternativa satisfatória ao esfregaço convencional e tem uma grande perspectiva de aplicação para os exames de prevenção de câncer cervical. Não é uma tecnologia empregada no Brasil, estudos não comprovaram a superioridade da citologia em base líquida em comparação ao teste convencional e os custos foram considerados muito altos (BRASIL, 2019).

Felizmente, o câncer de colo do útero pode ser curado quando alterações cancerosas precoces ou lesões precursoras causadas pelo Papilomavírus Humano (HPV) são detectadas pelo teste convencional de Papanicolau. No entanto, a taxa de cura está intimamente relacionada ao estágio da doença no momento do diagnóstico, com uma probabilidade muito alta de fatalidade se não for tratada. Portanto, a identificação precoce dos casos positivos é crucial.

Desde a criação da metodologia para um teste de triagem, nomeadamente o teste de Papanicolau, introduzido pelo Dr. George Papanicolau na década de 1940, observou-se uma diminuição substancial na taxa de câncer de colo do útero e na mortalidade relacionada. O teste de Papanicolau tem sido o teste de detecção de câncer mais eficaz de todos os tempos e ainda permanece a modalidade crucial na detecção das lesões precursoras para o câncer de colo do útero. O teste é baseado na obtenção de células do colo do útero e na sua transferência para lâminas de vidro para exame microscópico a fim de detectar os efeitos do HPV. As lâminas são coradas usando o método de Papanicolau, onde diferentes componentes das células mostram cores diferentes, de modo que seu exame se torna mais fácil (GENÇTAV; AKSOY; ÖNDER, 2012).

Existem certos fatores associados à sensibilidade do teste de Papanicolau e, portanto, à confiabilidade do diagnóstico. A sensibilidade do teste é prejudicada

principalmente pela qualidade da amostragem (por exemplo, número de células) e do esfregaço (por exemplo, presença de elementos obscuros como sangue, muco e células inflamatórias, ou fixação inadequada das amostras). As variabilidades de interpretação de amostras anormais dependendo do observador também contribuem para a ampla variação nos resultados falso-negativos. Outros fatores incluem o tempo excessivamente longo na interpretação das amostras, a natureza subjetiva da interpretação e a impossibilidade de assegurar a invasão ou extensão da invasão no caso de lesão maligna (LIMA *et al.*,2012). A promessa de diagnóstico precoce e as dificuldades associadas ao processo de triagem manual, tornaram o desenvolvimento de sistemas automatizados ou semiautomatizados que analisam imagens adquiridas usando uma câmera digital conectada ao microscópio um importante problema de pesquisa, onde se espera que um exame mais robusto, consistente e quantificável dos esfregaços aumente a confiabilidade do diagnóstico (GENÇTAV; AKSOY; ÖNDER, 2012).

2.1.2 Imagem Médica

A imagem médica é uma técnica de captura da morfologia e fisiologia de diferentes partes do corpo humano, através de imagens e compilação de dados gráficos para análise clínica adequada e intervenção médica. As ondas eletromagnéticas ou acústicas são transmitidas através do corpo humano para capturar uma representação visual da anatomia. A imagem médica também inclui técnicas de medição e registro que não produzem imagens, mas geram dados em uma representação gráfica como o eletrocardiograma. O campo das imagens médicas remonta à história do século XIX, com a introdução dos raios-X por Wilhelm C. Röntgen em 1895. O princípio básico da técnica era a penetração variável dos raios-X através da carne e dos ossos humanos. Desde então, muitos avanços no campo foram feitos, que são amplamente classificados em técnicas de imagem morfológica, anatômica e molecular. As primeiras incluem ultrassom, tomografia computadorizada (CT) e ressonância magnética (RM), que observam principalmente as mudanças estruturais. As últimas incluem técnicas como imagem óptica, tomografia por emissão de pósitrons (PET) e tomografia por emissão de fóton único

(SPECT), que escaneiam detalhes moleculares do sistema biológico dado. Ambos os tipos de técnicas têm suas vantagens e, portanto, agora são usados em conjunto para permitir imagens de alta resolução estruturada como PET-CT ou PET-RM. Ao contrário deles, o ultrassom usa ondas sonoras de alta frequência para se propagar dentro de sistemas biológicos, e os ecos retornados são analisados para formação de imagem. A última década viu alguns avanços promissores no fMRI funcional existente, que melhoraram os limites de resolução espacial e temporal em muitas vezes. A análise das imagens obtidas por fMRI é usada para visualização 3D e estudos anatômicos. Agora, com a introdução do elemento tempo em 3D, a imagem 4D está ganhando popularidade e, quando combinada com métodos de TC, RM, PET e SPECT, pode ser usada no estudo da dinâmica de fluidos dos órgãos internos humanos e gerar modelos de malha estruturada deles. Os avanços nas modalidades de imagem médica ao longo do tempo levaram a uma aquisição significativa de dados e a um aumento da carga sobre os especialistas médicos para analisá-los, o que os tornou mais propensos a erros. Isso levou à era da Inteligência Artificial, para tarefas relativamente mais simples como 'classificação' já em 1995. Posteriormente, surgiram ferramentas de Aprendizado de Máquina e depois tecnologias mais avançadas de Aprendizado Profundo para a remoção de ruídos e outras etapas de pré-processamento antes da análise real da imagem (SINGHA; THAKUR; PATEL, 2021).

Houve muitos avanços em termos de aplicações para Aprendizado Profundo em imagens médicas a nível celular também. Pré-processamento, segmentação, detecção são alguns passos comuns geralmente implementados, seguidos por uma análise das morfologias celulares. Estruturas de detecção são efetivamente usadas no diagnóstico de tumores também. Zeune *et al.* (2020) propuseram uma rede de classificação multiclasse de autoencoder CNN juntamente com técnicas de visualização fluorescente para detectar e contar células circulantes com base em sua morfologia e origem com uma precisão, especificidade e sensibilidade de mais de 96%. Os resultados do modelo mostram que ele pode prever com precisão a sobrevida global de pacientes com câncer.

Redes neurais profundas (*Deep Neural Networks* - DNN) também são úteis na detecção de mitose, que é uma tarefa desafiadora para o olho humano. Uma

estimativa de produtos mitóticos é um indicativo de potencial desenvolvimento de câncer em lâminas de histologia. Em 2012, uma DNN treinada em inúmeras imagens foi usada para detectar mitose em células de câncer de mama. Ela trabalhou diretamente com dados RGB obtidos de um ponto específico da imagem de origem e os analisou com base em características visuais aprendidas do conjunto de dados de treinamento mitótico. Hoje (2024), redes neurais R-CNN e deep CNN são usadas para detectar mitose em imagens de histopatologia de câncer, e são rápidas e precisas (CIREŞAN *et al.*, 2013).

Com a integração de lâminas digitais no fluxo de trabalho da patologia também, algoritmos avançados e técnicas de diagnóstico auxiliadas por computador estendem as fronteiras da visão do patologista para além de uma lâmina microscópica e possibilitam a verdadeira utilização e integração do conhecimento além dos limites e fronteiras humanas. A IA já permite que patologistas identifiquem marcadores de imagem únicos associados a processos de doenças com o objetivo de melhorar a detecção precoce, determinar prognósticos e selecionar tratamentos com maior probabilidade de serem eficazes. Isso permite que os patologistas atendam mais pacientes enquanto mantêm a precisão diagnóstica e prognóstica. Essa integração é especialmente importante considerando o número crescente de pacientes idosos e o fato de que menos de 2% dos graduados em medicina escolhem a patologia devido ao gerenciamento de informações, integração e acessibilidade de mídia digital aspecto do processo de revisão e assinatura de casos (NIAZI; PARWANI; GURCAN, 2019).

A patologia digital e a IA podem ter um potencial imenso para a oncologia e a medicina de precisão. Assim como a evolução da eficiência e eficácia na radiologia, a pressão sobre os patologistas para reduzir o tempo de resposta e desenvolver fluxos de trabalho mais eficientes está tendendo para a digitalização. Inteligência Artificial, redes neurais e a técnica de Aprendizado Profundo serão explicadas com maior detalhe na próxima seção.

2.2 Inteligência Artificial

O termo Inteligência Artificial (IA) foi criado em uma conferência por John McCarthy e foi lá que toda a área de pesquisa se originou. Em 1955, o Projeto de Pesquisa Dartmouth definiu a IA como o problema de fazer uma máquina se comportar de maneiras que seriam chamadas de inteligentes se um humano estivesse se comportando assim. De maneira semelhante, o cientista cognitivo Marvin Minsky considerou a IA como a ciência de fazer máquinas fazerem coisas que exigiriam inteligência se feitas por homens. A partir desse contexto, Kaplan e Haenlein (2019) definem a IA como a capacidade de um sistema interpretar corretamente dados externos, aprender com esses dados e usar essas aprendizagens para alcançar objetivos e tarefas específicas por meio de flexibilidade e adaptação.

A Inteligência Artificial, que engloba o Aprendizado de Máquina (*Machine Learning*) entre outros, é desenhada para aprender como uma criança. O Aprendizado de Máquina é composto de uma série de algoritmos. A partir de um conjunto de dados, uma Inteligência Artificial pode encontrar padrões e formular lógicas baseadas nestas descobertas. Chamado de aprendizado baseado em modelo, ele permite à Inteligência Artificial tomar melhores decisões do que humanos, porque ela pode levar em consideração um maior número de fatores e analisá-los em milissegundos (GONFALONIERI, 2019).

Os algoritmos de Aprendizado de Máquina são divididos em 3 classes: Supervisionados, Não-Supervisionados e Aprendizado por Reforço. Essas classes por sua vez possuem diversas subclasses, e às vezes podem até se misturar em soluções híbridas (GONFALONIERI, 2019).

O Aprendizado de Máquina formula lógicas, reavalia o modelo e os dados, tudo sem intervenção humana. Um programador não precisa escrever códigos para toda e qualquer ação ou reação. O sistema de Aprendizado de Máquina é capaz de descobrir todos os padrões possíveis em uma velocidade inalcançável por um humano. Há com isso uma diferença essencial entre algoritmos tradicionais e algoritmos de Aprendizado de Máquina.

O algoritmo tradicional na Figura 6 recebe uma entrada e uma lógica na forma de código, e com isso produz uma saída. Ou seja, ele produz uma saída baseado nos passos descritos no algoritmo. Ao receber uma entrada, o algoritmo produz a saída baseada nas regras e parâmetros descritos no código desenvolvido pelo programador.

Figura 6 - Algoritmo Tradicional



O algoritmo de Aprendizado de Máquina na Figura 7, no caso um algoritmo supervisionado, recebe uma entrada e uma saída, e constrói uma lógica com a qual recebe novas entradas para gerar uma saída. A lógica desenvolvida é o que o faz Aprendizado de Máquina.

Figura 7 - Algoritmo de Aprendizado de Máquina



O algoritmo de Aprendizado de Máquina faz a estimativa de uma saída com base no aprendizado através das entradas fornecidas a ele. Esse aprendizado por entradas é chamado de processo de treinamento. Dê ao algoritmo dados para aprender, e ele ajustará parâmetros para explicar os dados. Com esses parâmetros ajustados, o algoritmo pode então explicar ou fazer estimativas para novos dados.

2.2.1 Visão Computacional

A visão computacional é a tecnologia de processamento e análise de imagens por meio de computadores. É uma tecnologia importante no campo da IA, que se baseia em Aprendizado Profundo (*Deep Learning*). O ser humano está constantemente analisando o mundo a sua volta. Sem fazer esforço, faz análises sobre tudo o que vê e age de acordo. Quando vê alguma coisa, rotula cada objeto baseado no que aprendeu no passado. A profunda estrutura hierárquica de neurônios e conexões no cérebro fazem a maior parte desse processo de relembrar e rotular objetos.

O objetivo da visão computacional é permitir que os computadores reconheçam e entendam o mundo através da visão, como os humanos fazem. Ela utiliza principalmente algoritmos para identificar e analisar imagens. As visões computacionais mais amplamente utilizadas são o reconhecimento facial e o reconhecimento de imagem.

Desde 2015, métodos de processamento de imagem que fazem uso do Aprendizado Profundo para classificação de imagens foram muito utilizados. A rede neural desses modelos é construída com múltiplas camadas de neurônios, possuindo uma função de ajuste não linear através da função de ativação. Uma vez que a necessidade de projetar a entrada e a saída para o modelo é atendida, o modelo pode aprender automaticamente, realizando o processo de extração de características e treinamento de classificação. O uso do Aprendizado Profundo simplifica o processo de classificação de imagem que é normalmente mais demorado e trabalhoso e melhora tanto o efeito quanto a eficiência da classificação de imagem. Como exemplo, a rede neural convencional VGG, a ResNet (Rede Neural Residual) e a Startup são arquiteturas muito utilizadas na engenharia. Para selecionar um modelo é necessário considerar tanto a eficiência quanto a eficácia, ou seja, ele deve garantir precisão ao mesmo tempo que garante a agilidade. Após o treinamento do modelo, o modelo pode ser refinado e comprimido. Como exemplo, Faster R-CNN, Mask-RCNN e YOLO (You Only Look Once) que são hoje modelos de rede neural muito utilizados. As características comuns desses modelos são a

sua alta precisão e rápida velocidade. No campo do reconhecimento facial, esses modelos podem detectar e obter resultados em tempo real (ZHANG; LU, 2021).

A visão computacional divide a imagem em pixels e então processa os pixels, e ela evoluiu bastante ao longo dos anos, como com a segmentação semântica. O significado da segmentação semântica é entender o significado dos pixels após a segmentação. Isso permite o reconhecimento, por exemplo, de pessoas, motocicletas, carros e semáforos em uma imagem. As redes neurais convolucionais promoveram o desenvolvimento de algoritmos de segmentação semântica e sua arquitetura também evoluiu (ZHAO, R. *et al.*, 2019). O método mais básico da segmentação semântica era realizar uma classificação de imagem, ou seja, reconhecer o objeto na imagem, através de uma janela deslizante na imagem ou uma caixa delimitadora (*bounding box*). Com o surgimento da Rede Totalmente Convolucional (*Fully Convolutional Network*) as camadas totalmente conectadas das redes neurais foram substituídas. Com base na FCN, a arquitetura *Encoder-Decoder* foi desenvolvida posteriormente. Como o codificador falhava em processamento de imagens muito pequenas, o decodificador foi utilizado para restaurar as imagens para um tamanho maior com informações detalhadas. Depois disso, a operação de *pooling* também foi substituída pelas operações de convolução dilatada (*Dilated* ou *Atrous Convolution*). A vantagem da convolução dilatada é que a resolução espacial pode ser mantida. Além do método anterior, surgiu também um novo método chamado Campo Aleatório Condicional (*Conditional Random Fields*) que pode melhorar o efeito da segmentação (ZHAO, Z. Q. *et al.*, 2019).

2.2.2 Aprendizado Profundo

Originado das Redes Neurais Artificiais, o Aprendizado Profundo (*Deep Learning*) é um ramo da aprendizado de máquina que se destaca por ter múltiplas camadas de processamento não lineares e tenta aprender representações hierárquicas dos dados. Até o momento, existem várias arquiteturas de aprendizado profundo e este tópico de pesquisa está em rápido crescimento, com novos modelos sendo desenvolvidos até mesmo a cada semana. A comunidade é bastante aberta

e há uma série de tutoriais e livros de alta qualidade sobre aprendizado profundo (ZHAO, R. *et al*., 2019).

2.2.3 Redes Neurais Convolucionais

Dentro do Aprendizado Profundo, a arquitetura para as redes neurais convolucionais (CNNs) foram inicialmente propostas por LeCun *et al*. (1990) para o processamento de imagens, caracterizando-se por duas propriedades-chave: pesos compartilhados espacialmente e pooling espacial. Lenet definiu pela primeira vez a estrutura da rede CNN em 1998 (LECUN *et al.*, 1998), e não foi até a publicação do AlexNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012) em 2012 que a CNN começou a se tornar gradualmente uma tendência. Desde então, muitas redes neurais convolucionais eficientes e profundas foram propostas. Por exemplo, VGG (SIMONYAN; ZISSERMAN, 2014), *ResNet* (HE *et al.*, 2016), *DenseNet* (HUANG *et al.*, 2017), *GoogleNet* (SZEGEDY *et al.*, 2015), InceptionV3 (SZEGEDY *et al.*, 2016) e *EfficientNet* (TAN; LE, 2019). Essas redes apresentavam bom desempenho em várias aplicações. Com o desenvolvimento da CNN, a U-Net foi proposta e amplamente utilizada em tarefas de segmentação devido à sua estrutura simples, bons efeitos e forte generalização (ZHOU *et al.*, 2018). Os modelos de CNN demonstraram seu sucesso em várias aplicações de visão computacional onde os dados de entrada são geralmente dados de duas dimensões. A CNN também foi introduzida para lidar com dados sequenciais, incluindo Processamento de Linguagem Natural e Reconhecimento de Fala.

A CNN tem como objetivo aprender características abstratas alternando e empilhando camadas convolucionais e de pooling. Na CNN, as camadas convolucionais (kernels convolucionais) convolvem múltiplos filtros locais com dados de entrada brutos em três dimensões e geram características locais invariantes à translação, e as subsequentes camadas de pooling extraem características de comprimento fixo sobre janelas deslizantes dos dados de entrada brutos, seguindo várias regras, como média, máximo e assim por diante (ZHAO, R. *et al*., 2019).

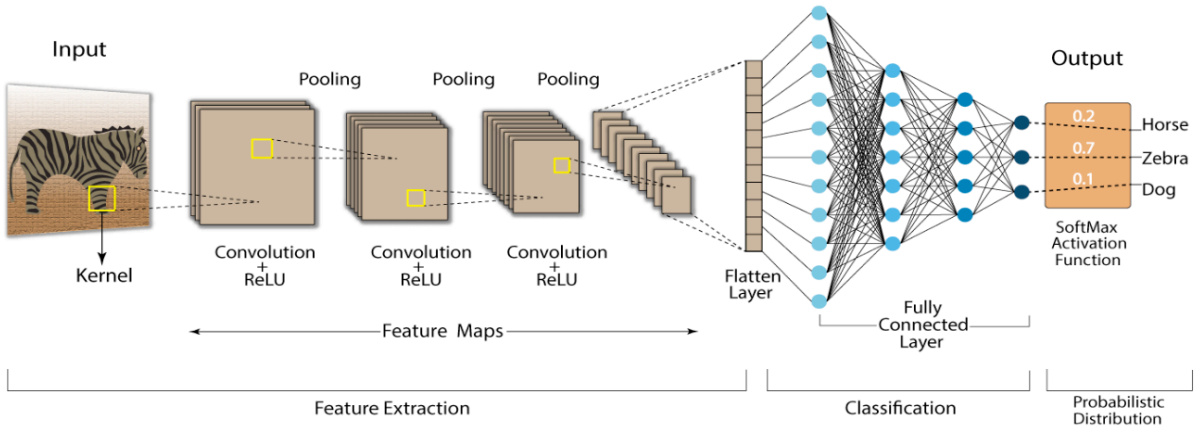
A equação 1 que descreve a operação de uma CNN é:

$$x_{ji} = f\left(\sum_{i \epsilon M j} x_i^{l-1} * w_{ji}^{l} + b_j^{l}\right) \qquad (1)$$

Onde M_j representa a seleção do mapa de características de entrada, x_i^{l-1} é a saída da camada anterior, ou seja, o mapa de características. w_{ji}^l é o peso do núcleo de convolução da camada de número l, f é a função de ativação e, b_j^lé o viés da camada de número l (SINGHA; THAKUR; PATEL, 2021).

Os elementos da arquitetura típica de uma CNN para treinar uma imagem de entrada de 256 × 256 com uma função objetivo de aprendizado supervisionado são detalhados a seguir. A função objetivo ou função de perda é utilizada para otimização no treinamento da CNN para uma aplicação específica de processamento de imagem. As diferentes operações que ocorrem na arquitetura da CNN na Figura 8 são convolução, *pooling* e transformação não linear usando a função de ativação (SINGHA; THAKUR; PATEL, 2021).

Figura 8 - Arquitetura da Rede Neural Convolucional (CNN)



Fonte: (KUMAR; LUO; SHAUKAT, 2023)

Convolução

O filtro de convolução consiste em núcleos de peso que detectam uma característica particular da imagem. O filtro de convolução de tamanho fixo é transladado por toda a imagem, e a saída é a multiplicação matricial dos pesos do filtro de convolução com a parte correspondente da matriz de imagem de entrada (SINGHA; THAKUR; PATEL, 2021).

Pooling

O operador de *pooling* é aplicado a um mapa de características para redução da resolução espacial. O mapa de características torna-se espacialmente invariante à distorção e translação dos dados de entrada quando sua resolução espacial é reduzida (SINGHA; THAKUR; PATEL, 2021).

Função de Ativação

A função de ativação transforma a saída convolucionada realizando uma transformação não linear. As funções de ativação mais comumente usadas são as unidades lineares retificadoras, ou *Rectified Linear Unit*, abreviada como *ReLU*. Além desta, outras também muito utilizadas são a função hiperbólica tangente, a função *softmax* e a sigmoidal (KUO, 2016).

Funções de Perda

Ao compilar o modelo, antes de começar o treinamento, é necessário adicionar uma função de perda (*loss function*). Através dela, a rede neural convolucional consegue balancear a perda e a acurácia entre o conjunto de imagens para treinamento e o conjunto de imagens para validação. Essa função inicia a Retropropagação (*backpropagation*), que é quando os dados retornam para a camada inicial e os pesos de cada neurônio são recalibrados pela função de otimização de gradiente descendente (ou estocástico). Esse ciclo se repete até a função de perda alcançar o menor valor aceitável (LECUN *et al*., 1998; WU, 2019).

No início do processo de retropropagação, o valor de saída é normalmente muito pequeno, muito menor do que o valor desejado. O gradiente geralmente também é muito baixo, aumentando a dificuldade para a rede utilizar os dados que ela possui para ajustar os pesos e se otimizar. A função *Cross-Entropy*, através do seu logaritmo, deixa a rede neural avaliar esses pequenos erros e trabalhar para eliminá-los, a tornando a ferramenta de otimização mais eficiente. Até duas classes de saída, a função *binary_crossentropy* é utilizada, para mais de duas classes, a função *categorical_crossentropy* (WU, 2019).

2.2.4 Vision Transformer

Os *Vision Transformers* (ViTs) emergiram recentemente como uma alternativa competitiva às Redes Neurais Convolucionais (CNNs) que atualmente são o estado da arte em diferentes tarefas de reconhecimento de imagem em visão computacional (DOSOVITSKIY *et al.*, 2021).

Os modelos de Transformers tornaram-se o padrão na prática no Processamento de Linguagem Natural (NLP). Por exemplo, o popular chatbot de IA *ChatGPT* é um modelo de linguagem baseado em *Transformer*. Ele é baseado na arquitetura GPT (*Generative Pre-trained Transformer*) que utiliza mecanismos de autoatenção para modelar as dependências entre palavras em um texto (JAVAID; HALEEM; SINGH, 2023).

Enquanto a arquitetura *Transformer* é muito utilizada para tarefas envolvendo Processamento de Linguagem Natural (NLP), seus casos de uso relacionados à Visão Computacional ainda são poucos. Em muitas tarefas de visão computacional, usamos o mecanismo de atenção tanto em conjunto com redes neurais convolucionais (CNN) quanto para substituir certos aspectos das CNNs, mantendo sua composição integral. Modelos populares de reconhecimento de imagem incluem *ResNet, VGG, YOLOv3, YOLOv7* ou *YOLOv8* e *Segment Anything* (SAM) (ZHANG; LU, 2021).

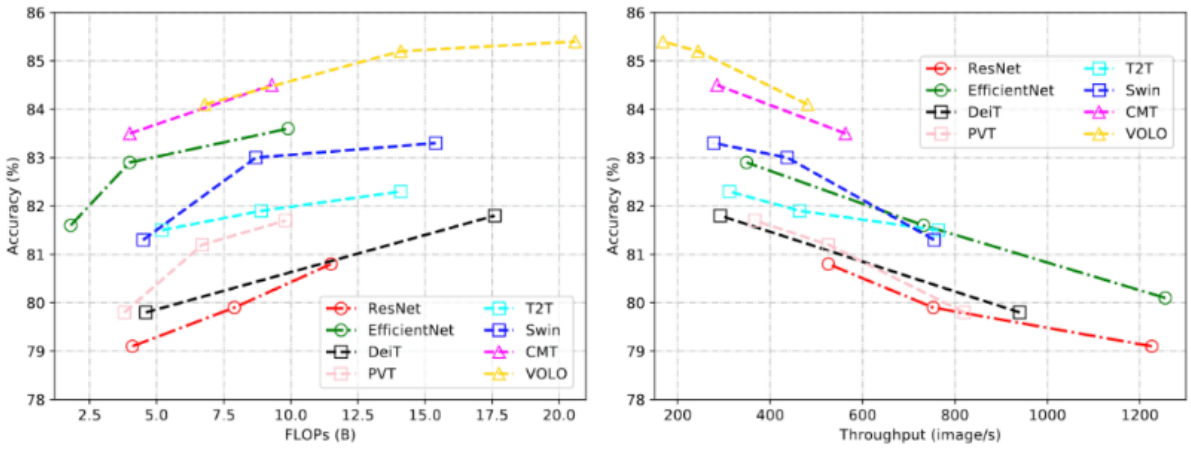
Os ViTs alcançaram recentemente um desempenho altamente competitivo em benchmarks para várias aplicações de visão computacional, como classificação de imagens, detecção de objetos e segmentação semântica de imagens. Como por exemplo com o modelo *CSWin Transformer*, ou com o modelo *CrossViT*, um modelo ViT de atenção cruzada para classificação de imagens. Esse último com estudos que indicam que quando pré-treinados com uma quantidade suficiente de dados, ele pode apresentar tão boa performance quando modelos de CNNs como o *ResNet* (BHOJANAPALLI *et al.*, 2021).

O ViT alcança resultados notáveis em comparação com as CNNs, enquanto utilizando substancialmente menos recursos computacionais para pré-treinamento. Em comparação com as CNNs, o ViT mostra um viés indutivo geralmente mais fraco, resultando em um aumento da dependência de regularização do modelo ou de *data*

augmentation ao treinar com conjuntos de dados menores (CHEN; HSIEH; GONG, 2022).

O ViT é um modelo visual baseado na arquitetura de um *Transformer* originalmente projetado para tarefas baseadas em texto. O modelo ViT representa uma imagem de entrada como uma série de trechos da imagem, semelhante à série de incorporações de palavras usadas ao aplicar *Transformers* a textos, e prevê diretamente os rótulos de classificação para a imagem. O ViT exibe um bom desempenho quando treinado com um quantidade suficiente de dados, superando o desempenho de uma CNN similar com 4 vezes menos recursos computacionais. Na Figura 9 pode-se ver a comparação de performance entre 7 modelos: CNNs: ResNet, EfficientNet; Puro Transfomer: DeiT, T2T, PVT, Swin; CNN + Transformer: CMT, VOLO (HAN *et al.*, 2023).

Figura 9 - CNNs x ViTs



Fonte: (HAN *et al.*, 2023)

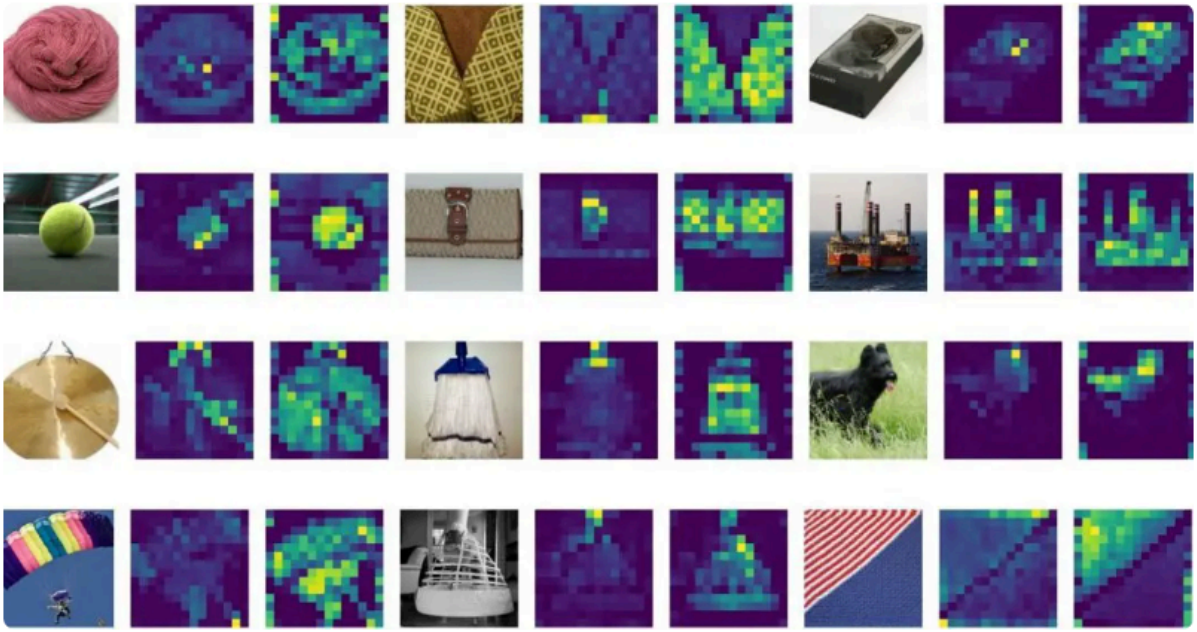
Os modelos ViT superam as atuais CNNs em quase 4 vezes em termos de eficiência computacional e precisão (PAUL; CHEN, 2022). As CNNs utilizam matrizes de pixels, enquanto os ViTs dividem as imagens de entrada em chaves visuais. O ViT divide uma imagem em recortes de tamanho fixo, incorpora cada um deles e inclui a posição como uma entrada para o codificador do *Transformer*. A camada de autoatenção no ViT torna possível incorporar informações de forma global por toda a imagem. O modelo também aprende com os dados de treinamento a codificar a

localização relativa dos pedaços da imagem para reconstruir a estrutura da imagem (DOSOVITSKIY *et al.*, 2021).

O mecanismo de autoatenção é essencial na arquitetura do *Transformer*, permitindo que o modelo identifique e priorize informações contextuais e dependências de longo alcance nos dados de entrada. Em um modelo de ViT, a autoatenção capacita o foco em regiões variadas dos dados conforme sua relevância para a tarefa específica. Funciona calculando uma soma ponderada das entradas, onde os pesos são derivados da similaridade entre as características de entrada. Isso possibilita ao modelo enfatizar as características mais pertinentes, melhorando a captura de representações informativas dos dados. A autoatenção, como uma ferramenta computacional, quantifica as interações entre entidades, auxiliando a rede a aprender sobre a hierarquia e o alinhamento inerentes aos dados. Esse mecanismo tem se mostrado um componente crucial para aprimorar a robustez das redes de visão computacional (VASWANI *et al.*, 2017; DOSOVITSKIY *et al.*, 2021).

Os mapas de atenção na Figura 10 do modelo ViT são ferramentas que indicam a relevância de diferentes regiões de uma imagem para as representações que o modelo aprende. Inicialmente, a imagem é dividida em recortes que são linearizados e inseridos no codificador do *Transformer*. Os mapas de atenção surgem do mecanismo de autoatenção, onde cada pedaço da imagem é comparado com todos os outros, gerando pesos de atenção que indicam a importância relativa entre eles. Esses pesos podem ser visualizados como mapas de calor, onde cores mais intensas indicam maior atenção entre as partes. Analisando esses mapas, é possível entender quais áreas da imagem são mais influentes para a classificação que o modelo realiza (CHEN; HSIEH; GONG, 2022).

Figura 10 - Imagens e seus respectivos mapas de atenção

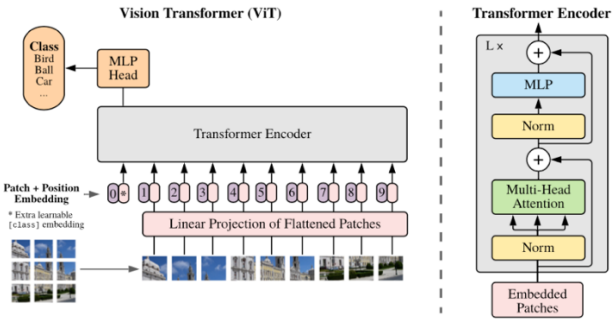


Fonte: (CHEN; HSIEH; GONG, 2022)

Arquitetura do modelo de Vision Transformer

Na literatura são encontradas muitas propostas para a arquitetura do modelo de *Vision Transformer*, mas a arquitetura na Figura 11 desenvolvida pelo Google é a mais citada em artigos facilitando o entendimento de sua estrutura básica (DOSOVITSKIY *et al.*, 2021).

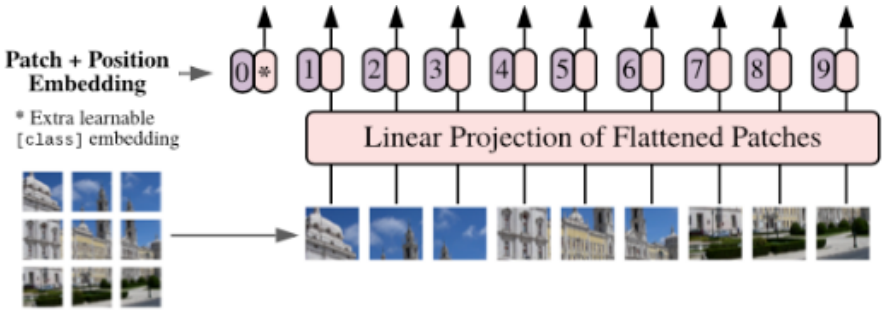
Figura 11 - Arquitetura do modelo de Vision Transformer (ViT)



Fonte: (DOSOVITSKIY *et al.*, 2021)

No primeiro passo na Figura 12, uma imagem de entrada com formato (*altura, largura, canais*) é incorporada em um vetor de características com formato $(\boldsymbol{n}+\mathbf{1}, \boldsymbol{d})$, seguindo uma sequência de transformações.

Figura 12 - ViT: Primeiro passo



Fonte: **(**DOSOVITSKIY *et al.*, 2021**)**

Isso corresponde à equação 2 do artigo:

$$\boldsymbol{z_0} = \left[\boldsymbol{x_{class}};\ \boldsymbol{x_p^1}\mathbf{E};\ \boldsymbol{x_p^2}\mathbf{E};\ \ldots;\ \boldsymbol{x_p^n}\right] + \mathbf{E}_{\boldsymbol{pos}},\ \mathbf{E} \in \mathbb{R}^{\left(\boldsymbol{p^2 * c}\right)\times \boldsymbol{D}}, \mathbf{E}_{\boldsymbol{pos}} \in \mathbb{R}^{(\boldsymbol{N+1})\times \boldsymbol{D}} \qquad (2)$$

1. A imagem é dividida em $\boldsymbol{n}$ recortes quadrados de formato $(\boldsymbol{p}, \boldsymbol{p}, \boldsymbol{c})$, onde $\boldsymbol{p}$ é um parâmetro pré-definido, em ordem de varredura (da esquerda para a direita, de cima para baixo).

2. Os recortes são achatados, resultando em $\boldsymbol{n}$ vetores de linha de formato $(\mathbf{1}, \boldsymbol{p}^2 * \boldsymbol{c})$.

3. Os recortes achatados são multiplicados por um tensor de incorporação treinável de formato $(\boldsymbol{p}^2 * \boldsymbol{c}, \boldsymbol{d})$, que aprende a projetar linearmente cada recorte achatado para a dimensão $\boldsymbol{d}$. Essa dimensão $\boldsymbol{d}$ é constante na arquitetura e é usada na maioria dos componentes. O resultado são $\boldsymbol{n}$ recortes incorporados de formato $(\mathbf{1}, \boldsymbol{d})$.

4. Um token $[\boldsymbol{cls}]$ aprendiz de formato $(\mathbf{1}, \boldsymbol{d})$ é adicionado antes da sequência de incorporações de recortes. A ideia deste token vem do artigo do BERT (DEVLIN *et al.*, 2019), onde apenas a última representação correspondente a este token (saída do Transformer $\boldsymbol{L}$) é alimentada pelas camadas de classificação.

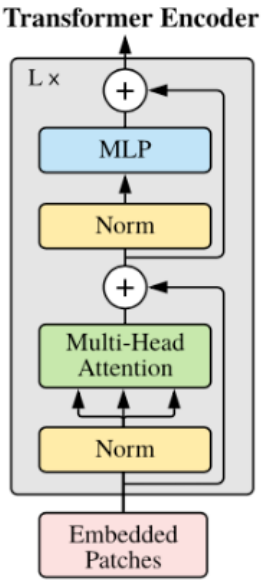
Intuitivamente, isso representa um agrupamento das representações dos recortes.

5. Um tensor de incorporação posicional treinável, $\mathbf{E}_{pos}$, com o mesmo formato, $(\boldsymbol{n+1}, \boldsymbol{d})$, é adicionado à sequência concatenada de projeções. Esse tensor aprende informações posicionais de uma dimensão para cada um dos recortes, a fim de adicionar uma representação espacial de cada recorte dentro da sequência.

O resultado, $\boldsymbol{z}_0$, é a primeira entrada para os codificadores dos *Transformers* empilhados. Os $\boldsymbol{L}$ codificadores empilhados representam o segundo componente da arquitetura. Cada *Transformer* recebe como entrada características representadas como um tensor $(\boldsymbol{n+1}, \boldsymbol{d})$ e produz uma saída de mesma dimensão.

No segundo passo na Figura 13, a rede aprende características mais abstratas dos recortes de imagem embutidos, usando uma pilha de $\boldsymbol{L}$ codificadores de *Transformers*.

Figura 13 - ViT: Segundo passo



Fonte: **(**DOSOVITSKIY *et al.*, 2021**)**

Isso corresponde às equações 3 e 4 do artigo a seguir:

$$\boldsymbol{z'_{\ell} = MSA\big(LN(z_{\ell} - 1)\big) + z_{\ell} - 1,\ \ell = 1 \ldots L} \qquad (3)$$

$$\boldsymbol{z_{\ell} = MLP\big(LN(z'_{\ell})\big) + z'_{\ell},\qquad \ell = 1 \ldots L} \qquad (4)$$

O componente codificador contém um mecanismo de atenção multi-cabeças, ou *Multi-Head Attention* (MHA), e um *perceptron* multicamadas, ou *Multilayer Perceptron* (MLP), de 2 camadas, com normalização de camada e conexões residuais entre eles.

A normalização de camada ajuda a estabilizar a dinâmica do estado oculto e a reduzir o tempo de treinamento. Isso é feito escalando com a média e o desvio padrão para cada exemplo de treinamento (ao contrário da normalização em lote, onde isso é feito por característica). As características resultantes são multiplicadas por um fator de escala e adicionadas a um fator de deslocamento, ambos aprendidos durante o treinamento.

Conexões residuais oferecem caminhos alternativos para os gradientes, a fim de resolver o problema dos gradientes desaparecendo em arquiteturas muito profundas.

Os pesos treináveis neste componente estão dentro do mecanismo MHA e dos pesos do MLP. Como o MLP tem 2 camadas (oculta e de saída), haverá duas matrizes de pesos: W_h e W_o.

A etapa de atenção multi-cabeças (MHA), incluída em cada um dos L *Transformers* empilhados, corresponde às equações 5, 6, 7 e 8 do apêndice do artigo:

$$[\mathbf{q}, \mathbf{k}, \mathbf{v}] = \mathbf{z}\mathbf{U}_{qkv} \quad \mathbf{U}_{qkv} \in \mathbb{R}^{D \times 3D_h}, \qquad (5)$$

$$A = \mathrm{softmax}\big(\mathbf{q}\mathbf{k}^{\mathrm{T}}/\sqrt{D_h}\big) \quad A \in \mathbb{R}^{N \times N}, \qquad (6)$$

$$\mathrm{SA}(\mathbf{z}) = \mathrm{Av}. \qquad (7)$$

$$\mathrm{MSA}(\mathbf{z}) = [\mathrm{SA}_1(z); \mathrm{SA}_2(z); \ldots; \mathrm{SA}_{\mathrm{k}}(z)]\mathbf{U}_{msa} \quad \mathbf{U}_{msa} \in \mathbb{R}^{kD_h \times D} \qquad (8)$$

O estado oculto do codificador anterior é dividido em k cabeças, resultando em k tensores de características de forma (n, d_h). A partir da intuição da atenção

multi-cabeças, múltiplas cabeças permitem que o mecanismo aprenda a partir de diferentes aspectos da representação abstrata.

Cada uma é multiplicada por 3 matrizes treináveis Qi, Ki, Vi, de formato (d_h, d_h). Isso é equivalente à primeira linha da equação, pois em $\mathbf{U} = (d, 3d_h)$ existem exatamente 3 matrizes para cada cabeça, cada uma de formato (d_h, d_h).

Qi, Ki e Vi representam a projeção da entrada em 3 subespaços. Podemos pensar em cada linha em Q como uma projeção aprendida do pedaço que nos interessa, e linhas em K como outros pedaços com os quais comparamos Q. V e K são aprendidos para expressar importância, ou pesos, para características em V ao calcular a "atenção" final.

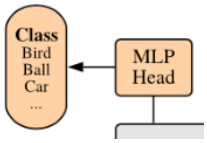
Em seguida, em cada cabeça, o tensor de atenção do produto escalar dimensionado (A) é calculado como uma função *softmax* da multiplicação entre as matrizes Ki e Qi, normalizado com a raiz quadrada da dimensão da cabeça. A linha número i nesta matriz é uma função de distribuição de probabilidade da atenção para direcionar a i, significando para as chaves de quais outros recortes o direcionamento ao recorte i é mais semelhante.

A auto-atenção é o produto entre A e v, que tem a forma $(n+1, d_h)$. O elemento na linha i e coluna j é a média ponderada da característica j pela função de distribuição de probabilidade na linha i em A.

As matrizes de auto-atenção são concatenadas na segunda dimensão, resultando em um tensor $(n+1, d)$, que é então processado por uma única camada linear, efetivamente multiplicando-o por um tensor treinável (d, d). Esta camada linear é muito importante, pois permite que as características sejam aprendidas como agrupamentos de todas as cabeças.

Como mencionado anteriormente, apenas a última representação do token $[cls]$ é utilizada na cabeça de classificação na Figura 14. Para o pré-treinamento, é utilizada uma MLP de 2 camadas, portanto existem duas matrizes de pesos — W_h com formato (d, d_{mlp}) e W_o com formato (d_{mlp}, d). Para o ajuste fino (*fine-tuning*), é utilizada uma única camada linear, portanto existe apenas um único tensor, com formato (d, n_cls). Em cada caso, a saída final da rede é um vetor com formato $(1, n_cls)$, contendo as probabilidades associadas a cada uma das classes.

Figura 14 - ViT: Último passo



Fonte: **(**DOSOVITSKIY *et al.*, 2021**)**

2.2.5 Explicabilidade para modelos de IA

Embora pareçam poderosos em termos de resultados e previsões, os algoritmos de IA sofrem de falta de transparência, são caixas pretas, ou seja, é difícil obter um entendimento sobre o seu mecanismo interno de funcionamento, especialmente com os algoritmos de Inteligência Artificial. Confiar decisões importantes a um sistema que não pode se explicar apresenta perigos óbvios. Para abordar essa questão, a Inteligência Artificial Explicável (XAI) propõe uma mudança em direção a uma IA mais transparente. O objetivo é aplicar um conjunto de técnicas que produzam modelos explicáveis, mantendo os altos níveis de desempenho. O objetivo de possibilitar a explicabilidade ou interpretabilidade em modelos de Inteligência Artificial, é garantir que as decisões do algoritmo, bem como quaisquer dados que impulsionem essas decisões, possam ser explicados aos usuários finais e outros interessados em termos não técnicos (ADADI; BERRADA, 2018).

O modelo de diagnóstico médico é responsável pela vida humana. Mas, é muito difícil ter confiança suficiente para tratar um paciente conforme instruído por um modelo de caixa preta. Em meados da década de 1990, uma rede neural artificial foi treinada para prever quais pacientes com pneumonia deveriam ser admitidos em hospitais e quais poderiam ser tratados como pacientes externos. Descobertas iniciais indicavam que as redes neurais eram muito mais precisas do que os métodos estatísticos clássicos. No entanto, após um teste extensivo, descobriu-se que a rede neural havia inferido que pacientes com pneumonia que também tinham asma apresentavam um risco menor de morte e, portanto, não deveriam ser admitidos. Medicamente, isso é contraintuitivo; no entanto, refletia um padrão real nos dados de treinamento — pacientes com asma e pneumonia geralmente eram admitidos não apenas no hospital, mas diretamente na UTI, recebiam tratamento agressivo e

sobreviviam. Então, decidiu-se abandonar o sistema de IA porque era muito perigoso usá-lo clinicamente. Apenas interpretando o modelo, podemos descobrir um problema tão crucial e evitá-lo (CARUANA *et al.*, 2015).

A explicabilidade é importante. Para construir confiança em sistemas automatizados e avançar em direção à sua integração significativa ao cotidiano, está claro que devemos construir modelos transparentes que tenham a capacidade de explicar por que eles preveem o que preveem. De modo geral, essa transparência e capacidade de explicação são úteis em três diferentes estágios da evolução da IA. Primeiro, quando a IA está errando mais que os humanos e ainda não é confiável para ser implantada, o objetivo da transparência e explicações é identificar porque o modelo está falhando, ajudando assim os pesquisadores a focarem seus esforços no problema. Segundo, quando a IA está acertando e errando tanto quanto os humanos e é confiável para ser implantada, o objetivo é estabelecer confiança e segurança nos usuários. Terceiro, quando a IA acerta muito mais que os humanos, o objetivo das explicações está no ensino de máquina – ou seja, uma máquina ensinando um humano sobre como tomar melhores decisões (SELVARAJU *et al.*, 2019).

Zhou *et al.* (2015) propuseram uma técnica chamada *Class Activation Mapping* (CAM) para identificar regiões discriminativas usadas por uma classe restrita de CNNs de classificação de imagens que não continham camadas totalmente conectadas. Em essência, este trabalho trocava a complexidade do modelo e desempenho por mais transparência no funcionamento do modelo. Selvaraju *et al.* generalizaram o CAM em 2019 para que fosse aplicável a uma gama maior de modelos de CNN, sem exigir mudanças arquitetônicas ou retreinamento. Eles propuseram uma técnica para produzir explicações visuais para decisões de uma ampla classe de modelos de CNN, tornando-os mais transparentes e explicáveis. Essa abordagem foi chamada de *Gradient-weighted Class Activation Mapping* (Grad-CAM). Ela utilizava os gradientes de qualquer alvo fluindo para a última camada convolucional para produzir um mapa de localização grosseiro destacando as regiões importantes na imagem para prever a classe.

3 MATERIAL E MÉTODOS

A metodologia aplicada na pesquisa e desenvolvimento do projeto foi dividida em etapas, sendo elas:

- Etapa 1 – Definição de projeto a partir de revisão bibliográfica de artigos publicados e arquiteturas de modelos de Visão Computacional;
- Etapa 2 – Classificação de imagens de exames de citologia cervical em artigos com seus respectivos laudos e patologias identificadas;
- Etapa 3 – Desenvolvimento de um modelo de classificação de Rede Neural Convolucional (CNN) para fins de comparação como ponto de partida;
- Etapa 4 – Desenvolvimento de um modelo de classificação de Vision Transformer (ViT) para fins de comparação como ponto de partida;
- Etapa 5 – Desenvolvimento de um modelo combinando a arquitetura de Rede Neural Convolucional com Vision Transformer (CNN+ViT);
- Etapa 6 – Gerar métricas de Sensibilidade, Especificidade e AUC (*Area Under the Curve*);
- Etapa 7 – Implementar técnicas de explicabilidade para validar resultados dos modelos visualmente.

3.1 Ferramentas e Tecnologias

Para o desenvolvimento do modelo foram utilizadas as ferramentas e tecnologias citadas a seguir:

- **Python:** linguagem de programação utilizada para escrever o código.
- **PyTorch:** biblioteca Python para a criação de modelos de rede neural.
- **OpenCV:** biblioteca para processamento de imagens e visão computacional.
- **EfficientNet-b4:** modelo pré-treinado de rede neural convolucional.
- **DeiT3-base:** modelo pré-treinado de Vision Transformer.
- **Grad-CAM**: biblioteca com explicabilidade para modelos de visão computacional.
- **Jupyter Notebook:** plataforma interativa de código.

3.2 Conjuntos de Imagens

No início do desenvolvimento de métodos automáticos para análise de exames de citologia cervical, muitos se dedicaram à coleta de imagens de citologia cervical, pois os métodos de análise automática dependiam de grandes quantidades de dados rotulados e existiam poucos conjuntos de dados públicos disponíveis. Esses conjuntos de dados públicos de citologia cervical podem ser utilizados para desenvolver algoritmos de análise automática para múltiplas tarefas, incluindo classificação de imagens, detecção de objetos e segmentação semântica.

Para o desenvolvimento desse projeto foram coletados 4 conjuntos de dados públicos:

1. ***Herlev*** (Jantzen *et al.*, 2005) - Herlev é um dos conjuntos de dados mais utilizado para a análise de citologia cervical, que consiste em 917 imagens cervicais de esfregaços de Papanicolaou divididos em 7 classes (3 classes normais e 4 classes anormais) baseadas na regra de classificação do sistema de displasia de 3 níveis. Todas as imagens são coletadas usando um microscópio conectado a um capturador de quadros com uma resolução de 0,201 μm/pixel no Hospital Universitário de Herlev (Dinamarca). Cada imagem celular é segmentada manualmente em fundo, citoplasma e núcleo para posterior extração de características. As imagens possuem formatos variados em pixels como 71 x126 ou 76 x 106 ou 69 x 70, não seguem um padrão, sugerindo que foram recortadas manualmente, porém mantendo suas dimensões próximas a isso.
2. ***SIPaKMeD*** (Plissiti *et al.*, 2018) - Esta base de dados consiste em 4049 imagens de células cervicais isoladas, que foram adquiridas da Universidade de Ioannina, Grécia, através de uma câmera CCD adaptada a um microscópio óptico (OLYMPUS BX53F). As células são anotadas por citopatologistas experientes divididas em cinco classes (superficial-intermediária, parabasal, coilocitótica, disceratótica e metaplásica), dependendo da sua aparência citológica e morfologia. Entre essas cinco classes, as células superficial-intermediária e parabasal são normais. Coilocitóticas (LSIL) e disceratóticas (HPV) são células anormais, mas não malignas, enquanto metaplásicas pertencem ao grupo benigno. Em cada imagem da base de dados SIPaKMeD, as áreas do

citoplasma e do núcleo são definidas manualmente e as células são recortadas em dimensões muito próximas às imagens do conjunto de imagens Herlev. O conjunto de imagens também possui as imagens originais não recortadas em dimensões de 2048×1536 pixels.

3. ***CRIC*** (Rezende *et al.*, 2021) - O conjunto de dados CRIC possui 400 imagens de esfregaços cervicais convencionais e 11.534 células classificadas. Os esfregaços de Papanicolau são coletados de 118 pacientes do sexo feminino na região Sudeste do Brasil e preparados e analisados no Laboratório de Citologia da Escola de Farmácia, Universidade Federal de Ouro Preto, Minas Gerais, Brasil. Todas as imagens são capturadas por microscopia de campo claro convencional com um objetivo de 40× e uma ocular de 10×, utilizando uma câmera digital Zeiss AxionCam MRc acoplada ao Zeiss AxioImager. O tamanho da imagem é de 1376 × 1020 e a resolução é de 0,228 μm/pixel. A coleção CRIC abrange seis classes baseadas na nomenclatura do sistema Bethesda (Bethesda, 2015): NILM (6779), ASC-US (606), LSIL (1360), ASC-H (925), HSIL (1703) e SCC (161). As imagens possuem dimensões de 1376×1020 pixels.
4. ***Mendeley LBC*** (Hussain *et al.*, 2020) - Este conjunto de dados possui um total de 460 espécimes de três centros de diagnóstico médico na Índia, incluindo Babina Diagnostic em Imphal, a escola médica e hospital Gauhati em Guwahati e o Instituto do Câncer Dr. B. Barooah em Guwahati. O conjunto de dados consiste em um total de 963 imagens de citologia em base líquida (LBC) capturadas por um microscópio Leica ICC50 HD coletadas com ampliação de 400× (lente objetiva de 40× e ocular de 10×). O tamanho de cada imagem é de 2048 × 1536. As imagens foram subdivididas em quatro categorias pelo sistema Bethesda: NIML (613), LSIL (163), HSIL (113) e SCC (74). As imagens possuem dimensões de 2048×1536 pixels.

Mais imagens avulsas de exames de citologia cervical foram coletadas do site da Organização Mundial de Saúde (IARC, 2004) e do sistema Bethesda (BETHESDA, 2015), totalizando 531 imagens divididas em 3 classes: uma classe normal e duas anormais que poderão ser utilizadas para testes de avaliação dos

modelos. As imagens coletadas na internet possuem formatos variados mantendo suas dimensões próximas a 500×375 pixels.

3.3 Etapa 1 – Definição de projeto

Segundo a revisão da literatura, a combinação de modelos de rede neural convolucional (CNN) com modelos de *Vision Transformer* traz a vantagem de trabalhar tanto com a informação global quanto a local, o que resolveria problemas já existentes nesses modelos como perda de contexto em CNN ou perda de foco em ViT, melhorando sua performance também em imagens de não tão boa qualidade ou fora de padrão. Pensando nisso, para a construção do modelo inicialmente, a arquitetura apresentada por Liu *et al.* (2022) foi seguida como referência, com a adição de Grad-CAM para mapear a explicabilidade do modelo.

3.4 Etapa 2 – Classificação de imagens

Para a classificação das imagens foram consideradas duas classes para otimização de performance do modelo: normal e anormal, utilizando grupos de células dos 4 conjuntos de dados para o treinamento dos modelos. Para a classe anormal foram consideradas as lesões pré-cancerosas do colo uterino classificadas como lesões intraepiteliais escamosas de baixo grau (compreendendo as alterações pelo HPV e a displasia leve/NIC 1) e lesões intraepiteliais escamosas de alto grau (correspondendo à displasia moderada/ NIC 2 e carcinoma in situ/NIC 3). Segue a relação ao se dividir as imagens dos conjuntos de dados nessas duas classes:

1. ***Herlev*** - Recortes de células individuais.

Normal: Epitélio escamoso superficial e intermediário, epitélio colunar

Anormal: Displasia escamosa leve, moderada e grave não queratinizante, carcinoma de células escamosas in situ intermediário

2. ***SIPaKMeD*** - Recortes de células individuais e grupo de células.

Normal: Superficial-intermediária, parabasal

Anormal: Coilocitótica, disceratótica

Benigna: Metaplásica (essa categoria será incluída como normal por não apresentar risco à saúde)

3. ***CRIC*** - Grupo de células.

Normal: NILM

Anormal: ASC-H, ASC-US, HSIL, LSIL, SCC

4. ***Mendeley LBC*** - Grupo de células.

Normal: NILM

Anormal: HSIL, LSIL, SCC

Sendo que imagens de recortes de células individuais só podem ser classificadas em conjunto com imagens do mesmo tipo (Herlev, SIPaKMeD), e imagens de células agrupadas só podem ser classificadas em conjunto com imagens de células agrupadas (SIPaKMeD, CRIC, Mendeley LBC) totalizando 2.648 imagens. As imagens do conjunto de Mendeley LBC pertencem a um outro tipo de exame, citologia de base líquida, que difere do exame convencional de Papanicolau. Elas serão utilizadas para verificar a capacidade do modelo detectar anormalidades mesmo com a variação de padrão de imagens e exames de citologia.

As imagens foram todas formatadas para as dimensões de 384 x 384 pixels para servir para o modelo e foram divididas em um conjunto de treino e validação, sendo 80% das imagens reservadas para treinamento do modelo e 20% das imagens reservadas para teste e refinamento do modelo. Além disso imagens avulsas que não tiveram contato com o modelo, coletadas na web, foram utilizadas para validar a capacidade do modelo de classificar imagens que nunca foram visualizadas.

3.5 Etapa 3 – CNN

Um modelo CNN pré-treinado com o conjunto de mais de um milhão de imagens da ImageNet (KRIZHEVSKY *et al.*, 2012) foi selecionado observando-se as melhores métricas em estudos previamente publicados na classificação de imagens para o conjunto de imagens de citologia cervical do SIPaKMeD. O modelo de rede neural convolucional EfficientNet-b4 foi o que apresentou as melhores métricas de Acurácia, Precisão, Recall e F1-score. Porém modelos como VGG13, DenseNet169,

MobileNetv3-Large, RexNet-300, entre outros, também apresentaram métricas semelhantes (PACAL, 2024).

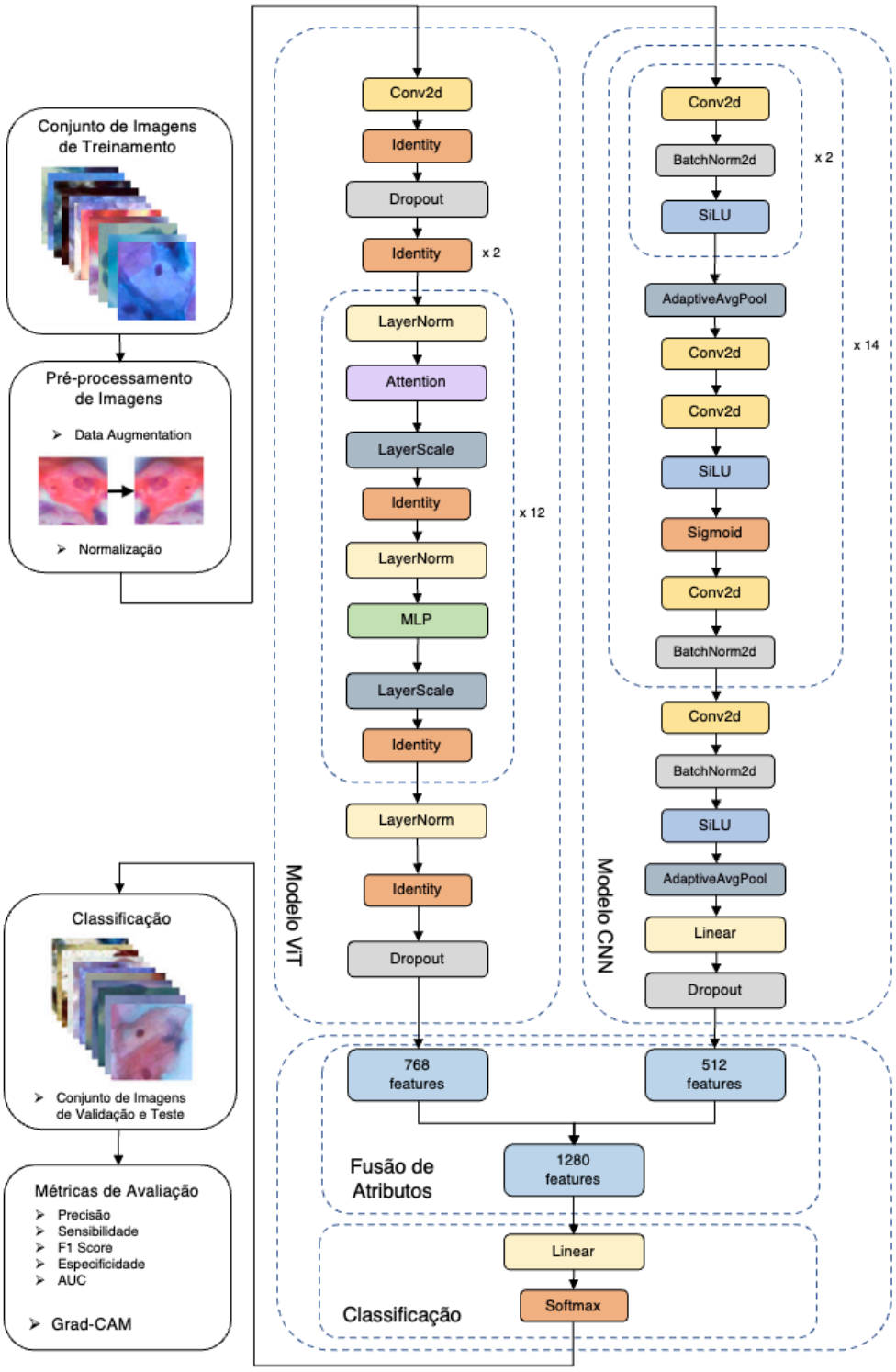
3.6 Etapa 4 – ViT

Seguindo a mesma estratégia um modelo de Vision Transformer foi selecionado também treinado com o conjunto de imagens da ImageNet. O modelo DeiT3-Base foi o que apresentou as melhores métricas de Acurácia, Precisão, Recall e F1-score no conjunto de imagens de citologia cervical do SIPaKMeD. Outros modelos também apresentaram boa performance como ViT-Large-Patch32 e modelos Swin e MaxViT (PACAL, 2024).

3.6 Etapa 5 – CNN + ViT

Os modelos CNN e ViT foram então combinados utilizando como referência a arquitetura do modelo CVM-Cervix (LIU et al., 2022) para extrair características locais e globais e obter melhor performance do modelo na classificação de imagens de exames de citologia cervical em categorias normais e anormais. Na figura 15 segue a representação da arquitetura e das camadas utilizadas na fusão dos modelos.

Figura 15 - Fusão dos Modelos: CNN + ViT



Fonte: Elaborado pela própria autora.

3.7 Etapa 6 – Métricas

Métricas foram utilizadas para averiguar qualidade do modelo. Primeiramente ao se treinar o modelo, uma otimização é realizada para maximizar a acurácia e ao mesmo tempo minimizar o valor de perda. Esses valores podem variar de acordo com a inicialização aleatória dos pesos ou com o número de épocas em que o modelo foi treinado, depende das configurações utilizadas para treinar o modelo. Ao encontrar os valores ótimos, o modelo está pronto para ser avaliado.

Para analisar a performance e a qualidade do modelo, outras métricas são utilizadas como matriz de confusão, Precisão, Recall e F1-score.

A métrica de Precisão na equação 9 calcula o total de positivos verdadeiros estimados em relação a todos os positivos estimados. É usada para mensurar o impacto dos falsos positivos.

$$Precisão = \frac{Verdadeiros\ Positivos}{Verdadeiros\ Positivos + Falsos\ Positivos} \quad (9)$$

A métrica de Recall ou Sensibilidade na equação 10 calcula quantos positivos reais o modelo estima em relação a todas as estimativas de cada classificação. É usada para mensurar o impacto dos falsos negativos.

$$Sensibilidade = \frac{Verdadeiros\ Positivos}{Verdadeiros\ Positivos + Falsos\ Negativos} \quad (10)$$

A métrica F1-score na equação 11 junta as métricas de Precisão e Recall, gerando um equilíbrio entre as duas métricas. Ela mensura o impacto dos falsos positivos e dos falsos negativos estimados pelo modelo, focando nos erros das classificações do modelo.

$$F1\ Score = 2 * \frac{Precisão * Sensibilidade}{Precisão + Sensibilidade} \quad (11)$$

As métricas de Sensibilidade e Especificidade foram priorizadas por serem consideradas altamente relevantes na área médica por fornecerem informações cruciais sobre o desempenho de testes diagnósticos, procedimentos de triagem e modelos preditivos em termos de sua capacidade de identificar corretamente as condições de saúde.

A Especificidade de um teste na equação 12 é a sua capacidade de identificar corretamente os indivíduos que não têm a doença (verdadeiros negativos). Um teste com alta especificidade é importante para evitar o diagnóstico incorreto de pessoas

saudáveis como doentes (falsos positivos). Isso é especialmente importante para evitar tratamentos desnecessários ou invasivos e a ansiedade associada a um diagnóstico incorreto.

$$Especificidade = \frac{Verdadeiros\ Negativos}{Falsos\ Positivos + Verdadeiros\ Negativos} \quad (12)$$

Também utilizada, a métrica AUC (*Area Under the Curve*), se refere à precisão do modelo. Um modelo cujas previsões estão 100% erradas tem uma AUC de 0, enquanto um modelo cujas previsões são 100% corretas tem uma AUC de 1. A curva AUC é derivada da curva ROC (*Receiver Operating Characteristic*) que utiliza uma relação entre a taxa de verdadeiro positivo na equação 13 e a taxa de falso positivo na equação 14.

$$Taxa\ de\ Verdadeiro\ Positivo = \frac{Verdadeiros\ Positivos}{Verdadeiros\ Positivos + Falsos\ Negativos} \quad (13)$$

$$Taxa\ de\ Falso\ Positivo = \frac{Falsos\ Positivos}{Falsos\ Positivos + Verdadeiros\ Negativos} \quad (14)$$

3.8 Etapa 7 – Explicabilidade

Para explicabilidade dos modelos de Visão Computacional, foi empregado um pacote com os métodos do estado da arte em Pytorch, o Grad-CAM. O Grad-CAM pesa as ativações em duas dimensões pelo gradiente médio, destacando o foco e a área de interesse que gerou uma classificação específica.

Ao verificar que o modelo está de fato destacando a área relevante para sua respectiva classificação, o aprendizado do modelo pode ser validado.

4 RESULTADOS

4.1 Modelo CNN

O modelo pré-treinado de Rede Neural Convolucional EfficientNet-b4 foi configurado e retreinado para classificar imagens de citologia cervical em normal ou anormal, uma classificação binária. Ele foi treinado por 15 épocas com 2117 imagens para treinamento e 531 imagens para validação. As imagens passaram por um pré-processamento com redimensionamento e *Data Augmentation*, uma técnica para gerar novos exemplares de dados de treinamento a fim de melhorar a generalização do modelo.

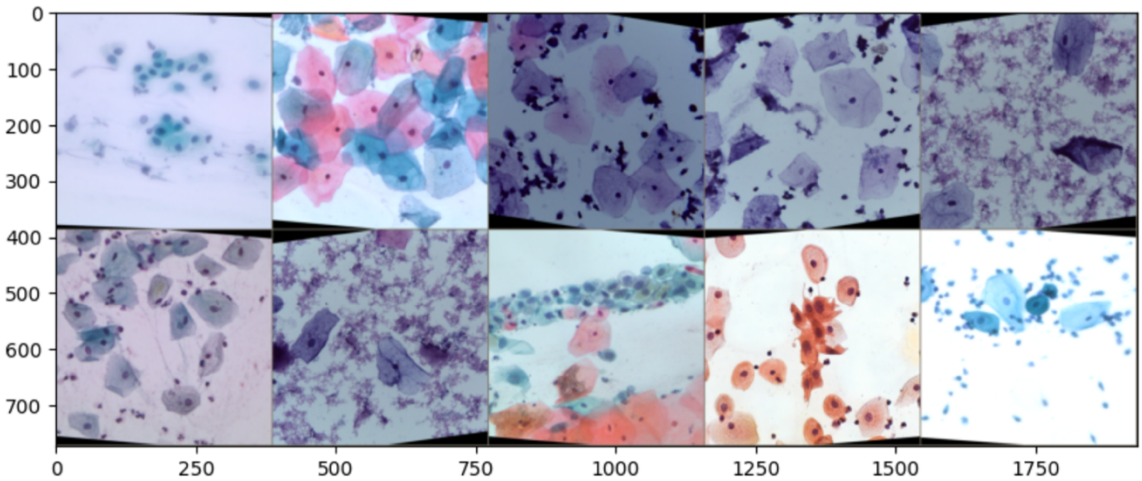
Os seguintes processos, Figura 16, foram utilizados em *Data Augmentation*:

- Rotação aleatória: rotacionando a imagem mais ou menos 10 graus;

- Giro horizontal aleatório: invertendo 50% das imagens;

- Normalização: modelos pré-treinados com imagens da ImageNet possuem parâmetros recomendados de média = [0.485, 0.456, 0.406] e desvio padrão = [0.229, 0.224, 0.225].

Figura 16 - *Data Augmentation*

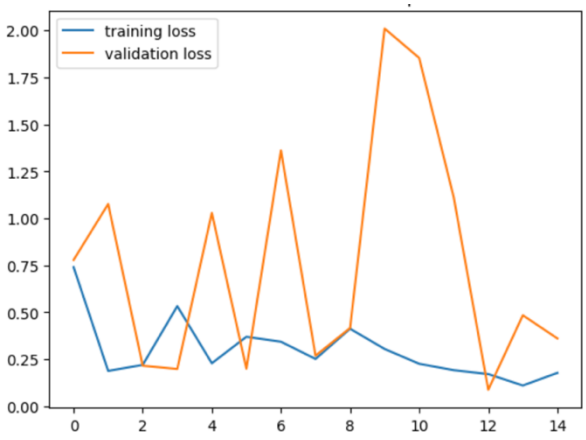


Fonte: Elaborado pela própria autora.

Ao final das 15 épocas, levando em torno de 13h 50min, o modelo na iteração que apresentava a melhor acurácia foi salvo para ser reutilizado na avaliação. Para treinamento do modelo foi utilizado um MacBook Pro M1 com 32 GB de memória

RAM e placa de vídeo integrada conhecida por superar o desempenho de algumas placas de vídeo da Nvidia e Radeon. Métricas de acompanhamento foram geradas ao longo do treinamento do modelo. As imagens a seguir mostram a comparação da performance de classificação dos conjuntos de imagem de treino e validação durante o treinamento. Na Figura 17, o erro que se mantinha entre 0,5% e 1,5%, dispara na nona e décima iteração se perdendo, mas logo retorna à base. Sua melhor performance é atingida na décima terceira época.

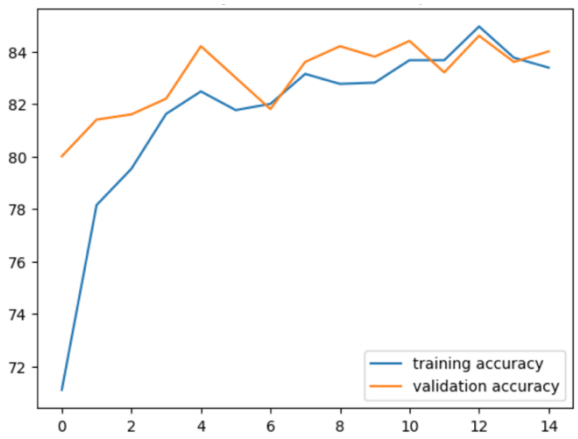
Figura 17 - CNN: Erro por época



Fonte: Elaborado pela própria autora.

Na Figura 18 é possível acompanhar a curva de aprendizado do modelo a partir da classificação das imagens do conjunto de treinamento. O modelo atinge sua melhor performance quando as métricas de acerto do conjunto de treino e do conjunto de validação chegam ao mesmo nível, o que significa que o modelo está aprendendo e generalizando. O modelo pode então continuar a ser treinado, e pode encontrar grande variação, melhorando ou piorando, mas esse experimento foi limitado a 15 épocas por questão de tempo.

Figura 18 - CNN: Acurácia por época



Fonte: Elaborado pela própria autora.

4.1.1 Modelo CNN: Métricas de Avaliação

Para avaliar a qualidade de cada modelo treinado foram utilizados dois conjuntos de imagens. O primeiro conjunto de imagens é formado por dados públicos SIPaKMeD, CRIC, Mendeley LBC, também utilizado no treinamento dos modelos (Plissiti *et al.*, 2018; Rezende *et al.*, 2021; Hussain *et al.*, 2020). O segundo conjunto é constituído de imagens coletadas do site da Organização Mundial de Saúde (IARC, 2004) e do sistema Bethesda (BETHESDA, 2015).

Imagens de validação utilizadas no treino

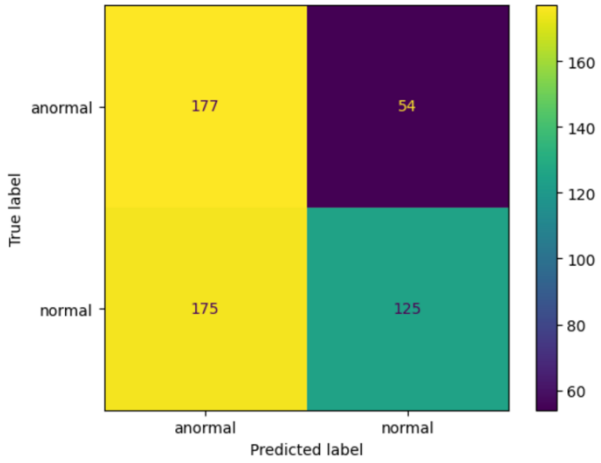
531 imagens:

- 231 imagens classificadas como anormais (“positivo”)
- 300 imagens classificadas como normais (“negativo”)

Ou seja, se uma imagem é “anormal” e é classificada como “anormal”, é um “Verdadeiro Positivo”. Se ela é “anormal”, mas foi classificada incorretamente como “normal”, é um “Falso Negativo”. Da mesma forma, se uma imagem é “normal” e é classificada como “normal”, é um “Verdadeiro Negativo”. Se ela é “normal”, mas foi classificada incorretamente como “anormal”, é um “Falso Positivo”.

A Matriz de Confusão mostra que a maioria das imagens está sendo classificada como “anormal” e a maior parte das imagens com diagnóstico “normal” está sendo erroneamente classificada como “anormal”, Figura 19. O modelo está enviesado, tendendo a classificar todas as imagens como anormais.

Figura 19 - CNN Validação: Matriz de Confusão



Fonte: Elaborado pela própria autora.

Na Tabela 1, o erro dos Falsos Positivos pode ser visualizado pela métrica de Sensibilidade de 0,42, então mais da metade das imagens de classificação “normal” está sendo classificada como “anormal”. O que torna esse erro mais grave é que a maioria das imagens originalmente deveria ser classificada como “normal”, considerando que são 300 imagens com diagnóstico “normal” e apenas 231 como diagnóstico anormal.

Tabela 1 - CNN Validação: Métricas por classe

Classificação	**Precisão**	**Sensibilidade**	**F1 Score**	**Qtd. Imagens**
Anormal:	0,50	0,77	0,61	231
Normal:	0,70	0,42	0,52	300

Fonte: Elaborado pela própria autora.

A Tabela 2 ilustra que o modelo é razoavelmente bom em classificar os Verdadeiros Positivos (diagnóstico “anormal”) com a métrica de Sensibilidade de 0,766, mas muito ruim em classificar os Verdadeiros Negativos (diagnóstico “normal”) com Especificidade de 0,417.

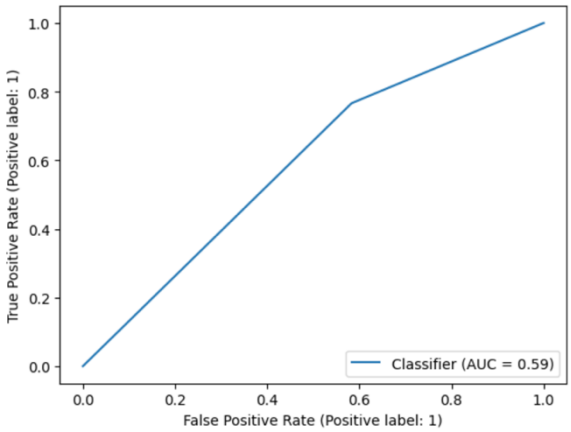
Tabela 2 - CNN Validação: Métricas

Métricas de Avaliação	
Especificidade	0,417
Sensibilidade	0,766
Acurácia	0,569
Precisão	0,503
F1 Score	0,607
Acurácia Balanceada	0,591
AUC	0,59

Fonte: Elaborado pela própria autora.

A Figura 20 mostra que a performance do modelo é acima da média, mas ainda assim muito próximo à média com uma métrica AUC de 0,59 ou 59%. Quanto mais próximo de 50% em uma classificação binária, a probabilidade do modelo errar ou acertar se torna próxima à chance de se tirar cara ou coroa ao se jogar uma moeda.

Figura 20 - CNN Validação: Área sob a curva ROC (AUC)



Fonte: Elaborado pela própria autora.

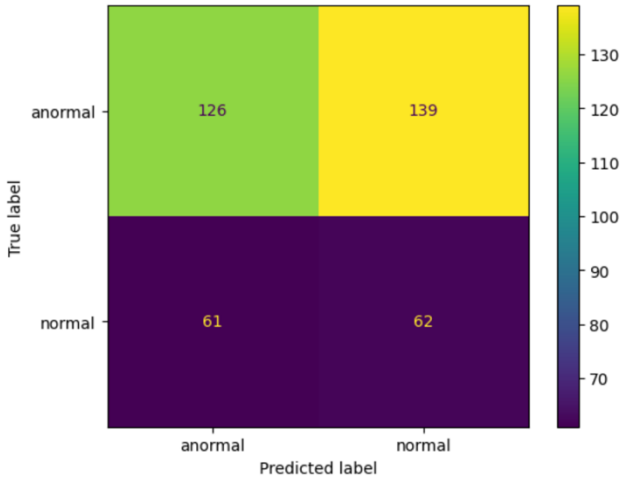
Imagens de teste reservadas fora do treino

388 imagens:

- 265 imagens classificadas como anormais (“positivo”)
- 123 imagens classificadas como normais (“negativo”)

A Matriz de Confusão na Figura 21 mostra que ao classificar as imagens que foram deixadas fora do treino, o modelo erra ou acerta a classificação quase na mesma proporção, sendo que ele erra levemente a mais do que acerta classificando imagens de diagnóstico “anormal” como “normais”.

Figura 21 - CNN Teste: Matriz de Confusão



Fonte: Elaborado pela própria autora.

Na Tabela 3 pode-se perceber que as métricas se encontram muito baixas, especialmente para a classificação de imagens com diagnóstico “normal”, muitas imagens de diagnóstico “anormal” foram erroneamente classificadas como “normais”. O modelo está tendendo a classificar as imagens como “normais”, mesmo havendo mais imagens de diagnóstico “anormal”, ele se comporta de forma completamente oposta ao classificar imagens desse conjunto de imagens de teste em comparação ao conjunto de imagens de validação com o qual o modelo foi treinado.

Tabela 3 - CNN Teste: Métricas por classe

Classificação	**Precisão**	**Sensibilidade**	**F1 Score**	**Qtd. Imagens**
Anormal:	0,67	0,48	0,56	265
Normal:	0,31	0,50	0,38	123

Fonte: Elaborado pela própria autora.

A Tabela 4 também mostra métricas muito baixas atestando pela baixa qualidade de classificação do modelo, todas muito próximas a 50%.

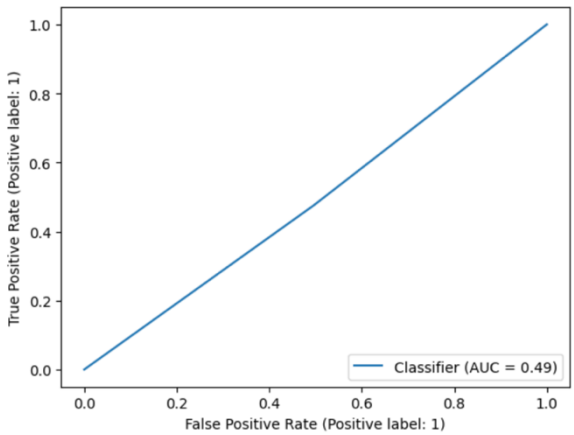
Tabela 4 - CNN Teste: Métricas

Métricas de Avaliação	
Especificidade	0,504
Sensibilidade	0,475
Acurácia	0,485
Precisão	0,674
F1 Score	0,558
Acurácia Balanceada	0,490
AUC	0,49

Fonte: Elaborado pela própria autora.

A Figura 22 mostra a performance do modelo levemente abaixo da média, o que significa que modelo tende a errar mais do que acertar na classificação das imagens do conjunto de teste.

Figura 22 - CNN Teste: Área sob a curva ROC (AUC)



Fonte: Elaborado pela própria autora.

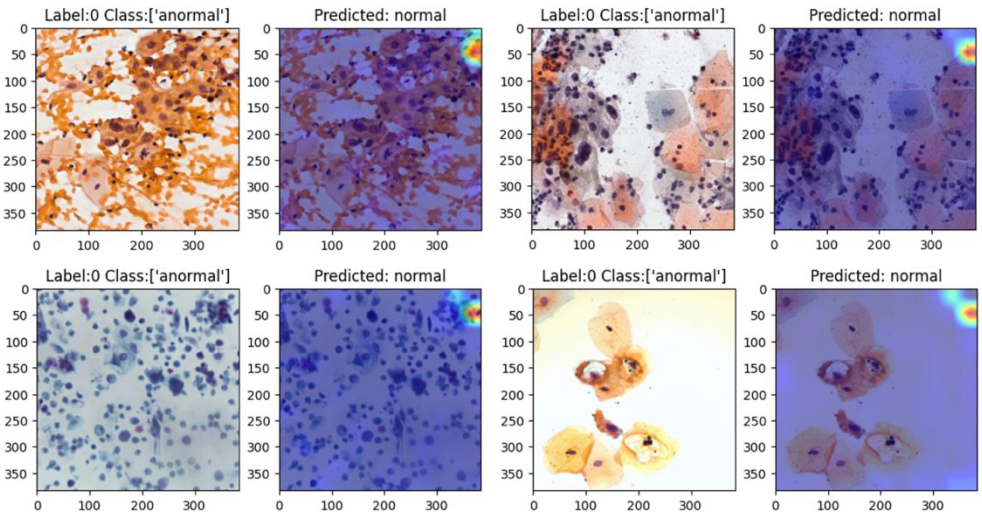
4.1.2 Modelo CNN: Explicabilidade

Com a biblioteca Grad-CAM é possível localizar na imagem e destacar as áreas de interesse, às quais o modelo atribui o maior peso para classificar uma imagem com o diagnóstico “normal” ou “anormal”. Dessa forma pode-se visualizar se o modelo está errando ou até entender por que ele está errando. A seguir as imagens originais são acompanhadas de uma cópia delas com as partes da imagem utilizadas para a classificação do modelo destacadas em laranja e amarelo.

Imagens de validação utilizadas no treino

Na Figura 23 pode-se ver a atenção do modelo voltada ao canto superior direito das imagens ao classificar. Nessa área não há nenhuma informação de interesse para classificar a imagem com diagnóstico “normal”, nem mesmo “anormal”. Ao utilizar Grad-CAM na rede CNN percebeu-se que o modelo estava constantemente focando neste canto da imagem ao invés de focar nas células, na maioria das vezes errando o diagnóstico, e acertando algumas poucas vezes. Isso demonstrava que o modelo não estava devidamente treinado, e estava gerando classificações aleatórias.

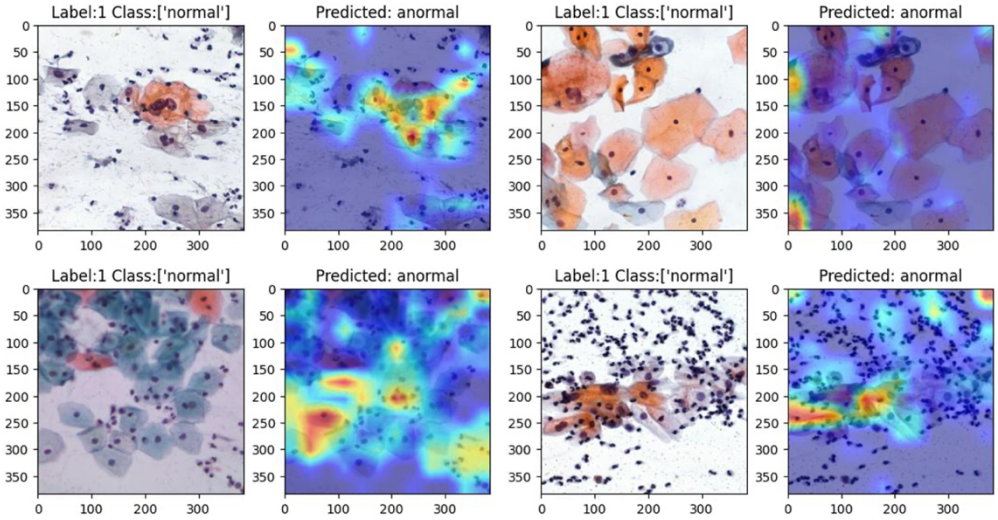
Figura 23 - Classificações de Validação incorretas: Falsos Negativos (Total: 54)



Fonte: Plissiti *et al.*, 2018; Rezende *et al.*, 2021; Hussain *et al.*, 2020; própria autora.

Na Figura 24 percebe-se o modelo voltando sua atenção para áreas importantes da imagem, mas essas imagens com diagnóstico “normal” possuem características muito similares às imagens de diagnóstico “anormal”, por sua tonalidade e excesso de células. Isso acaba por confundir o modelo que classifica as imagens erroneamente como “anormais”.

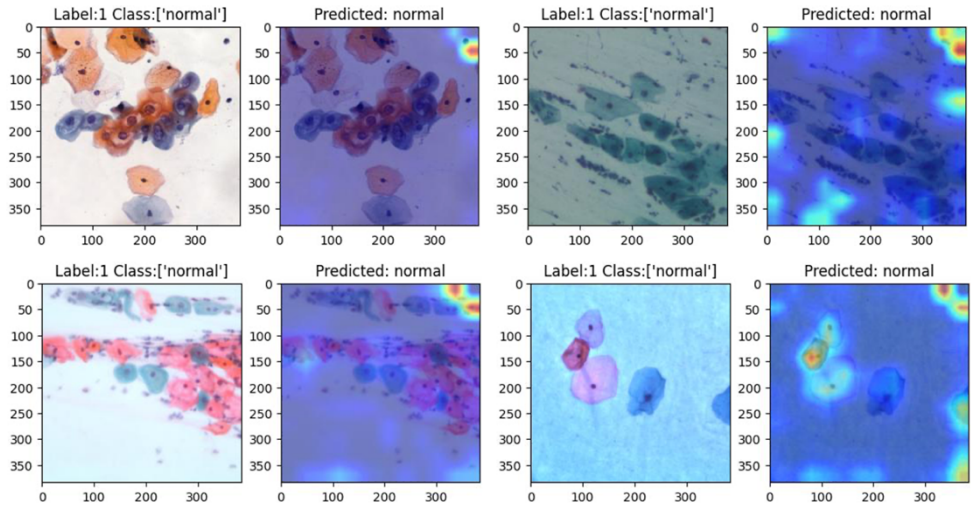
Figura 24 - Classificações de Validação incorretas: Falsos Positivos (Total: 175)



Fonte: Plissiti *et al.*, 2018; Rezende *et al.*, 2021; Hussain *et al.*, 2020; própria autora.

Na Figura 25, novamente pode-se ver o modelo voltando sua atenção para o canto superior direito onde não há informação de interesse para classificar a imagem. Ele gera as classificações aleatoriamente, mas por sorte consegue acertar. Já na quarta classificação abaixo à direita, o modelo destaca áreas mais importantes da imagem como as células, mas ainda assim sua atenção continua fortemente voltada ao canto superior direito, indicando que essa classificação possa ter sido também aleatória.

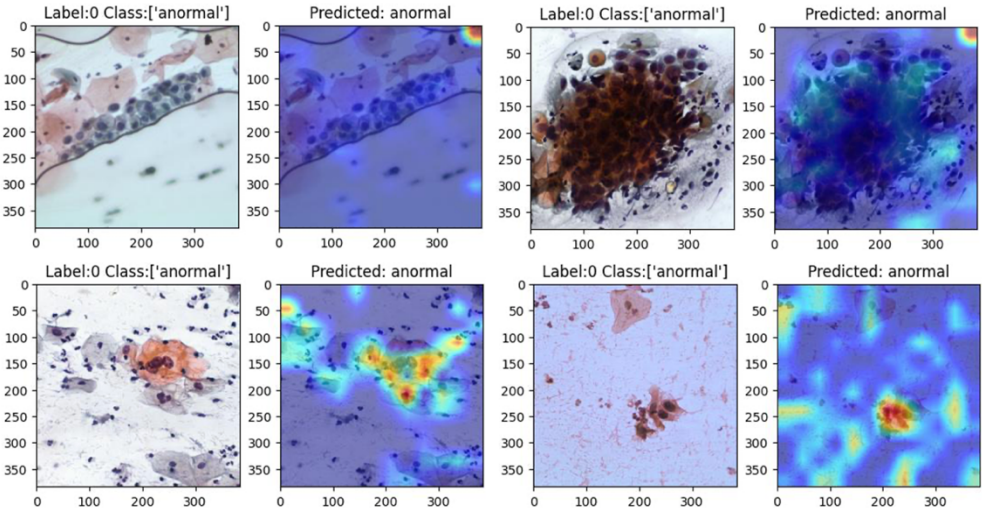
Figura 25 - Classificações de Validação corretas: Verdadeiros Negativos (Total: 125)



Fonte: Plissiti *et al.*, 2018; Rezende *et al.*, 2021; Hussain *et al.*, 2020; própria autora.

Na Figura 26, pode-se ver novamente o modelo voltando sua atenção para o canto superior direito, indicando que a classificação tenha sido aleatória, apesar de correta. Já na terceira e quarta classificação, o modelo destaca áreas mais importantes das imagens e não foca mais tanto no canto superior direito, indicando que essa classificação foi acertada, o modelo estava olhando as características importantes para classificar a imagem corretamente. Verificou-se que isso ocorreu mais frequentemente nas imagens com diagnóstico anormal.

Figura 26 - Classificações de Validação corretas: Verdadeiros Positivos (Total: 177)

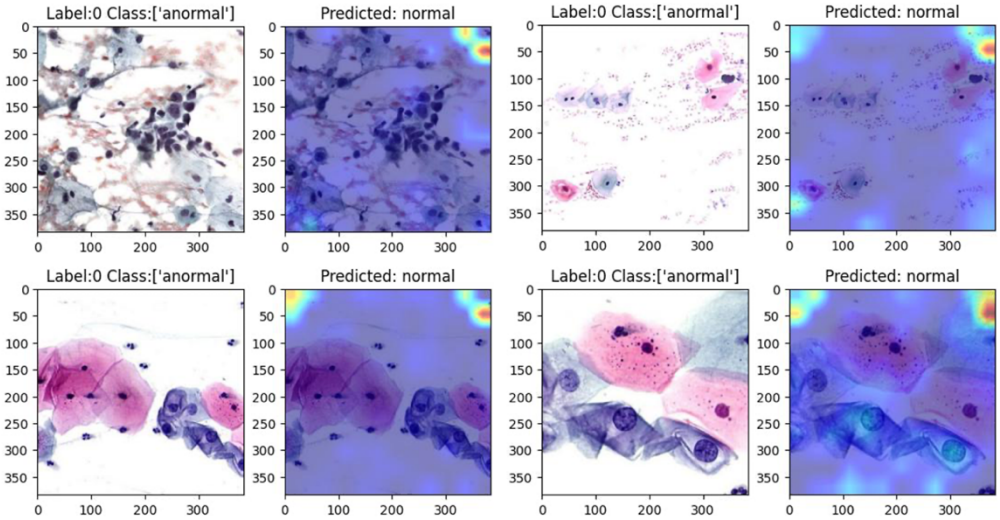


Fonte: Plissiti *et al.*, 2018; Rezende *et al.*, 2021; Hussain *et al.*, 2020; própria autora.

Imagens de teste reservadas fora do treino

Na Figura 27, assim como no conjunto de imagens de validação, as imagens de teste que foram deixadas fora do treino do modelo são incorretamente classificadas. O modelo foca no canto superior direito e às vezes no canto superior esquerdo, onde não há informação relevante para a classificação da imagem, indicando que o modelo está classificando de maneira aleatória. E ele erra ao classificar a maioria das imagens originalmente de diagnóstico “anormal”.

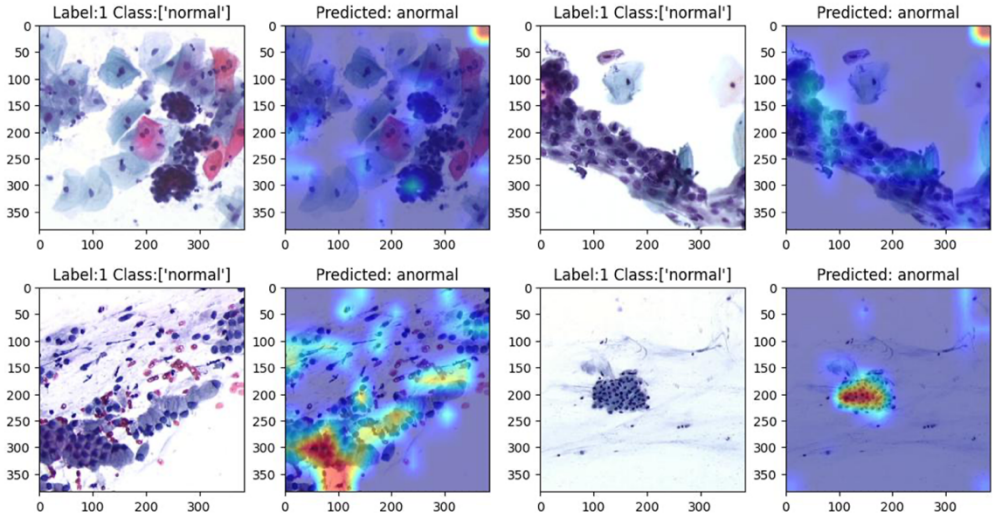
Figura 27 - Classificações de Teste incorretas: Falsos Negativos (Total: 139)



Fonte: IARC, 2004; Bethesda, 2015; própria autora.

Na Figura 28, o modelo classifica erroneamente imagens de diagnóstico “normal” como “anormal”, e como anteriormente ele foca no canto superior direito da imagem quando sua classificação é aleatória. Mas, também há casos em que o modelo confunde imagens de diagnóstico “normal” com “anormal” pela similaridade de características, como as tonalidades e agrupamentos densos de células.

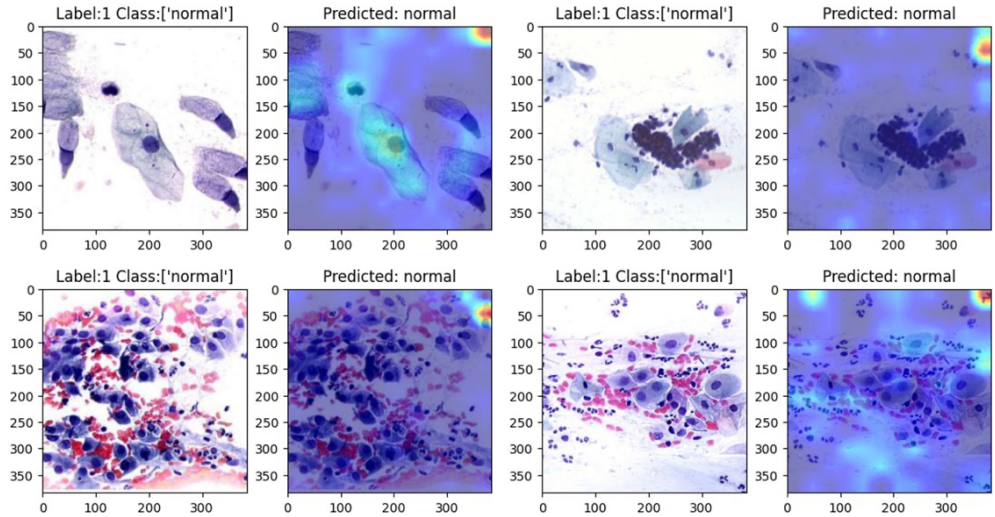
Figura 28 - Classificações de Teste incorretas: Falsos Positivos (Total: 61)



Fonte: IARC, 2004; Bethesda, 2015; própria autora.

Nos exemplos da Figura 29, pode-se ver que o modelo foca algumas poucas vezes nas áreas relevantes para a classificação da imagem, mas na maioria das vezes foca no canto superior direito, indicando que a classificação foi aleatória.

Figura 29 - Classificações de Teste corretas: Verdadeiros Negativos (Total: 62)

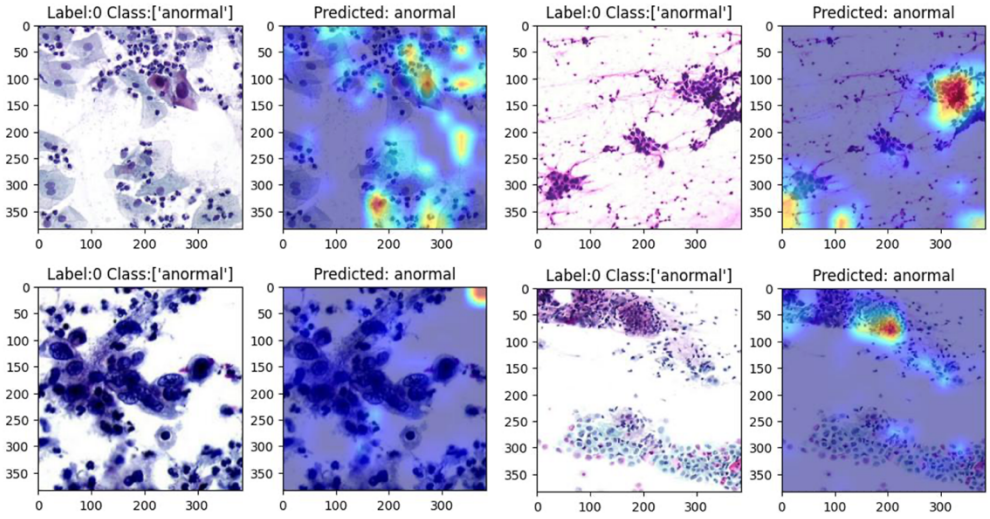


Fonte: IARC, 2004; Bethesda, 2015; própria autora.

Nos exemplos da Figura 30, percebe-se que o modelo foca em áreas relevantes da imagem para gerar sua classificação, mas algumas vezes ele ainda

se perde ao gerar classificações aleatórias, focando no canto superior direito da imagem onde não há informação relevante, acertando por sorte.

Figura 30 - Classificações de Teste corretas: Verdadeiros Positivos (Total: 126)



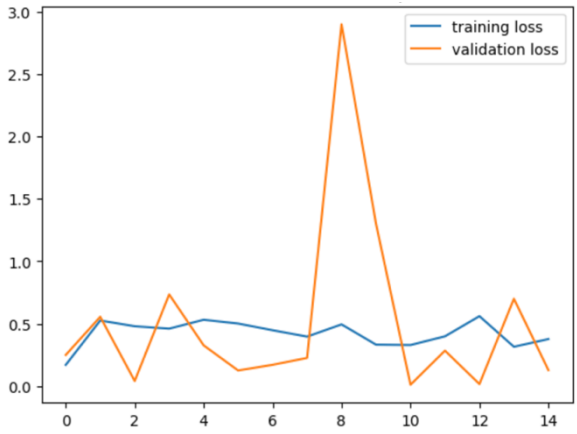
Fonte: IARC, 2004; Bethesda, 2015; própria autora.

4.2 Modelo ViT

O modelo pré-treinado de Vision Transformer DeiT3-base foi configurado e retreinado para classificar imagens de citologia cervical em normal ou anormal. Assim como o modelo anterior, ele foi treinado por 15 épocas com 2117 imagens para treinamento e 531 imagens para validação, e passou pelo mesmo pré-processamento.

Ao final das 15 épocas, levando em torno de 12h 42min, o modelo na iteração que apresentava a melhor acurácia foi salvo para ser reutilizado na avaliação. Métricas de acompanhamento foram geradas ao longo do treinamento do modelo. As imagens a seguir mostram a comparação da performance de classificação dos conjuntos de imagem de treino e validação durante o treinamento. Na Figura 31, o erro que se mantinha entre 0,5% e 1,0%, sofre um desvio se perdendo na oitava época, mas se recupera. Sua melhor performance é atingida na décima época.

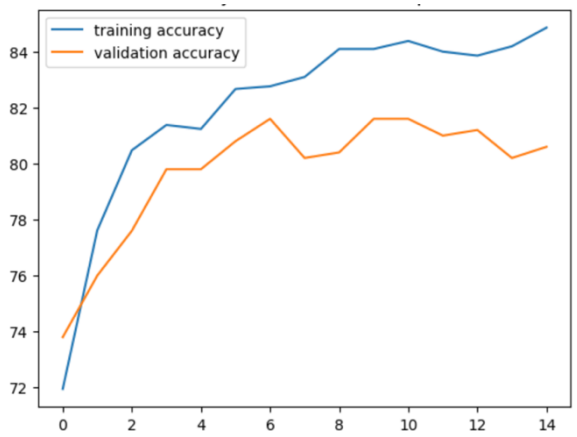
Figura 31 - ViT: Erro por época



Fonte: Elaborado pela própria autora.

Na Figura 32 pode-se observar a curva de aprendizado crescente do modelo e sua tendência a *overfitting*, quanto mais ele é treinado mais ele pode viciar no conjunto de imagens de treino perdendo sua capacidade de generalizar, classificando incorretamente imagens do conjunto de validação. Técnicas de regularização podem suavizar esse problema, com as camadas de *Dropout* o modelo treinado é forçado a esquecer alguns dos pesos atribuídos.

Figura 32 - ViT: Acurácia por época



Fonte: Elaborado pela própria autora.

4.2.1 Modelo ViT: Métricas de Avaliação

Novamente são utilizados os dois conjuntos de imagens para avaliar a qualidade do modelo de *Vision Transformer*. O primeiro conjunto de imagens é formado por dados públicos SIPaKMeD, CRIC, Mendeley LBC, também utilizado no treinamento dos modelos (Plissiti *et al.*, 2018; Rezende *et al.*, 2021; Hussain *et al.*, 2020). O segundo conjunto é constituído de imagens coletadas do site da Organização Mundial de Saúde (IARC, 2004) e do sistema Bethesda (BETHESDA, 2015).

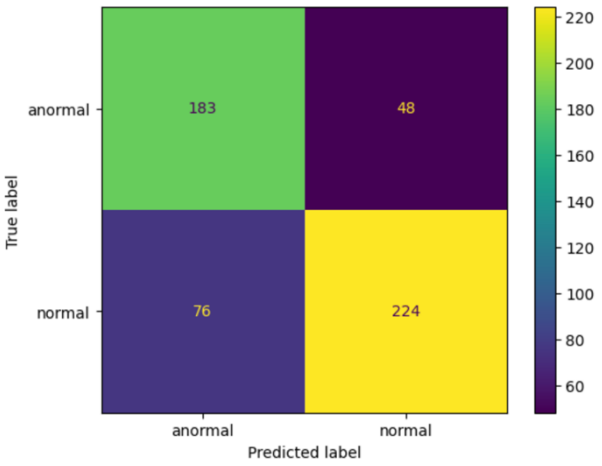
Imagens de validação utilizadas no treino

531 imagens:

- 231 imagens classificadas como anormais (“positivo”)
- 300 imagens classificadas como normais (“negativo”)

A Matriz de Confusão na Figura 33 mostra uma melhora considerável em relação ao modelo CNN treinado anteriormente, com a maioria das imagens de diagnóstico “anormal” sendo corretamente classificada como “anormal”, e a maioria das imagens de diagnóstico “normal” sendo corretamente classificada como “normal”.

Figura 33 - ViT Validação: Matriz de Confusão



Fonte: Elaborado pela própria autora.

Na Tabela 5, também há uma melhora considerável nas métricas, com todas elas se mantendo acima de 70%. As métricas por classe demonstram a capacidade do modelo de classificar corretamente imagens de diagnóstico “normal” com uma precisão de 0,82 ou 82%. Demonstra também que o número de imagens de diagnóstico “anormal” classificadas erroneamente como “normal” está muito baixo com uma métrica de Sensibilidade de 0,79.

Tabela 5 - ViT Validação: Métricas por classe

Classificação	Precisão	Sensibilidade	F1 Score	Qtd. Imagens
Anormal:	0,71	0,79	0,75	231
Normal:	0,82	0,75	0,78	300

Fonte: Elaborado pela própria autora.

A Tabela 6 mostra novamente que as métricas de classificação do modelo se mantiveram acima de 70%, com destaque para Sensibilidade de 0,792 demonstrando que o modelo não está deixando muitas imagens com diagnóstico positivo para anormalidades serem classificadas erroneamente como saudáveis.

Tabela 6 - ViT Validação: Métricas

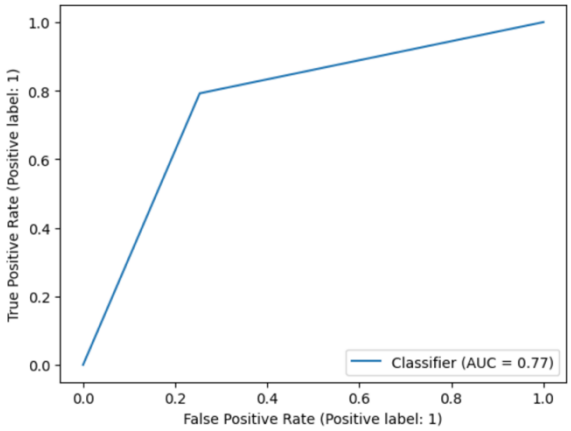
Métricas de Avaliação	
Especificidade	0,747
Sensibilidade	0,792
Acurácia	0,766
Precisão	0,707
F1 Score	0,747
Acurácia Balanceada	0,769
AUC	0,77

Fonte: Elaborado pela própria autora.

A Figura 34 mostra que a performance do modelo está acima da média, com uma métrica AUC de 0,77 ou 77%. Ele classifica imagens de diagnóstico positivo

para anormalidades com quase 80% de acurácia, classificando erroneamente em média 2 imagens de diagnóstico “normal” como “anormal” a cada 10.

Figura 34 - ViT Validação: Área sob a curva ROC (AUC)



Fonte: Elaborado pela própria autora.

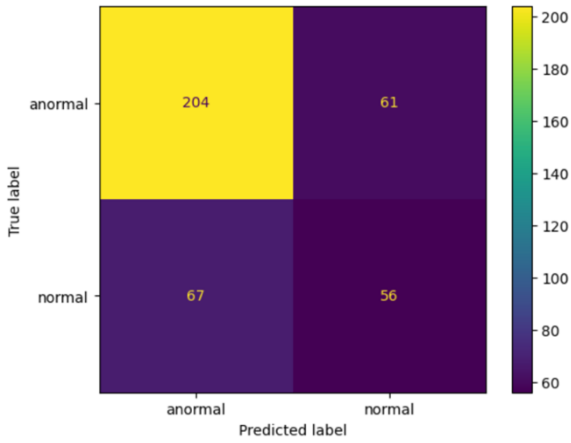
Imagens de teste reservadas fora do treino

388 imagens:

- 265 imagens classificadas como anormais (“positivo”)
- 123 imagens classificadas como normais (“negativo”)

Na Matriz de Confusão na Figura 35, percebe-se que a maioria das imagens está sendo classificada como “anormal”, mesmo que erroneamente.

Figura 35 - ViT Teste: Matriz de Confusão



Fonte: Elaborado pela própria autora.

Na Tabela 7, as métricas para classificação de diagnóstico “anormal” se mostram muito boas, enquanto as métricas para classificação de diagnóstico “normal” se mostram muito ruins. Considerando que a maioria das imagens do conjunto de teste é originalmente de diagnóstico “anormal”, o modelo performa bem, mesmo tendo sido treinado em mais imagens de diagnóstico “normal” do que “anormal”, o que significa que ele está conseguindo generalizar bem apesar dos conjuntos de imagens serem levemente desbalanceados. Mas, ele mostra também uma tendência em classificar erroneamente imagens saudáveis como anormais, o que pode ser devido à similaridade de caraterísticas entre os dois diagnósticos em algumas das imagens.

Tabela 7 - ViT Teste: Métricas por classe

Classificação	**Precisão**	**Sensibilidade**	**F1 Score**	**Qtd. Imagens**
Anormal:	0,75	0,77	0,76	265
Normal:	0,48	0,46	0,47	123

Fonte: Elaborado pela própria autora.

A Tabela 8 mostra que o modelo é razoavelmente bom em classificar os Verdadeiros Positivos (diagnóstico “anormal”) com a métrica de Sensibilidade de 0,770, mas muito ruim em classificar os Verdadeiros Negativos (diagnóstico “normal”) com Especificidade de 0,455. Resultado semelhante ao modelo CNN.

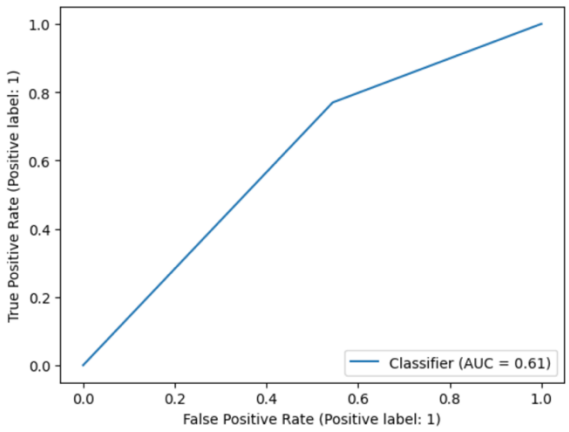
Tabela 8 - ViT Teste: Métricas

Métricas de Avaliação	
Especificidade	0,455
Sensibilidade	0,770
Acurácia	0,670
Precisão	0,753
F1 Score	0,761
Acurácia Balanceada	0,613
AUC	0,61

Fonte: Elaborado pela própria autora.

A Figura 36 mostra que a performance do modelo é acima da média, com uma métrica AUC de 0,61 ou 61%. Ele está um pouco melhor que o modelo CNN para classificar as imagens de teste, mas ainda assim está apresentando muito erro classificando imagens saudáveis como anormais.

Figura 36 - ViT Teste: Área sob a curva ROC (AUC)



Fonte: Elaborado pela própria autora.

4.2.2 Modelo ViT: Explicabilidade

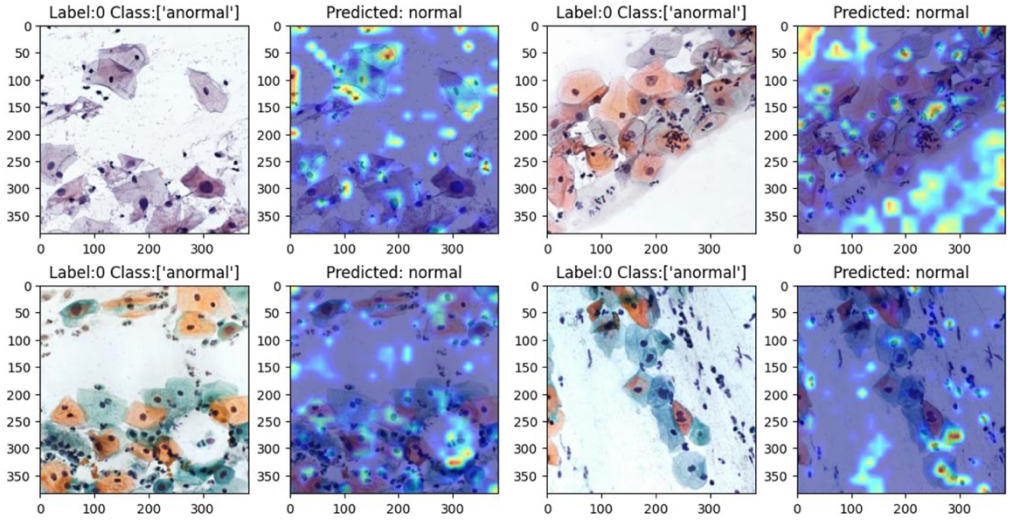
Com o auxílio da biblioteca Grad-CAM, as imagens dos conjuntos de validação e de teste foram analisadas para verificar quais características da imagem o modelo estava levando em consideração para a classificação. As imagens a seguir estão acompanhadas de uma cópia delas com as características da imagem destacadas em laranja e amarelo, características estas que o modelo atribuiu maior peso para a classificação.

Imagens de validação utilizadas no treino

Na Figura 37, observa-se que o modelo foca nas partes normais das imagens e por engano classifica as imagens com diagnóstico “normal” ao invés de “anormal”, o que ocorre em 48 imagens do conjunto de imagens. O modelo precisaria ser treinado a priorizar as partes anormais da imagem na sua classificação, podendo ser

treinado também com células anormais isoladas e recebendo maior penalidade quando deixa células anormais passarem despercebidas.

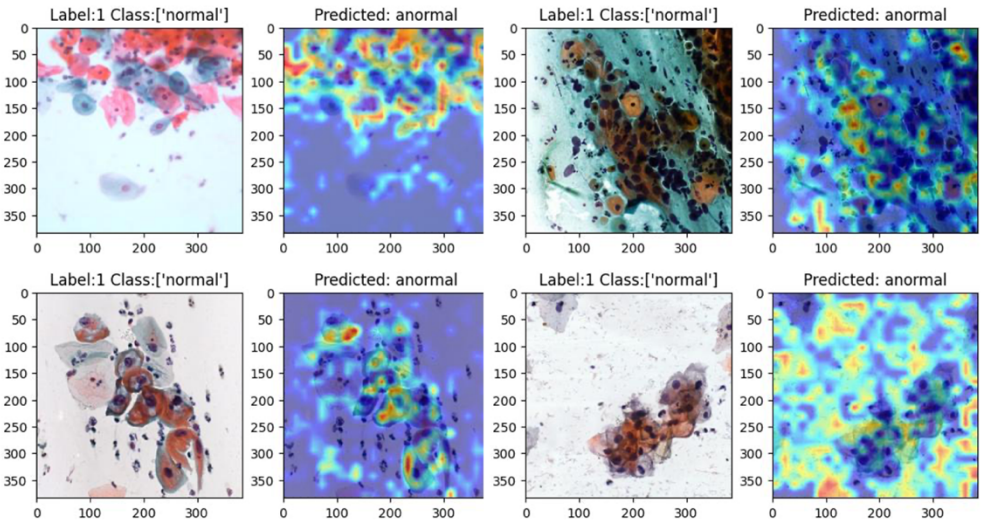
Figura 37 - Classificações de Validação incorretas: Falsos Negativos (Total: 48)



Fonte: Plissiti *et al.*, 2018; Rezende *et al.*, 2021; Hussain *et al.*, 2020; própria autora.

Na Figura 38, o modelo confunde imagens de diagnóstico “normal” com diagnóstico “anormal” pela similaridade de características de células de aspecto “anormal”, focando principalmente nos contornos das células, e manchas irregulares na lâmina.

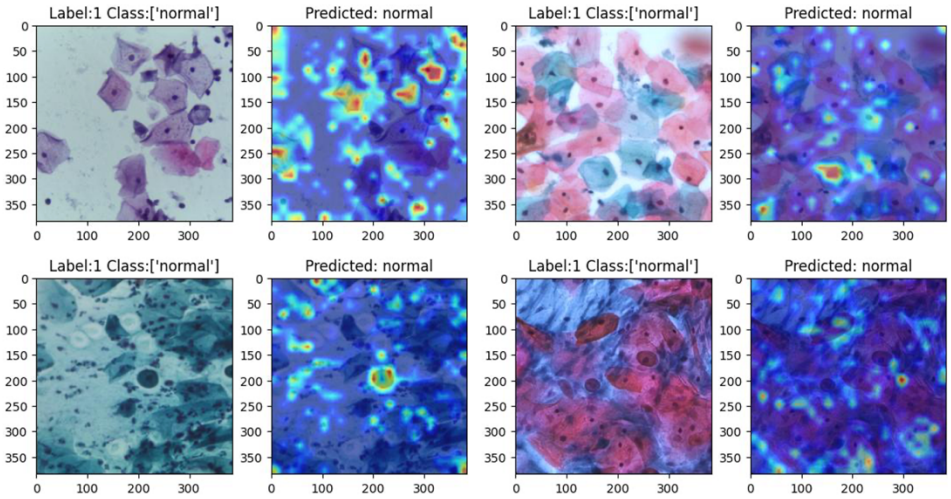
Figura 38 - Classificações de Validação incorretas: Falsos Positivos (Total: 76)



Fonte: Plissiti *et al.*, 2018; Rezende *et al.*, 2021; Hussain *et al.*, 2020; própria autora.

Na Figura 39, o modelo classifica corretamente todas as imagens de diagnóstico “normal” destacando aspectos importantes das imagens como núcleos das células e contornos das células, não se deixando enganar pelas tonalidades escuras ou aglomerações de células. Esses exemplos mostram que o modelo está focando no que realmente deveria para a classificação correta.

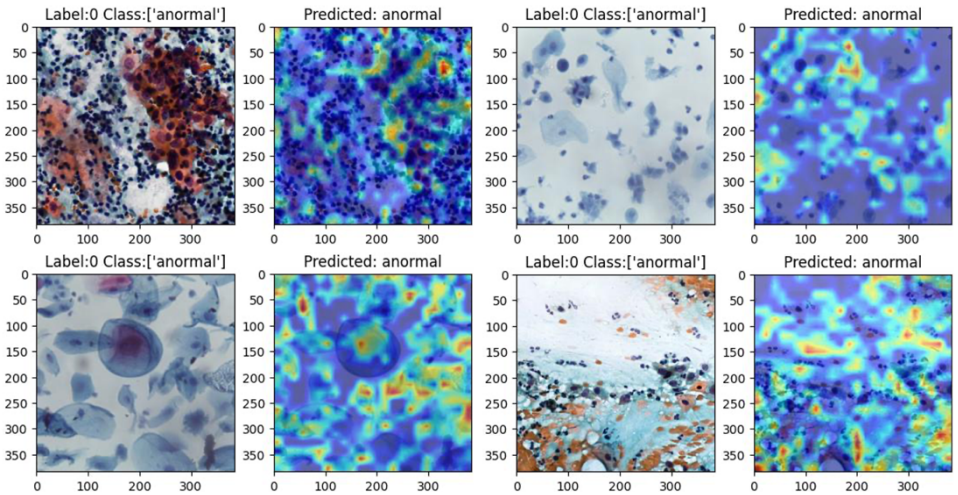
Figura 39 - Classificações de Validação corretas: Verdadeiros Negativos (Total: 224)



Fonte: Plissiti *et al.*, 2018; Rezende *et al.*, 2021; Hussain *et al.*, 2020; própria autora.

Na Figura 40, estão exemplos de classificações corretas com diagnóstico “anormal” e observa-se que o modelo destaca áreas relevantes para a classificação como os núcleos e contornos das células.

Figura 40 - Classificações de Validação corretas: Verdadeiros Positivos (Total: 183)

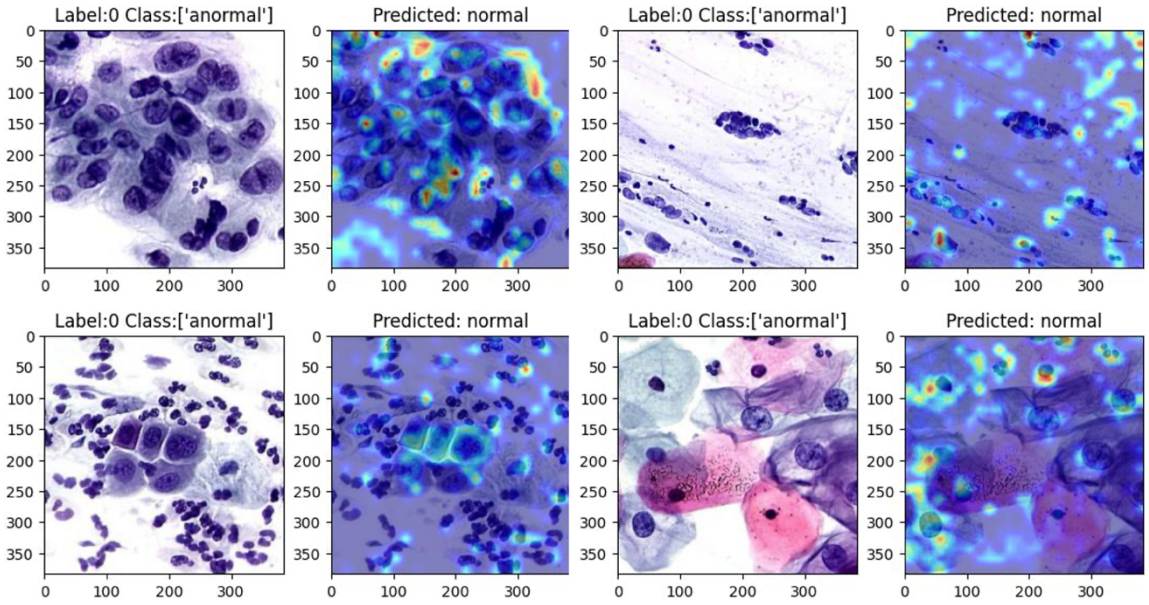


Fonte: Plissiti *et al.*, 2018; Rezende *et al.*, 2021; Hussain *et al.*, 2020; própria autora.

Imagens de teste reservadas fora do treino

Na Figura 41, agora com o conjunto de imagens de teste que foram deixadas fora do treino, o modelo erra ao classificar imagens de diagnóstico “anormal” como “normal” ao destacar partes irrelevantes das imagens para gerar sua classificação, assim como nos testes com o conjunto de imagens de validação anteriormente. O modelo precisa ser treinado para dar maior peso para às anormalidades na imagem.

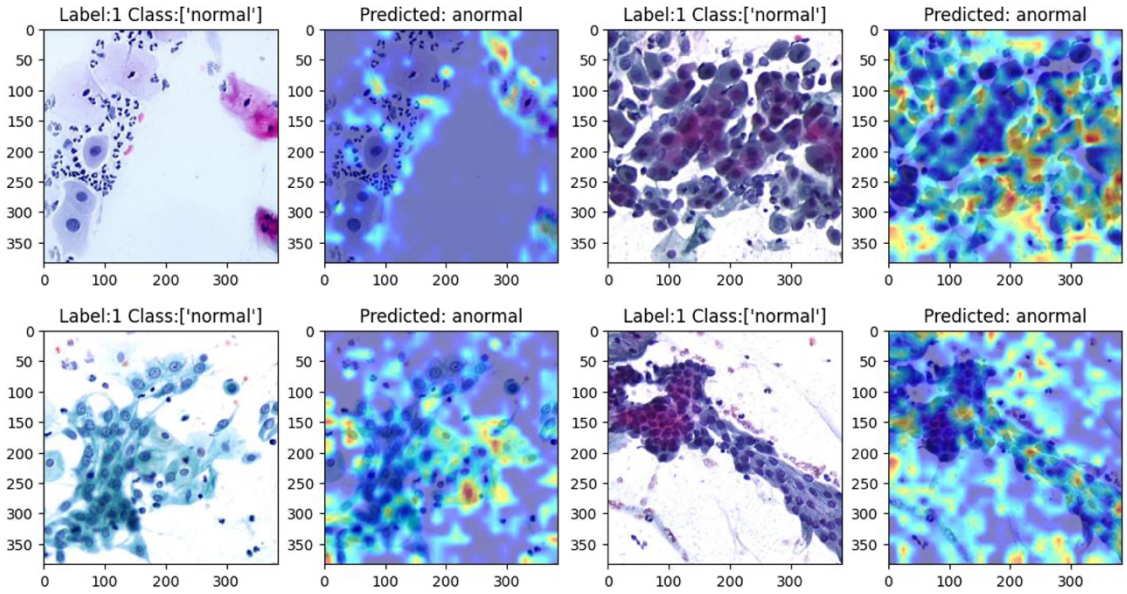
Figura 41 - Classificações de Teste incorretas: Falsos Negativos (Total: 61)



Fonte: IARC, 2004; Bethesda, 2015; própria autora.

Na Figura 42, o modelo erra novamente classificando imagens de diagnóstico “normal” como “anormal”, são imagens que apresentam muitas similaridades com imagens de diagnóstico “anormal”. O modelo destaca irregularidades nas lâminas e contornos das células.

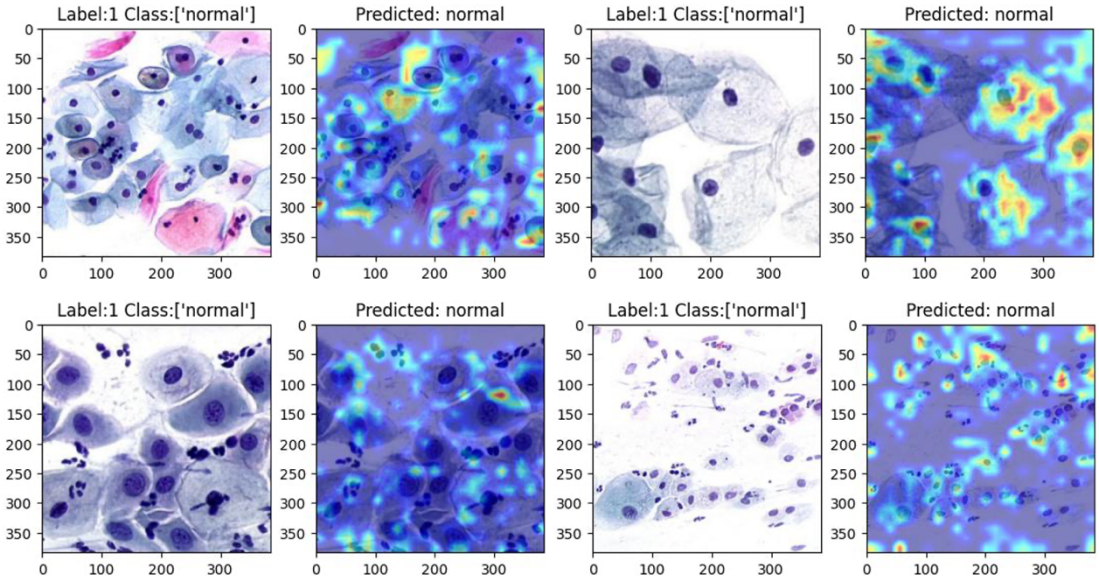
Figura 42 - Classificações de Teste incorretas: Falsos Positivos (Total: 67)



Fonte: IARC, 2004; Bethesda, 2015; própria autora.

Na Figura 43, observa-se que o modelo classificou corretamente as imagens de diagnóstico “normal” do conjunto de teste destacando os contornos e interiores das células, sendo essas imagens relativamente mais fácil para o modelo classificar, não contendo características que possam confundir.

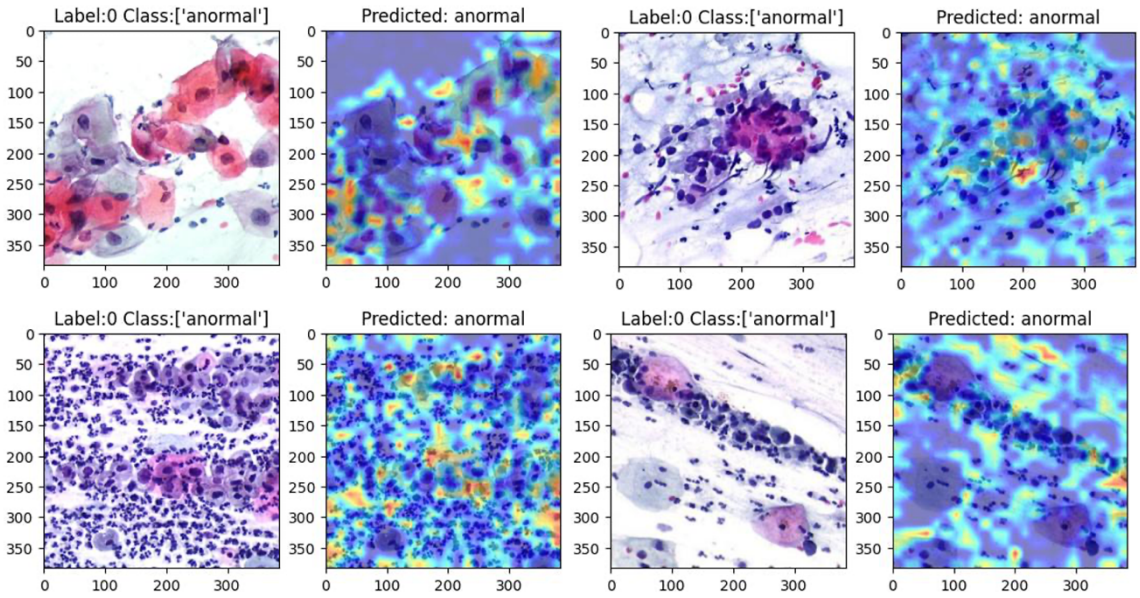
Figura 43 - Classificações de Teste corretas: Verdadeiros Negativos (Total: 56)



Fonte: IARC, 2004; Bethesda, 2015; própria autora.

Na Figura 44, observa-se que o modelo classificou corretamente as imagens do conjunto de teste de diagnóstico “anormal”, sendo essas a maior parte do conjunto de imagens, mesmo o modelo tendo sido treinado com mais imagens de diagnóstico “normal” do que “anormal”. O modelo aprendeu bem a identificar as características relevantes para classificar uma imagem como “anormal” destacando os contornos e interiores das células.

Figura 44 - Classificações de Teste corretas: Verdadeiros Positivos (Total: 204)



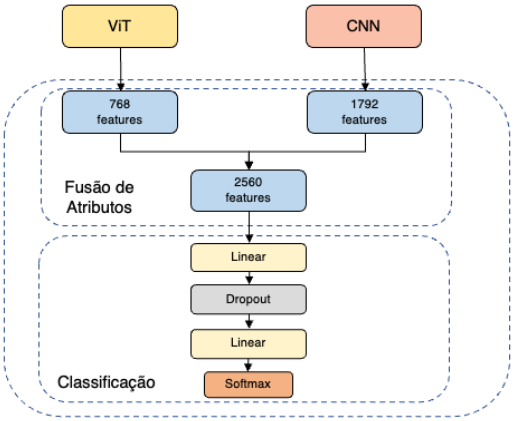
Fonte: IARC, 2004; Bethesda, 2015; própria autora.

4.3 Modelo de Fusão

Primeiramente o modelo de Fusão foi testado ao se simplesmente combinar a rede neural convolucional (CNN) e o modelo Vision Transformer (ViT), treinados previamente nas etapas anteriores, e apresentou uma acurácia no conjunto de imagens de validação de apenas 56,1%. Mesmo sendo essa apenas uma comparação inicial da qualidade do modelo treinado, ela ficou abaixo da acurácia do modelo CNN sozinho de 56,9% e do modelo ViT sozinho de 76,6%. O modelo de fusão combinando os outros dois foi então retreinado por 30 épocas utilizando a configuração na Figura 45, com as 768 *features* de saída do modelo ViT e 1792 as *features* de saída do modelo CNN. Essas *features* são características selecionadas

da imagem aos quais os modelos atribuem maior peso para auxiliar a classificação correta.

Figura 45 - Primeira configuração testada do modelo de fusão



Fonte: Elaborado pela própria autora.

O modelo demorava mais para ser treinado devido a sua maior complexidade e para validar esse experimento era preciso que o modelo atingisse uma performance superior. Considerando que eram dois modelos treinados em paralelo, o número de épocas também foi dobrado. Ao final de 30 épocas, levando em torno de 4 dias e 1h, o modelo foi salvo na sua melhor iteração e gráficos de acompanhamento foram gerados. Na figura 46 é possível observar grande variabilidade, mas mesmo ao se perder o modelo se manteve mais estável do que anteriormente, nunca ultrapassando a margem de erro de 1,75%.

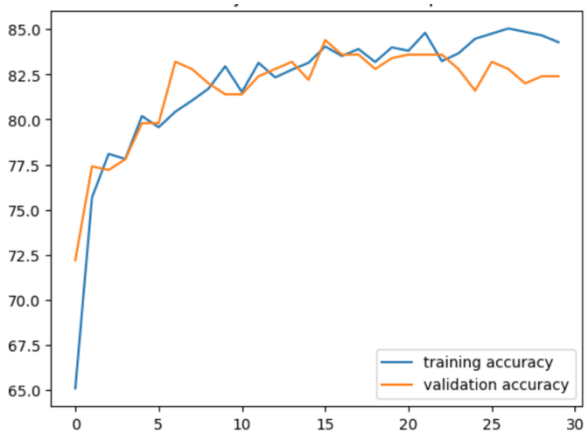
Figura 46 - Modelo de Fusão 1: Erro por época



Fonte: Elaborado pela própria autora.

Já ao observar a figura 47, percebe-se que a curva de aprendizado das imagens do conjunto de treino e de validação andam praticamente juntas, até chegarem a um patamar, onde o modelo passa a ficar muito melhor em classificar imagens do conjunto de treino em detrimento das imagens do conjunto de validação, caracterizando tendência a *overfitting*.

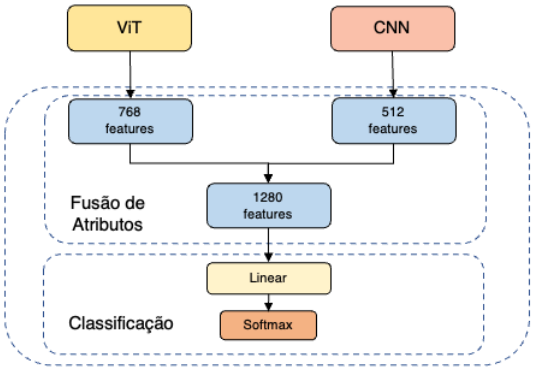
Figura 47 - Modelo de Fusão 1: Acurácia por época



Fonte: Elaborado pela própria autora.

Além da curva de aprendizado do modelo tender ao *overfitting* como o modelo ViT, a métrica de acurácia continuava tão ruim quanto a do modelo CNN sozinho, chegando a apenas a 57,1 % mesmo após o treinamento. Chegou-se à conclusão de que o modelo de fusão estava pegando as piores características de cada modelo. Então o desenho de saída da rede foi reformulado para balancear as características extraídas de cada modelo. A Figura 48 demonstra como a rede neural de fusão foi reestruturada, com um número de *features* de saída de cada modelo mais balanceado: 768 *features* de saída do modelo ViT e 512 *features* de saída do modelo CNN, priorizando o modelo ViT que apresentava inicialmente melhor acurácia.

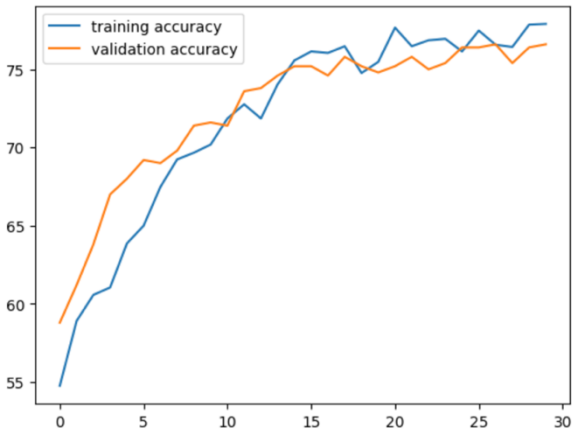
Figura 48 - Segunda configuração testada do modelo de fusão



Fonte: Elaborado pela própria autora.

Os testes já demonstraram melhora com o modelo obtendo uma acurácia de 63,3% depois do mesmo treino. O modelo foi retreinado por 30 épocas com uma taxa de aprendizado ("*learning rate*") de 10^{-5}, o que demorou mais quatro dias inteiros, e era possível ver que o modelo ainda tinha a possibilidade de aprender ainda mais se fosse treinado por mais tempo como mostra a Figura 49. As acurácias dos conjuntos de imagens de treino e de teste vão crescendo juntas.

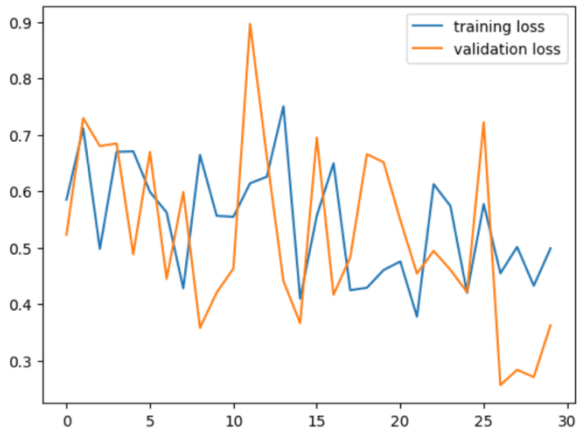
Figura 49 - Modelo de Fusão 2: Acurácia por época



Fonte: Elaborado pela própria autora.

E ao mesmo tempo o erro também gradualmente diminuindo apesar das ocasionais variações bruscas na Figura 50.

Figura 50 - Modelo de Fusão 2: Erro por época



Fonte: Elaborado pela própria autora.

4.3.1 Modelo de Fusão: Métricas de Avaliação

Finalmente para testar a qualidade do modelo de Fusão que combina os modelos CNN e ViT, são novamente utilizados os dois conjuntos de imagens. O primeiro conjunto de imagens de Validação com dados públicos SIPaKMeD, CRIC, Mendeley LBC, que também foi utilizado no treinamento dos modelos (Plissiti *et al.*, 2018; Rezende *et al.*, 2021; Hussain *et al.*, 2020). O segundo conjunto de Teste com imagens coletadas do site da Organização Mundial de Saúde (IARC, 2004) e do sistema Bethesda (BETHESDA, 2015), imagens que nunca foram visualizadas pelo modelo.

Imagens de validação utilizadas no treino

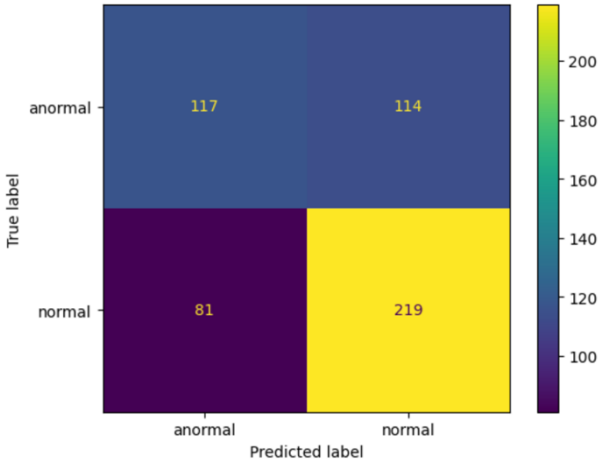
531 imagens:

- 231 imagens classificadas como anormais (“positivo”)
- 300 imagens classificadas como normais (“negativo”)

A Matriz de Confusão na Figura 51 mostra que o modelo segue a mesma tendência que o modelo ViT, classificando a maioria das imagens de diagnóstico “anormal” corretamente e a maioria das imagens de diagnóstico “normal” corretamente. E classificando ainda melhor as imagens de diagnóstico “normal”. O número alto de Falsos Negativos não condiz com um modelo otimizado para

diagnósticos na área de saúde, é mais importante que ele seja capaz de detectar anormalidades.

Figura 51 - Fusão Validação: Matriz de Confusão



Fonte: Elaborado pela própria autora.

Na Tabela 9, é possível perceber que as melhores métricas se encontram na classificação de imagens com diagnóstico “normal”. O que faz sentido já que o modelo possuía mais imagens de diagnóstico “normal” para seu treinamento.

Tabela 9 - Fusão Validação: Métricas por classe

Classificação	Precisão	Sensibilidade	F1 Score	Qtd. Imagens
Anormal:	0,59	0,51	0,55	231
Normal:	0,66	0,73	0,69	300

Fonte: Elaborado pela própria autora.

A Tabela 10 mostra uma métrica de Especificidade de 73% o que condiz com sua capacidade de diagnosticar imagens de diagnóstico “normal” melhor do que de diagnóstico “anormal”. Já a métrica de Sensibilidade em 50,6% mostra que o modelo deixou de diagnosticar corretamente um número considerável de anormalidades nas imagens.

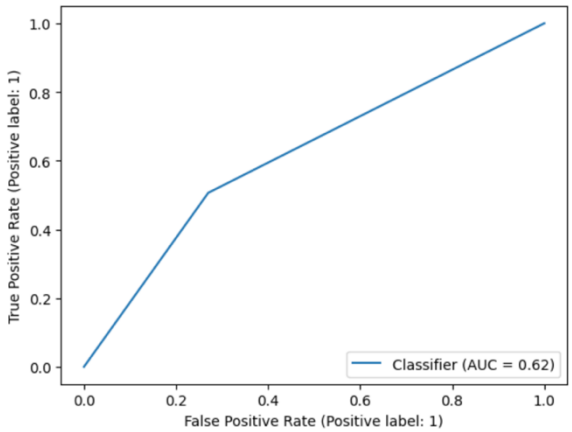
Tabela 10 - Fusão Validação: Métricas

Métricas de Avaliação	
Especificidade	0,730
Sensibilidade	0,506
Acurácia	0,633
Precisão	0,591
F1 Score	0,545
Acurácia Balanceada	0,618
AUC	0,62

Fonte: Elaborado pela própria autora.

A Figura 52 mostra uma métrica AUC de 0,62 acima da média, mas o modelo é melhor em classificar imagens com diagnóstico “normal” do que encontrar anormalidades. A métrica de Taxa de Verdadeiro Positivo (“*True Positive Rate*”) no eixo y mostra que o modelo classifica imagens de diagnóstico “anormal” corretamente em menos de 60% da vezes.

Figura 52 - Fusão Validação: Área sob a curva ROC (AUC)



Fonte: Elaborado pela própria autora.

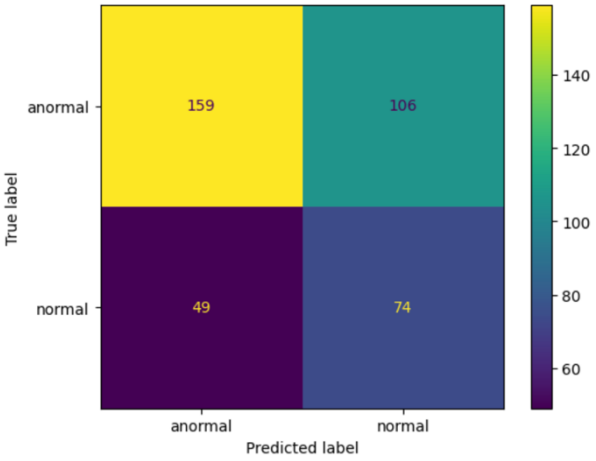
Imagens de teste reservadas fora do treino

388 imagens:

- 265 imagens classificadas como anormais (“positivo”)
- 123 imagens classificadas como normais (“negativo”)

A Matriz de Confusão na Figura 53 mostra que o modelo classifica corretamente a maioria das imagens de diagnóstico “anormal” como “anormal” e a maioria das imagens de diagnóstico “normal” como “normal”, mas o número de imagens de diagnóstico “anormal” diagnosticadas erroneamente como “normal” ainda é alto apesar de um pouco melhor do que no conjunto de validação.

Figura 53 - Fusão Teste: Matriz de Confusão



Fonte: Elaborado pela própria autora.

Na Tabela 11 a métrica de Precisão de 76% reflete a questão de que a maioria das imagens que são de diagnóstico “anormal” estão sendo corretamente classificadas, o que acaba melhorando as demais métricas, mas ainda são muitos Falsos Negativos de imagens de diagnóstico “anormal” sendo classificadas como normais. Já a métrica de Precisão na classificação de imagens de diagnóstico “normal” aparece mais baixa porque o número de imagens de diagnóstico “normal” no conjunto de imagens de teste é menor, o que proporcionalmente torna o erro de um Falso Negativo maior.

Tabela 11 - Fusão Teste: Métricas por classe

Classificação	Precisão	Sensibilidade	F1 Score	Qtd. Imagens
Anormal:	0,76	0,60	0,67	265
Normal:	0,41	0,60	0,49	123

Fonte: Elaborado pela própria autora.

A Tabela 12 mostra as métricas todas acima de 60% para a classificação das imagens do conjunto de teste, indicando que o modelo de Fusão consegue classificar melhor imagens do conjunto de teste do que as imagens do conjunto de validação.

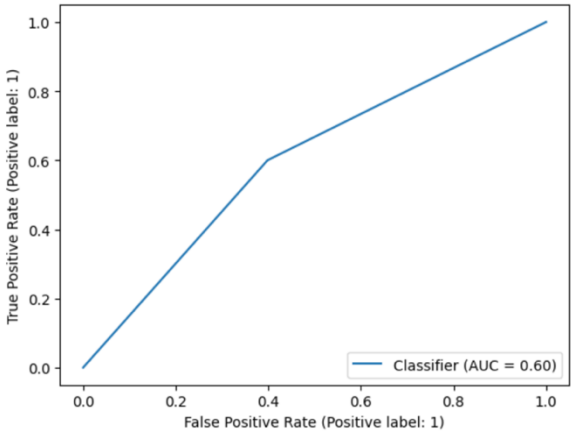
Tabela 12 - Fusão Teste: Métricas

Métricas de Avaliação	
Especificidade	0,602
Sensibilidade	0,600
Acurácia	0,601
Precisão	0,764
F1 Score	0,672
Acurácia Balanceada	0,601
AUC	0,60

Fonte: Elaborado pela própria autora.

A Figura 54 mostra uma métrica AUC de 0,60, acima da média, mostrando que o modelo tem performances parecidas tanto para o conjunto de teste quanto para o conjunto de validação. No entanto, o modelo se mostra melhor em classificar imagens de diagnóstico “anormal” do conjunto de teste do que do conjunto de validação.

Figura 54 - Fusão Teste: Área sob a curva ROC (AUC)



Fonte: Elaborado pela própria autora.

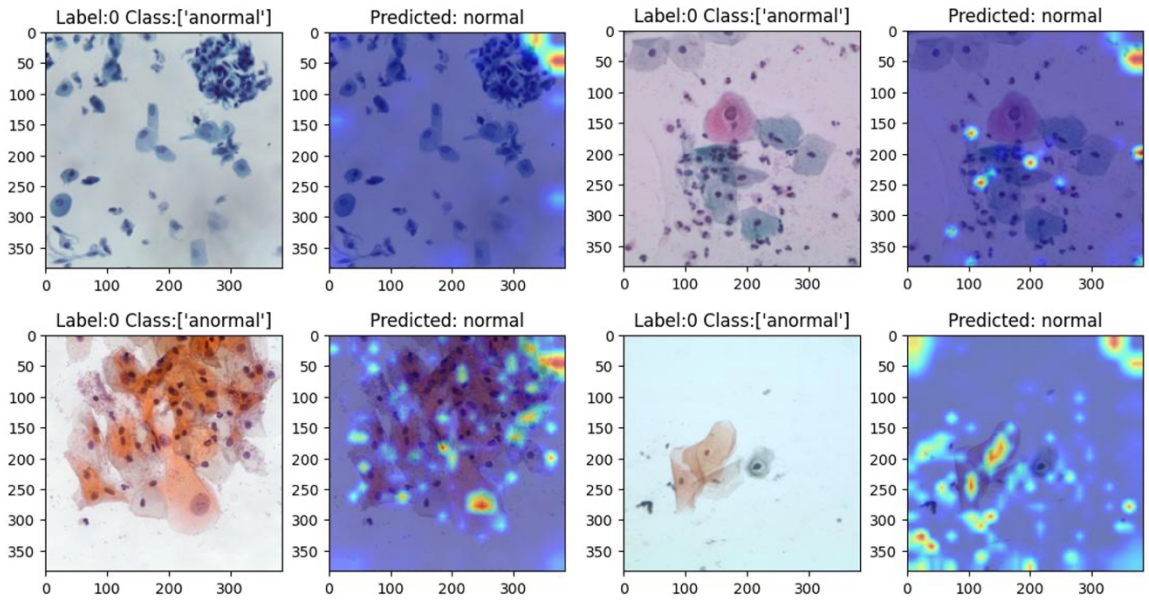
4.3.2 Modelo de Fusão: Explicabilidade

A biblioteca Grad-CAM foi utilizada para verificar o que o modelo de Fusão estava destacando ao classificar uma imagem com diagnóstico “normal” ou “anormal”. Essa análise e comparação pode servir para mostrar se o modelo melhorou a sua classificação e se está olhando características relevantes para o diagnóstico.

Imagens de validação utilizadas no treino

Na Figura 55, agora analisando a classificação do modelo de Fusão que combina a rede neural convolucional e o modelo *Vision Transformer*, nos erros das imagens de diagnóstico “anormal” sendo erroneamente classificadas como “normais”, é possível ver o modelo focando no canto superior direito da imagem assim como o modelo de rede neural convolucional previamente treinado estava fazendo, ou mesmo destacando as partes relevantes e classificando de maneira errada. O modelo de Fusão acabou por herdar a maneira aleatória que o modelo de rede neural convolucional estava classificando as imagens com muitos mais Falsos Negativos, o que aumentou o erro em relação ao modelo de Vision Transformer sozinho.

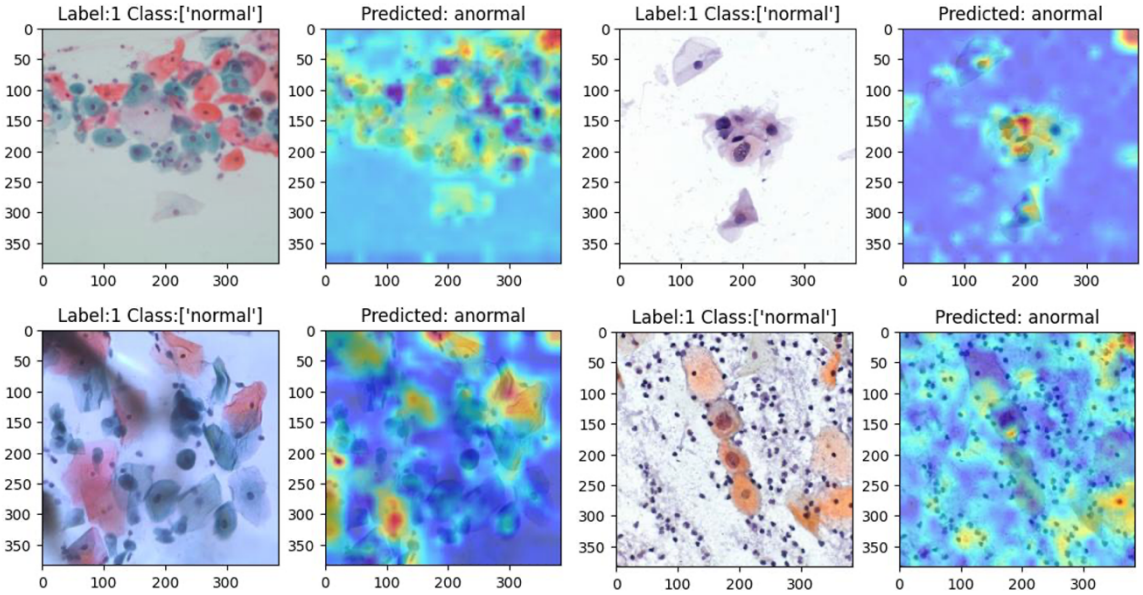
Figura 55 - Classificações de Validação incorretas: Falsos Negativos (Total: 114)



Fonte: Plissiti *et al.*, 2018; Rezende *et al.*, 2021; Hussain *et al.*, 2020; própria autora.

Na Figura 56, é possível ver o modelo destacando partes relevantes das imagens como os contornos das células e núcleos, mesmo assim ele classifica de maneira errada, sendo confundido por irregularidades nas imagens ou pelas similaridades de imagens de diagnóstico “anormal”.

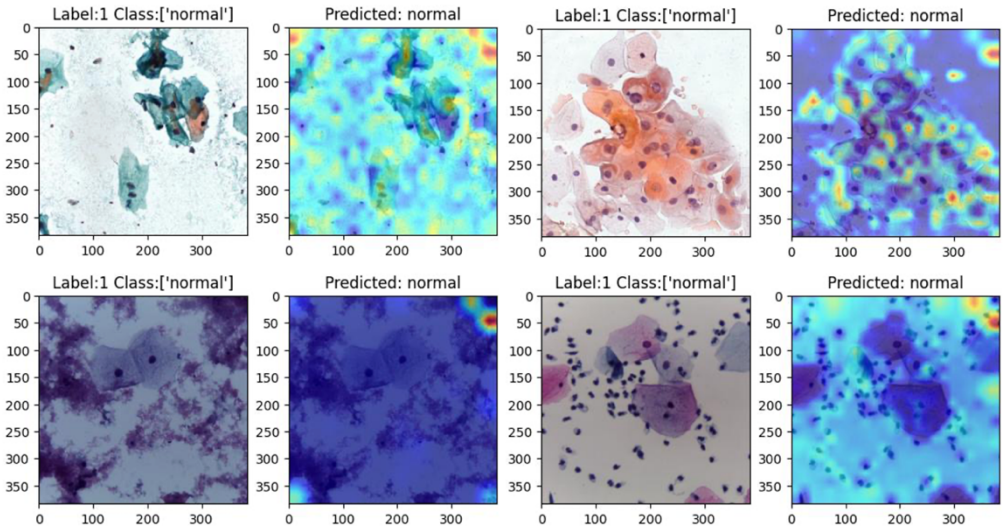
Figura 56 - Classificações de Validação incorretas: Falsos Positivos (Total: 81)



Fonte: Plissiti *et al.*, 2018; Rezende *et al.*, 2021; Hussain *et al.*, 2020; própria autora.

Na Figura 57, é possível ver que o modelo destaca partes relevantes das imagens para a sua classificação correta, e ao mesmo tempo são visíveis os vestígios da classificação aleatória da rede neural convolucional que foca no canto superior direito das imagens. O modelo de *Vision Transformer* ressalta outras áreas das imagens, mas os pesos que o modelo CNN atribui às áreas irrelevantes desequilibram a classificação. O modelo apenas consegue acertar nesse caso porque a maioria das imagens são originalmente de diagnóstico “normal”.

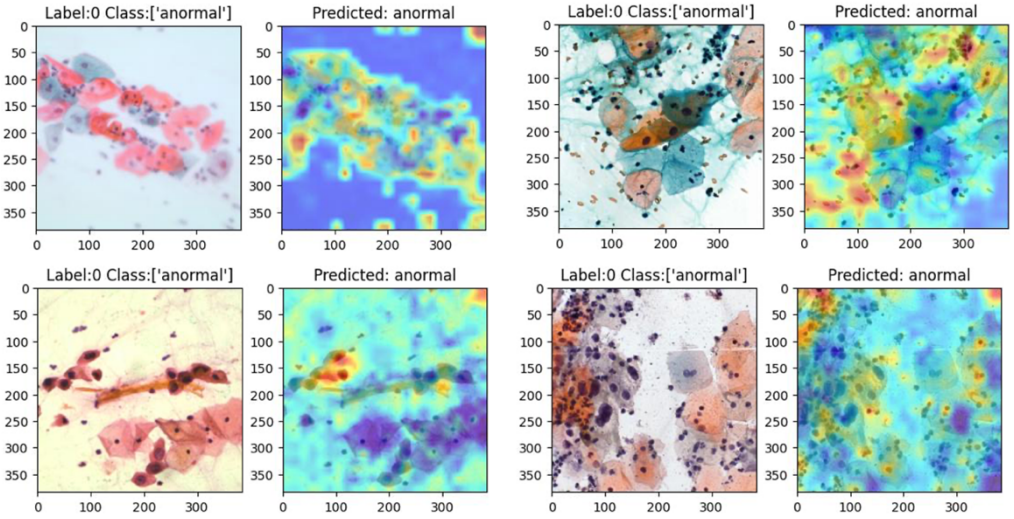
Figura 57 - Classificações de Validação corretas: Verdadeiros Negativos (Total: 219)



Fonte: Plissiti *et al.*, 2018; Rezende *et al.*, 2021; Hussain *et al.*, 2020; própria autora.

Na Figura 58, é possível ver que que os dois modelos combinados sincronizam as partes sendo destacadas, concentrando-se nos contornos da células e núcleos, apesar do modelo CNN ainda dar destaque desnecessário ao canto superior direito das imagens e outras áreas irrelevantes.

Figura 58 - Classificações de Validação corretas: Verdadeiros Positivos (Total: 117)

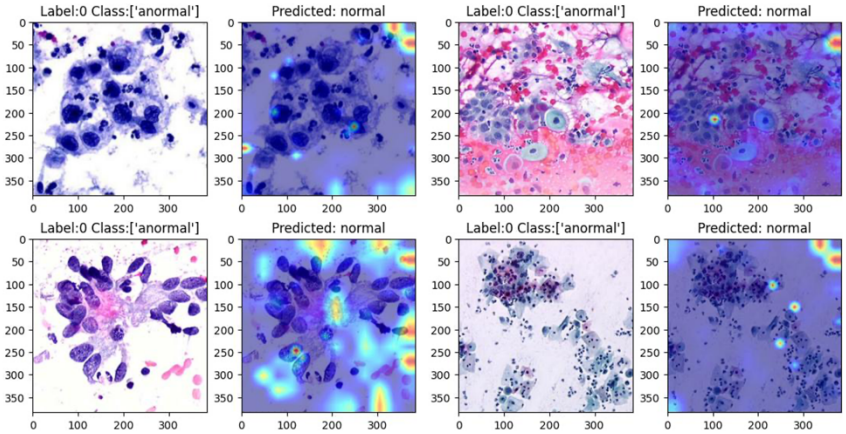


Fonte: Plissiti *et al.*, 2018; Rezende *et al.*, 2021; Hussain *et al.*, 2020; própria autora.

Imagens de teste reservadas fora do treino

Na Figura 59, ainda é possível ver o modelo de rede neural convolucional dando destaque ao canto superior direito da imagens e outras partes irrelevantes como irregularidades. O modelo de *Vision Transformer* perde destaque nessas imagens. Os Falsos Negativos possuem impacto muito negativo na medicina diagnóstica.

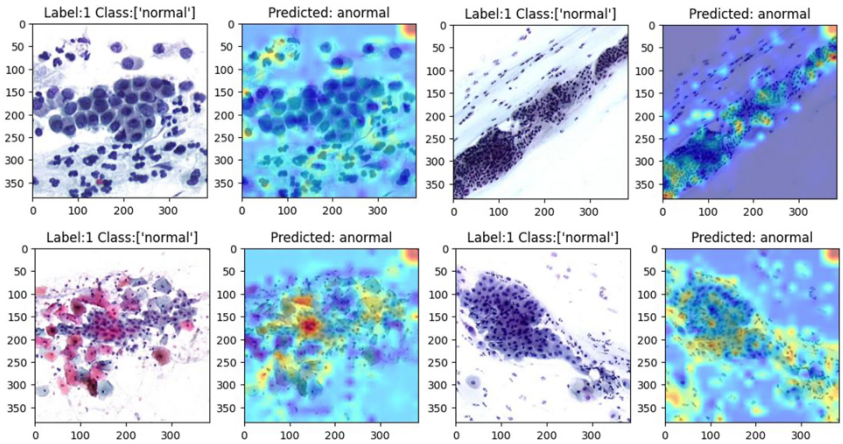
Figura 59 - Classificações de Teste incorretas: Falsos Negativos (Total: 106)



Fonte: IARC, 2004; Bethesda, 2015; própria autora.

Na Figura 60, o modelo destaca partes relevantes da imagem, mas parece se confundir devido às similaridades com imagens de diagnóstico “anormal” como o agrupamento denso de células.

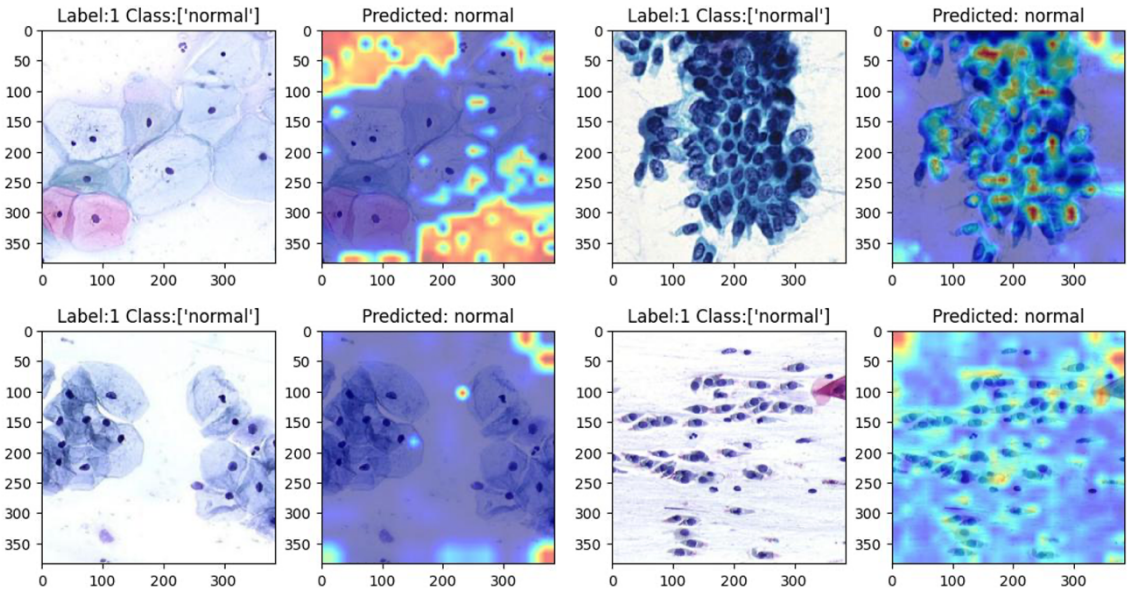
Figura 60 - Classificações de Teste incorretas: Falsos Positivos (Total: 49)



Fonte: IARC, 2004; Bethesda, 2015; própria autora.

Na Figura 61, é possível ver o modelo destacando fortemente as áreas exteriores às células para classificá-las como normais na primeira imagem, apesar dessa provavelmente não ser a informação mais relevante. Essa imagem evidencia uma característica que o modelo está enfatizando, mas pode também induzi-lo ao erro. Novamente também o foco desnecessário no canto superior direito. Nas outras imagens o modelo consegue classificar corretamente apesar da densa aglomeração de células e irregularidades.

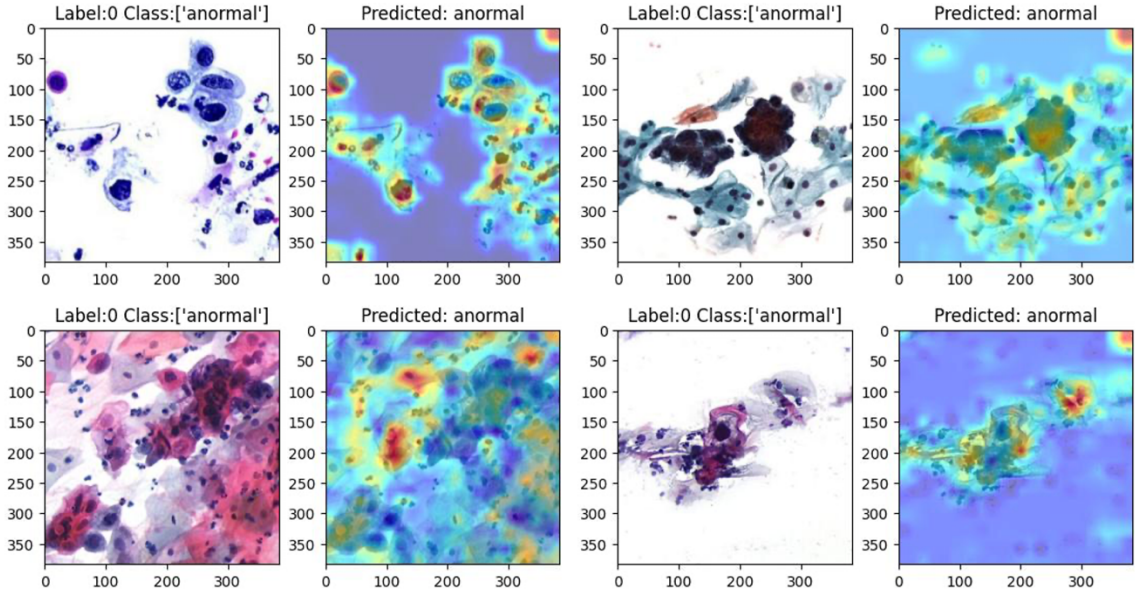
Figura 61 - Classificações de Teste corretas: Verdadeiros Negativos (Total: 74)



Fonte: IARC, 2004; Bethesda, 2015; própria autora.

Na Figura 62, o modelo destaca muito bem as partes relevantes para classificação de imagens de diagnóstico “anormal”. Os fortes tons alaranjados e amarelados indicam que os modelos combinados atribuem um peso maior a essas características. Nesses exemplo o modelo de Fusão destaca bem imagens de diagnóstico “anormal” em um conjunto de imagens nunca visualizado, apresentando também menos Falsos Positivos.

Figura 62 - Classificações de Teste corretas: Verdadeiros Positivos (Total: 159)



Fonte: IARC, 2004; Bethesda, 2015; própria autora.

4.4 Comparação dos Modelos

Na Tabela 13 os três modelos são comparados em termos de Sensibilidade, Especificidade e Área sob a Curva (AUC), analisando os pontos fortes e fracos de cada um ao classificar o conjunto de imagens de Validação.

Tabela 13 - Comparação dos 3 modelos no conjunto de Validação

Métricas	EfficientNet-b4	DeiT3-base	Fusão
Sensibilidade	0,766	0,792	0,506
Especificidade	0,417	0,747	0,730
AUC	0,59	0,77	0,62

Fonte: Elaborado pela própria autora.

O modelo de rede neural convolucional EfficientNet-b4 é bom em identificar corretamente as imagens de diagnóstico “anormal”. Fraco em identificar corretamente as imagens de diagnóstico “normal”. Desempenho global baixo, indicando baixa separabilidade entre as classes. Esse modelo tem um desempenho

moderado para identificar imagens de diagnóstico “anormal”, mas falha em reconhecer imagens de diagnóstico “normal”. O AUC baixo sugere que ele não é eficiente para discriminar entre as classes.

O modelo de *Vision Transformer* DeiT3-base é o melhor em identificar imagens de diagnóstico “anormal” entre os três modelos. Bom desempenho em identificar corretamente as imagens de diagnóstico “normal”. Desempenho global sólido, com boa capacidade de discriminação entre as classes. Esse modelo apresenta o melhor equilíbrio entre sensibilidade e especificidade, além de um AUC elevado, sugerindo que é o mais eficaz dos três para a tarefa de classificação.

Apesar de ser uma fusão dos dois modelos, o modelo de Fusão não conseguiu melhorar significativamente o desempenho global. O baixo valor de sensibilidade indica que ele não é eficaz para detectar imagens de diagnóstico “anormal”.

Na Tabela 14 os três modelos também são comparados na suas performances com o segundo conjunto de imagens de teste que não foram utilizadas no treino, considerando as métricas de Sensibilidade, Especificidade e AUC.

Tabela 14 - Comparação dos 3 modelos no conjunto de Teste

Métricas	**EfficientNet-b4**	**DeiT3-base**	**Fusão**
Sensibilidade	0,475	0,792	0,600
Especificidade	0,504	0,747	0,602
AUC	0,59	0,77	0,62

Fonte: Elaborado pela própria autora.

O modelo de rede neural convolucional EfficientNet-b4 apresenta baixo desempenho na identificação de imagens de diagnóstico “anormal”. É apenas ligeiramente melhor que um modelo aleatório na identificação de imagens de diagnóstico “normal”. E a métrica AUC abaixo de 0,5, indica que o modelo é ineficaz e não discrimina bem entre as classes. O modelo CNN tem um desempenho muito fraco no novo conjunto de imagens, falhando tanto em sensibilidade quanto em especificidade. O AUC abaixo de 0,5 sugere que ele não está capturando corretamente as características do conjunto.

O modelo de *Vision Transformer* DeiT3-base apresenta bom desempenho na identificação de imagens de diagnóstico “anormal”. Mas, tem desempenho fraco na identificação de imagens de diagnóstico “normal”. Desempenho moderado, tem uma capacidade de separação limitada, mas ainda melhor que o CNN. O modelo ViT continua tendo uma boa Sensibilidade para detectar imagens de diagnóstico “anormal”, mas a Especificidade baixa prejudica o equilíbrio geral. Seu AUC de 0,61 sugere um desempenho aceitável, mas longe do ideal.

O Modelo de Fusão apresenta desempenho intermediário na identificação de imagens de diagnóstico “anormal” e desempenho equilibrado e moderado na identificação de imagens de diagnóstico “normal”. Possui também um desempenho global comparável ao ViT, com um equilíbrio melhor entre as métricas. O modelo de fusão tem um desempenho mais equilibrado em termos de sensibilidade e especificidade. Embora não atinja os melhores valores isolados, oferece um desempenho global decente com um AUC ligeiramente inferior ao ViT.

Imagens de validação utilizadas no treino

- Melhor modelo em termos globais (AUC): ViT DeiT3-base (AUC = 0,77). Ele é o mais equilibrado e eficaz para discriminar entre as classes.
- Modelo com melhor sensibilidade: ViT DeiT3-base (79,2%).
- Modelo com melhor especificidade: ViT DeiT3-base (74,7%).
- Modelo com pior desempenho global: CNN EfficientNet-b4 (AUC = 0,59).
- Modelo de Fusão não superou o ViT: Embora a fusão combine os dois modelos, seu desempenho é inferior ao do ViT individual.

O ViT DeiT3-base é o mais indicado para a classificação das imagens, pois oferece o melhor equilíbrio entre sensibilidade e especificidade, além do maior AUC, indicando uma performance superior no cenário proposto.

Imagens de teste reservadas fora do treino

- Melhor modelo em termos globais (AUC): ViT DeiT3-base (AUC = 0,61). Ele mantém um bom desempenho na identificação de diagnóstico “anormal”, mas sofre no diagnóstico “normal”.

- Modelo mais equilibrado: Modelo de Fusão (Sensibilidade = 60,0%, Especificidade = 60,2%). Embora o AUC seja ligeiramente menor que o ViT, o modelo é mais consistente nas duas métricas principais.

- Pior desempenho global: CNN EfficientNet-b4 (AUC = 0,49). O modelo não foi eficaz nesse conjunto de dados.

Se o foco principal for a Sensibilidade (priorizando a identificação de imagens de diagnóstico “anormal”), o ViT DeiT3-base é o mais indicado.

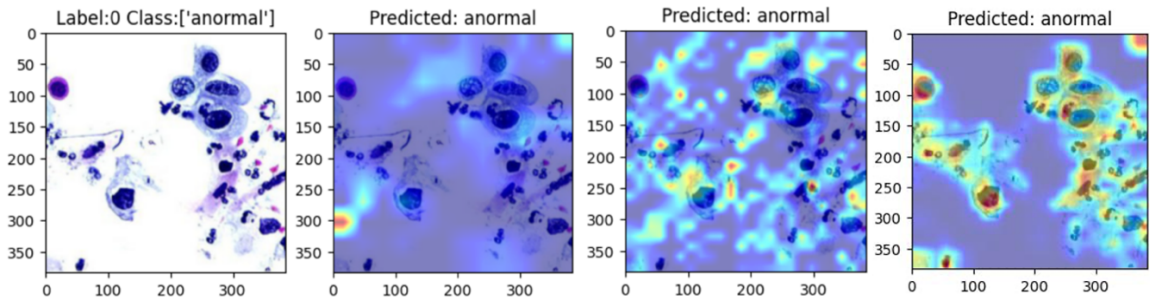
Se o objetivo for um desempenho mais equilibrado entre as classes, o Modelo de Fusão é a melhor escolha, pois alcança especificidade e sensibilidade razoáveis com desempenho consistente.

O CNN EfficientNet-b4 deveria ser descartado devido ao desempenho global insuficiente (AUC < 0,5).

Comparação dos destaques na classificação

Na figura 63, é possível visualizar como cada modelo classificou uma imagem de diagnóstico anormal corretamente e quais áreas o modelo estava destacando para essa classificação específica. Há uma melhora gradual na explicabilidade de cada modelo, com o modelo CNN destacando partes menos relevantes da imagem, o modelo ViT melhorando a relevância dos destaques, e finalmente, na fusão dos dois modelos, destacando detalhadamente todos os contornos das células e núcleos, que seriam realmente as áreas fundamentais para fornecer um diagnóstico correto. Demonstrando que com maior aprofundamento, mais tempo de treinamento dos modelos e uma seleção correta de configurações e redes, o modelo poderia fornecer uma classificação correta e explicável, podendo ainda ser visualmente validada por um profissional da área médica.

Figura 63 - Grad-CAM: CNN, ViT e Fusão



Fonte: IARC, 2004; Bethesda, 2015; própria autora.

4.5 Artigos

Nos apêndices constam os artigos publicados e submetidos a partir da pesquisa dessa dissertação.

5 DISCUSSÃO

Os resultados obtidos neste estudo destacam os desafios e as potencialidades associados à combinação de redes neurais convolucionais (CNNs) e *Vision Transformers* (ViTs) na classificação de imagens de citologia cervical. O estudo foi realizado com foco em imagens sem filtragem, abrangendo tanto células normais quanto anormais, com a intenção de melhorar a acurácia diagnóstica e minimizar a ocorrência de falsos negativos, que são críticos em contextos clínicos. Ao comparar os modelos individuais — EfficientNet-b4 e DeiT3-base — e o modelo de Fusão, observa-se uma diferença significativa no desempenho, especialmente em termos de Sensibilidade, Especificidade e Explicabilidade.

O modelo EfficientNet-b4 apresentou os piores resultados entre os três modelos, tanto no conjunto de validação quanto no de teste. Com um AUC de 0,59 e 0,49 nos respectivos conjuntos, fica evidente que este modelo não foi capaz de discriminar adequadamente entre as classes normais e anormais. A alta taxa de falsos negativos, particularmente no conjunto de teste (134 casos), é preocupante, pois sugere que este modelo não é confiável para aplicações clínicas, onde o erro pode resultar em riscos significativos para os pacientes.

Por outro lado, o modelo DeiT3-base demonstrou ser o mais eficaz, com uma sensibilidade de 79,2% e 77,0% e AUCs de 0,77 e 0,61 nos conjuntos de validação e teste, respectivamente. Esses resultados indicam que o ViT é mais robusto na identificação de anormalidades celulares, mesmo quando exposto a imagens de teste nunca vistas anteriormente. Essa robustez pode ser atribuída à sua capacidade de capturar relações globais nas imagens, evidenciada pela explicação gerada pelo Grad-CAM, que mostrou que o modelo focava em regiões relevantes, como contornos celulares e núcleos. Essa diferença pode ser atribuída às condições de treinamento, que no presente estudo foram propositalmente mais desafiadoras devido à utilização de imagens sem filtragem, o que reflete um cenário mais próximo da prática clínica.

O modelo de Fusão não alcançou o desempenho esperado. Quando os modelos foram combinados, a intenção era integrar a capacidade do EfficientNet de capturar características locais com a habilidade do DeiT3 de capturar características

globais. Embora tenha apresentado um AUC de 0,62 no conjunto de validação, comparável ao do DeiT3-base sozinho, sua Sensibilidade foi significativamente inferior (50,6%). Essa queda no desempenho pode ser atribuída à contribuição negativa do EfficientNet-b4 na combinação, uma vez que este modelo frequentemente focalizava regiões irrelevantes das imagens, conforme evidenciado pelas análises de explicação com Grad-CAM. A inclusão do EfficientNet no modelo combinado parece ter prejudicado a performance, possivelmente devido ao desbalanceamento entre as características locais e globais extraídas.

A aplicação do Grad-CAM revelou informações cruciais sobre como os modelos tomavam suas decisões. Enquanto o EfficientNet-b4 frequentemente se concentrava em áreas vazias ou irrelevantes das imagens, o DeiT3-base demonstrou um comportamento mais consistente, focando nas regiões de interesse, como os núcleos celulares e os contornos das células. Esses achados destacam a importância de incorporar ferramentas de explicação em sistemas de IA, especialmente em aplicações médicas, onde a compreensão do processo de classificação pode aumentar a confiança e a segurança no uso do modelo.

Embora as métricas globais de desempenho sejam importantes, elas não são suficientes para garantir a confiabilidade de um modelo. Um modelo com boas métricas, mas que não foca nas regiões relevantes para a classificação, pode produzir resultados inconsistentes ou até perigosos em cenários reais. Portanto, a explicação gerada pelo Grad-CAM desempenhou um papel fundamental neste estudo ao expor as limitações do EfficientNet-b4 e propor a adequação do DeiT3-base para a tarefa proposta.

Uma das principais limitações deste estudo foi o uso de imagens com várias células e sem filtros, em vez de células recortadas e selecionadas, como nos estudos de referência. Isso pode ter contribuído para o desempenho inferior dos modelos em relação à literatura. (PACAL, 2024) Além disso, é possível que outros modelos CNN, mais adequados para a tarefa, possam apresentar melhor desempenho quando integrados a um ViT. Investigar outras arquiteturas CNN, como ResNet ou DenseNet, em futuras pesquisas pode trazer insights valiosos.

Os achados deste trabalho também reforçam que, embora o uso de imagens sem filtragem seja desafiador, ele também expõe limitações importantes das

arquiteturas atuais. A dificuldade do EfficientNet em interpretar corretamente as imagens demonstra que redes convolucionais podem apresentar vulnerabilidades frente a ruídos e complexidades do mundo real. Por outro lado, o desempenho consistente do DeiT3-base sugere que modelos baseados em *Transformers* possuem maior potencial para aplicações clínicas em ambientes menos controlados.

Outra área promissora para trabalhos futuros é a otimização da estratégia de fusão. A abordagem utilizada neste estudo pode não ter sido ideal para explorar totalmente as complementaridades entre CNNs e ViTs. Futuras investigações podem explorar novos métodos de fusão que integrem as forças dos dois modelos de forma mais eficiente, além de incorporar técnicas para lidar com imagens não filtradas, ampliando as possibilidades de aplicação em contextos reais.

Por fim, a inclusão de mais dados de diferentes fontes e a avaliação de modelos em cenários reais são essenciais para validar os resultados obtidos e garantir sua generalização. O uso de técnicas de explicação como Grad-CAM deve ser ampliado para garantir a transparência e a segurança dos sistemas de IA em aplicações clínicas.

6 CONCLUSÃO

Este estudo investigou a aplicação de redes neurais convolucionais (EfficientNet-b4) e Vision Transformers (DeiT3-base) na classificação de imagens de citologia cervical, além da fusão dessas arquiteturas. Os resultados evidenciaram a superioridade do modelo DeiT3-base, que alcançou uma Sensibilidade de 79,2%, Especificidade de 74,7% e AUC de 0,77, destacando-se como o mais eficaz e robusto. Em contraste, o EfficientNet-b4 apresentou desempenho inferior, com Sensibilidade de 76,6%, Especificidade de 41,7% e AUC de 0,59, comprometendo também o desempenho do modelo combinado. A abordagem de fusão, que obteve Sensibilidade de 50,6%, Especificidade de 73,0% e AUC de 0,62, mostrou-se limitada, indicando a necessidade de aprimoramentos. A análise com Grad-CAM reforçou a importância da explicabilidade, evidenciando que o DeiT3-base focava em áreas relevantes das imagens, enquanto o comportamento do EfficientNet-b4 era mais aleatório. Este achado destaca a relevância de métodos de inteligência artificial responsável, que garantam decisões fundamentadas em características compreensíveis e confiáveis. Apesar dos avanços, o elevado número de falsos negativos aponta para desafios na adaptação dos modelos a imagens não filtradas e heterogêneas. Futuras pesquisas podem abordar essas limitações por meio da exploração de arquiteturas CNN mais adequadas, métodos de fusão mais eficientes e estratégias avançadas de pré-processamento. Além disso, a criação de conjuntos de dados mais diversificados e representativos pode contribuir para o desenvolvimento de modelos mais robustos e aplicáveis em condições reais. Em síntese, este estudo reforça a importância de uma abordagem integrada que equilibre desempenho, explicabilidade e robustez. Ele contribui para o avanço do uso de aprendizado profundo em aplicações clínicas e fornece diretrizes para futuros aprimoramentos nessa área.

REFERÊNCIAS

ADADI, A.; BERRADA, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). **IEEE Access**, v. 6, p. 52138-52160, 2018. DOI: 10.1109/ACCESS.2018.2870052.

BENGTSSON, E.; MALM, P. Screening for Cervical Cancer Using Automated Analysis of PAP-Smears. **Computational and Mathematical Methods in Medicine**, v. 2014, Article ID 842037, p.1-12, 2014. DOI: 10.1155/2014/842037. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3977449/. Acesso em: 30 jul. 2024.

BETHESDA. **For Reporting Cervical Cytology**: Online Atlas. 2015. Disponível em: https://bethesda.soc.wisc.edu/index.htm. Acesso em: 30 jul. 2024.

BHOJANAPALLI, S.; CHAKRABARTI, A.; GLASNER, D.; LI, D.; UNTERTHINER, T.; VEIT, A. Understanding Robustness of Transformers for Image Classification. **Proceedings of the IEEE/CVF International Conference on Computer Vision** (ICCV), p. 10231-10241, 2021. DOI: 10.48550/arXiv.2103.14586.

BISCOTTI, C. V.; DAWSON, A. E.; DZIURA, B.; GALUP, L.; DARRAGH, T.; RAHEMTULLA, A.; WILLS-FRANK, L. Assisted Primary Screening Using the Automated ThinPrep Imaging System. **American Society for Clinical Pathology**, v. 123, p. 281-287, 2005. DOI: 10.1309/AGB1MJ9H5N43MEGX.

BRAY, F.; LAVERSANNE, M.; SUNG, H.; FERLAY, J.; SIEGEL, R.L.; SOERJOMATARAM, I. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. **CA: A Cancer Journal for Clinicians**, v. 74, n. 3, p. 229–263, abr. 2024. DOI: 10.3322/caac.21834. Disponível em: https://pubmed.ncbi.nlm.nih.gov/38572751/. Acesso em: 30 jul. 2024.

CARUANA, R.; LOU, Y.; GEHRKE, J.; KOCH, P.; STURM, M.; ELHADAD, N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *In:* KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 21., 2015. **Anais** [...]. New York: American for Computing Machinery, ago. 2015. p. 1721-1730.

CHEN, X.; HSIEH, C.; GONG, B. When Vision Transformers Outperform ResNets without Pre-training or Strong Data Augmentations. International Conference on Learning Representations (ICLR), 2022.

CHEN, Y.; HUANG, P.; LIN, K.; LIN, H.; WANG, L.; CHENG, C.; CHEN, T.; CHAN, Y.; CHIANG, J. Y. Semi-Automatic Segmentation and Classification of Pap Smear Cells. **IEEE Journal of Biomedical and Health Informatics**, v. 18, n. 1, jan. 2014. DOI: 10.1109/JBHI.2013.2250984.

CIREŞAN, D.C.; GIUSTI, A.; GAMBARDELLA, L. M.; SCHMIDHUBER, J. Mitosis detection in breast cancer histology images with deep neural networks. **Medical Image Computing and Computer-Assisted Intervention**, v. 16, p. 411–418, 2013. DOI: 10.1007/978-3-642-40763-5_51.

DEVLIN, J.; CHANG, M.; LEE, K.; TOUTANOVA, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. **North American Chapter of the Association for Computational Linguistics** (NAACL), maio 2019. DOI: 10.48550/arXiv.1810.04805.

DOSOVITSKIY, A; BEYER, L.; KOLESNIKOV, A.; WEISSENBORN, D.; ZHAI, X.; UNTERTHINER, T.; DEHGHANI, M.; MINDERER, M.; HEIGOLD, G.; GELLY, S.; USZKOREIT, J.; HOULSBY, N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. **International Conference on Learning Representations** (ICLR), jun. 2021. DOI: 10.48550/arXiv.2010.11929

GENÇTAV, A.; AKSOY, S.; ÖNDER, S. Unsupervised segmentation and classification of cervical cell images. **Pattern Recognition**, v. 45, n. 12, p. 4151-4168, dec. 2012. DOI: 10.1016/j.patcog.2012.05.006.

GONFALONIERI, A. 2019. **What is an AI Algorithm?** *Medium.* Disponível em: https://medium.com/predict/what-is-an-ai-algorithm-aceeab80e7e3. Acesso em: 05 jul. 2024.

HAN, K.; WANG, Y.; CHEN, H.; CHEN, X.; GUO, J.; LIU, Z.; TANG, Y.; XIAO, A.; XU, C.; XU, Y.; YANG, Z.; ZHANG, Y.; TAO, D. A Survey on Visual Transformer. **IEEE Transaction on Pattern Analysis and Machine Intelligence**, jul. 2023. DOI: 10.48550/arXiv.2012.12556.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep Residual Learning for Image Recognition. **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**, p. 770-778, 2016. Disponível em: https://arxiv.org/abs/1512.03385. Acesso em: 30 jul. 2024. DOI: 10.48550/arXiv.1512.03385.

HUANG, G.; LIU, Z.; MAATEN, L.; WEINBERGER, K. Q. Densely Connected Convolutional Networks. **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition** (Honolulu, HI: IEEE), p. 4700–4708, 2017. Disponível em: https://arxiv.org/abs/1608.06993. Acesso em: 30 jul. 2024. DOI: 10.48550/arXiv.1608.06993.

HUSSAIN, E.; MAHANTA, L. B.; BORAH, H.; DAS, C. R. Liquid based-cytology Pap smear dataset for automated multi-class diagnosis of pre-cancerous and cervical cancer lesions. **Data in Brief**, v. 30, p. 105589, jun. 2020. Disponível em: https://www.data-in-brief.com/article/S2352-3409(20)30483-2/fulltext. Acesso em: 30 jul. 2024. DOI: 10.1016/j.dib.2020.105589.

FRAPPART, L. *et al*. **Histopathology and cytopathology of the uterine cervix**: digital atlas. Lyon, France: International Agency for Research on Cancer: World Health Organization, 2004. *E-book*. Disponível em: https://www.ncbi.nlm.nih.gov/nlmcatalog?cmd=PureSearch&term=101240825%5Bnlmid%5D. Acesso em: 30 jul. 2024.

JAVAID, M.; HALEEM, A.; SINGH, R.P.; ChatGPT for healthcare services: An emerging stage for an innovative perspective. **BenchCouncil Transactions on Benchmarks, Standards and Evaluations**, v. 3, p. 100105, 2023. DOI: 10.1016/j.tbench.2023.100105.

JANTZEN, J.; NORUP, J.; DOUNIAS, G.; BJERREGAARD, B. Pap-smear benchmark data for pattern classification. **Nature inspired Smart Information Systems** (NiSIS), p. 1–9, 2005. Disponível em: https://mde-lab.aegean.gr/index.php/downloads/. Acesso em: 30 jul. 2024.

JIANG, P.; LI, X.; SHEN, H.; CHEN, Y.; WANG, L.; CHEN, H.; FENG, J.; LIU, J. A systematic review of deep learning-based cervical cytology screening: from cell identification to whole slide image analysis. **Artificial Intelligence Review**, v. 56, p. S2687–S2758, 2023. DOI: 10.1007/s10462-023-10588-z.

KAPLAN, A.; HAENLEIN, M. Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. **Business Horizons**, v. 62, n. 1, p. 15–25, 2019. DOI: 10.1016/j.bushor.2018.08.004.

KESSLER, T. A. Cervical Cancer: Prevention and Early Detection. **Seminars in Oncology Nursing**, v. 33, n. 2, p. 172-183, maio 2017. DOI: 10.1016/j.soncn.2017.02.005.

KOLIOPOULOS, G.; ARBYNB, M.; MARTIN-HIRSCHD, P.; KYRGIOUE, M.; PRENDIVILLE, W.; PARASKEVAIDIS, E. Diagnostic accuracy of human papillomavirus testing in primary cervical screening: A systematic review and meta-analysis of non-randomized studies. **Gynecologic Oncology**, v. 104, p. 232-246, 2007. DOI: 10.1016/j.ygyno.2006.08.053.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. ImageNet Classification with Deep Convolutional Neural Networks. **Advances in Neural Information Processing Systems** (NIPS), p. 25, 2012. DOI: 10.1145/3065386.

KUMAR, P.; LUO, S.; SHAUKAT, K. A Comprehensive Review of Deep Learning Approaches for Animal Detection on Video Data. **International Journal of Advanced Computer Science and Applications**, v. 14, n. 11, 2023. DOI: 10.14569/IJACSA.2023.01411144.

KUO, C. C. J. Understanding Convolutional Neural Networks with a Mathematical Model. **Journal of Visual Communication and Image Representation**, p. 406-413, nov. 2016. DOI: 10.48550/arXiv.1609.04112.

LE CUN, Y.; BOSER, B.; DENKER, J. S.; HENDERSON, D.; HOWARD, R. E.; HUBBARD, W.; JACKEL, L. D. Handwritten digit recognition with a back-propagation network. **Advances in Neural Information Processing Systems**, p. 396-404, jun. 1990.

LE CUN, Y.; BOTTOU, L.; BENGIO, Y.; HAFFNER, P. Gradient-Based Learning Applied to Document Recognition. **Proceedings of the IEEE**, v. 86, n. 11, p. 2278-2324, nov. 1998. DOI: 10.1109/5.726791.

LIMA, D. N. O. **Atlas de citopatologia ginecológica.** Centro de Estudos e Pesquisa em Saúde Coletiva (CEPESC), 1ª Edição, 204p., 2012. *E-book*. Disponível em: https://bvsms.saude.gov.br/bvs/publicacoes/atlas_citopatologia_ginecologica.pdf. Acesso em: 30 jul. 2024.

LIMA, D. N. O.; BARROS, A. L. S.; OLIVEIRA, M. L.; AZEVEDO, M. D. **Caderno de referência 1: Citopatologia Ginecológica**, Centro de Estudos e Pesquisa em Saúde Coletiva (CEPESC), 194p., 2012. *E-book*. Disponível em: https://bvsms.saude.gov.br/bvs/publicacoes/tecnico_citopatologia_caderno_referencia_1.pdf. Acesso em: 30 jul. 2024.

LIU, W.; LI, C.; XU, N.; JIANG, T.; RAHAMAN, M.; SUN, H.; WU, X.; HU, W.; CHEN, H.; SUN, C.; YAO, Y.; GRZEGORZEK, M. CVM-Cervix: A hybrid cervical Pap-smear image classification framework using CNN, visual transformer and multilayer perceptron. **Pattern Recognition**, v. 130, p. 108829, jun. 2022. DOI: 10.48550/arXiv.2206.00971.

MATIAS, A. V.; AMORIM, J. G. A.; MACARINI, L. A. B.; CERENTINI, A.; ONOFRE, A. S. C.; ONOFRE, F. B. de M.; DALTOÉ, F. P.; STEMMER, M. R.; WANGENHEIM, A. von. What is the state of the art of computer vision-assisted cytology? A Systematic Literature Review**. Computerized Medical Imaging and Graphics**, v. 91, p. 101934, 2021. DOI: 10.1016/j.compmedimag.2021.101934.

MARINAKIS, Y.; DOUNIAS, G.; JANTZEN, J. Pap smear diagnosis using a hybrid intelligent scheme focusing on genetic algorithm based feature selection and nearest neighbor classification. **Computers in Biology and Medicine**, v. 39, p. 69-78, 2009. DOI: 10.1016/j.compbiomed.2008.11.006.

BRASIL. Ministério da Saúde. Coordenação de Monitoramento e Avaliação de Tecnologias em Saúde - CMATS/DGITIS/SCTIE/MS. **Relatório de recomendação**: citologia em meio líquido para rastreamento de câncer de colo de útero e lesões precursoras. Brasília, DF: Coordenação De Monitoramento e Avaliação de Tecnologias em Saúde - CMATS/ DGITIS/ SCTIE/ MS, 2019. Disponível em: https://www.gov.br/conitec/pt-br/midias/consultas/relatorios/2019/relatorio_citologialiquida_cancerutero_cp59_2019.pdf. Acesso em: 30 jul. 2024.

NIAZI, M. K. K.; PARWANI, A. V.; GURCAN, M. N. Digital pathology and artificial intelligence. **Lancet Oncology**, v. 20, p. e253-61, 2019. DOI: 10.1016/S1470-2045(19)30154-8.

PACAL, I. MaxCerVixT: A novel lightweight vision transformer-based Approach for precise cervical cancer detection. **Knowledge-Based Systems**, v. 289, p. 111482, fev. 2024. DOI: 10.1016/j.knosys.2024.111482.

PAUL, S.; CHEN, P. Vision Transformers are Robust Learners. **AAAI Conference on Artificial Intelligence**, 2022. DOI: 10.48550/arXiv.2105.07581.

PLISSITI, M. E.; DIMITRAKOPOULOS, P.; SFIKAS, G.; NIKOU, C.; KRIKONI, O.; CHARCHANTI, A. Sipakmed: A new dataset for feature and image based classification of normal and pathological cervical cells in pap smear images. **25th IEEE International Conference on Image Processing** (ICIP), p. 3144–3148, out. 2018. Disponível em: https://www.kaggle.com/datasets/prahladmehandiratta/cervical-cancer-largest-dataset-sipakmed. Acesso em: 30 jul. 2024. DOI: 10.1109/ICIP.2018.8451588.

REZENDE, M. T.; SILVA, R.; BERNARDO, F. O.; TOBIAS, A. H.; OLIVEIRA, P. H.; MACHADO, T. M.; COSTA, C. S.; MEDEIROS, F. N.; USHIZIMA, D. M.; CARNEIRO, C. M.; BIANCHI, A. G. C. Cric searchable image database as a public platform for conventional pap smear cytology data. **Scientific Data - Nature,** v. 8, n. 1, p. 151, jun. 2021. Disponível em: https://database.cric.com.br/downloads. Acesso em: 30 jul. 2024. DOI: 10.1038/s41597-021-00933-8.

SELVARAJU, R. R.; COGSWELL, M.; DAS, A.; VEDANTAM, R.; PARIKH, D.; BATRA, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. **Journal of Computer Vision**, dez. 2019. Disponível em: https://arxiv.org/abs/1610.02391. Acesso em: 30 jul. 2024. DOI: 10.48550/arXiv.1610.02391.

SIMONYAN, K.; ZISSERMAN, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. **Computer Vision and Pattern Recognition**, set. 2014. Disponível em: https://arxiv.org/abs/1409.1556. Acesso em: 30 jul. 2024. DOI: 10.48550/arXiv.1409.1556.

SINGHA, A.; THAKUR, R. S.; PATEL, T. Deep Learning Applications in Medical Image Analysis. **Biomedical Data Mining for Information Retrieval:** Methodologies, Techniques and Applications, p. 293-350, aug. 2021. Disponível em: https://www.researchgate.net/publication/353702181_Deep_Learning_Applications_in_Medical_Image_Analysis. Acesso em: 30 jul. 2024. DOI: 10.1002/9781119711278.ch11.

SZEGEDY, C.; LIU, W.; JIA, Y.; SERMANET, P.; REED, S.; ANGUELOV, D.; ERHAN, D.; VANHOUCKE, V.; RABINOVICH, A. Going Deeper with Convolutions. **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition** (Boston, MA: IEEE), p. 1–9, 2015. Disponível em:

https://arxiv.org/abs/1409.4842. Acesso em: 30 jul. 2024. DOI: 10.48550/arXiv.1409.4842.

SZEGEDY, C.; VANHOUCKE, V.; IOFFE, S.; SHLENS, J.; WOJNA, Z. Rethinking the Inception Architecture for Computer Vision. **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition** (Las Vegas, NV: IEEE), p. 2818–2826, 2016. Disponível em: https://arxiv.org/abs/1512.00567. Acesso em: 30 jul. 2024. DOI: 10.48550/arXiv.1512.00567.

TAN, M.; LE, Q. V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. **International Conference on Machine Learning** (Taiyuan: PMLR), p. 6105–6114, 2019. Disponível em: https://arxiv.org/abs/1905.11946. Acesso em: 30 jul. 2024. DOI: 10.48550/arXiv.1905.11946.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L.; POLOSUKHIN, I. Attention is All you Need. **Advances in Neural Information Processing Systems** (NIPS), v. 30, 2017. DOI: 10.48550/arXiv.1706.03762.

WAGGONER, S. E. Cervical cancer. **The Lancet**, v. 361, jun. 2003. DOI: 10.1016/S0140-6736(03)13778-6.

WILD, C.P.; WEIDERPASS, E.; STEWART, B.W. World Cancer Report: Cancer Research for Cancer Prevention. **International Agency for Research on Cancer,** Lyon, France, 2020. *E-book*. Disponível em: http://publications.iarc.fr/586. Acesso em: 30 jul. 2024.

WILLIAM, W.; WARE, A.; BASAZA-EJIRI, A. H.; OBUNGOLOCH, J. A review of image analysis and machine learning techniques for automated cervical cancer screening from pap-smear images. **Computer Methods and Programs in Biomedicine**, v. 164, p. 15-22, 2018. DOI: 10.1016/j.cmpb.2018.05.034.

WILLIAM, W.; WARE, A.; BASAZA-EJIRI, A. H.; OBUNGOLOCH, J. A pap-smear analysis tool (PAT) for detection of cervical cancer from pap-smear images. **BioMedical Engineering OnLine**, v. 18, p. 16, 2019. DOI: 10.1186/s12938-019-0634-5.

WU, J. Introduction to Convolutional Neural Networks**. National Key Lab for Novel Software Technology**, maio 2017.

ZEUNE, L.L.; BOINK, Y. E.; DALUM, G. van; NANOU, A.; WIT, S. de; ANDREE, K. C.; SWENNENHUIS, J. F.; GILS, S. A. van; TERSTAPPEN, L. W. M. M.; BRUNE, C. Deep learning of circulating tumour cells. **Nature Machine Intelligence**, v. 2, p. 124–133, fev. 2020. DOI: 10.1038/s42256-020-0153-x.

ZHANG, C.; LU Y. Study on artificial intelligence: The state of the art and future prospects. **Journal of Industrial Information Integration**, v. 23, 2021. DOI: 10.1016/j.jii.2021.100224.

ZHAO, R.; YAN, R.; CHEN, Z.; MAO, K.; WANG, P.; GAO, R. X. Deep learning and its applications to machine health monitoring. **Mechanical Systems and Signal Processing**, v. 115, p. 213-237, jan. 2019. DOI: 10.1016/j.ymssp.2018.05.050.

ZHAO, Z. Q.; ZHENG, P.; XU, S. T.; WU, X. Object detection with deep learning: a review. **IEEE Trans Neural Netw Learn Syst.,** v. 30, n. 11, p. 3212-3232, nov. 2019. DOI: 10.48550/arXiv.1807.05511.

ZHOU, B.; KHOSLA, A.; LAPEDRIZA, A.; OLIVA, A.; TORRALBA, A. Learning Deep Features for Discriminative Localization. **Computer Vision and Pattern Recognition**, dez. 2015. Disponível em: https://arxiv.org/abs/1512.04150. Acesso em: 30 jul. 2024. DOI: 10.48550/arXiv.1512.04150.

ZHOU, Z.; SIDDIQUEE, M. R.; TAJBAKHSH, N.; LIANG, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. **Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support** (Granada: Springer), p. 3–11, 2018. Disponível em: https://arxiv.org/abs/1807.10165. Acesso em: 30 jul. 2024. DOI: 10.48550/arXiv.1807.10165.

APÊNDICE A - Publicações da Dissertação

Como resultado da dissertação, o artigo intitulado “Inteligência artificial aplicada ao diagnóstico de câncer por exames de imagem” foi publicado, que reflete diretamente os achados obtidos ao longo das etapas iniciais descritas no trabalho**.** Especificamente, etapa 1 da revisão de literatura e definição do projeto, assim como etapa 2 na definição de conjuntos de imagens e etapa 6 na definição de métricas. O artigo aborda os principais resultados alcançados, reforçando a conexão entre a metodologia adotada e os dados obtidos durante a pesquisa. Essa publicação não só valida as conclusões apresentadas, mas também contribui para o avanço do conhecimento na área, destacando a relevância dos esforços empreendidos na dissertação. Publicado na **Revista Eletrônica Acervo Científico**, ISSN 2595-7899, Volume 47, 2024, classificada como Qualis B3 no período de 2017-2020, disponível em: https://doi.org/10.25248/reac.e16193.2024.

A seguir o texto do artigo publicado:

Inteligência Artificial Aplicada ao Diagnóstico de Câncer por Exames de Imagem

RESUMO

Objetivo: Revisar a aplicação da Inteligência Artificial (IA) no diagnóstico de câncer através de exames de imagem, analisando tendências e avanços recentes na intersecção entre IA e medicina diagnóstica. **Métodos:** Foi realizada uma pesquisa bibliográfica em bases de dados eletrônicas, selecionando artigos científicos, revisões sistemáticas e meta-análises publicados entre 2017 e 2023. Os estudos incluíram algoritmos de IA, técnicas de aprendizado de máquina e redes neurais aplicadas a exames de imagem para detecção, classificação e diagnóstico de câncer. **Resultados:** A análise focou em modelos de reconhecimento de imagem para diagnóstico de câncer, priorizando métricas de sensibilidade e especificidade. Foram destacados estudos que compararam o desempenho de radiologistas com sistemas de IA mostrando que em alguns casos a IA superou os profissionais e, em outros, melhorou significativamente o desempenho dos radiologistas quando usada como assistência. **Conclusão:** A IA mostrou-se uma ferramenta promissora no diagnóstico de câncer por imagem. A combinação de IA com dados clínicos pode melhorar as métricas de diagnóstico. Limitações incluem a qualidade e quantidade de imagens para treinamento, mas novas tecnologias como IA Generativa estão surgindo para superar esses desafios.

Palavras-chave: Câncer, Diagnóstico, Aprendizado Profundo, Inteligência Artificial Generativa.

ABSTRACT

Objective: To review the application of Artificial Intelligence (AI) in cancer diagnosis through imaging tests, analyzing trends and recent advances at the intersection of AI and diagnostic medicine. **Methods:** A bibliographic search was conducted in electronic databases, selecting scientific articles, systematic reviews, and meta-analyses published between 2017 and 2023. The studies included AI algorithms, machine learning techniques, and neural networks applied to imaging tests for cancer detection, classification, and diagnosis. **Results:** The analysis focused on image recognition models for cancer diagnosis, prioritizing sensitivity and specificity metrics. Studies that compared the performance of radiologists with AI systems were highlighted, showing that in some cases AI outperformed the professionals and, in others, significantly improved the performance of radiologists when used as assistance. **Conclusion:** AI has proven to be a promising tool in cancer diagnosis through imaging. The combination of AI with clinical data may improve diagnostic metrics. Limitations include the quality and quantity of images for training, but new technologies such as Generative AI are emerging to overcome these challenges.

Key words: Cancer, Diagnosis, Deep Learning, Generative Artificial Intelligence.

RESUMEN

Objetivo: Revisar la aplicación de la Inteligencia Artificial (IA) en el diagnóstico de cáncer a través de exámenes de imagen, analizando tendencias y avances recientes en la intersección entre IA y medicina diagnóstica. **Métodos:** Se realizó una búsqueda bibliográfica en bases de datos electrónicas, seleccionando artículos científicos, revisiones sistemáticas y metaanálisis publicados entre 2017 y 2023. Los estudios incluyeron algoritmos de IA, técnicas de aprendizaje automático y redes neuronales aplicadas a exámenes de imagen para la detección, clasificación y diagnóstico de cáncer. **Resultados:** El análisis se centró en modelos de reconocimiento de imagen para el diagnóstico de cáncer, priorizando métricas de sensibilidad y especificidad. Se destacaron estudios que compararon el rendimiento de radiólogos con sistemas de IA, mostrando que en algunos casos la IA superó a los profesionales y, en otros, mejoró significativamente el rendimiento de los radiólogos cuando se utilizó como asistencia. **Conclusión:** La IA se ha mostrado como una herramienta prometedora en el diagnóstico de cáncer por imagen. La combinación de IA con datos clínicos puede mejorar las métricas de diagnóstico. Las limitaciones incluyen la calidad y cantidad de imágenes para entrenamiento, pero nuevas tecnologías como la IA Generativa están surgiendo para superar estos desafíos.

Palabras clave: Cáncer, Diagnóstico, Aprendizaje Profundo, Inteligencia Artificial Generativa**.**

INTRODUÇÃO

O termo Inteligência Artificial (IA) foi criado em uma conferência em Dartmouth de 1956 por John McCarthy e foi lá que toda a área de pesquisa se originou. Kaplan e Haenlein definiram a IA como "a capacidade de processar dados externos de forma sistemática e aprender com eles para alcançar metas e tarefas específicas" (Puppe F, 1992; Kaplan A e Haenlein M, 2019).

A IA passa a ser aplicada na Medicina no início da década de 70, seu objetivo era aumentar a eficiência do diagnóstico e tratamento médico com este auxílio de sistemas de IA. A partir da década de 80, Liu R, et al. (2020) afirmam que a IA na Medicina poderia ser dividida aproximadamente em 4 estágios: a infância como primeiro estágio (1980) com árvores de decisão e o desenvolvimento das

redes neurais; o segundo, a adolescência (1990) com os algoritmos das Máquina de Vetores de Suporte (*Support Vector Machine - SVM*); a maturidade no terceiro estágio (2000) quando o conceito de *Deep Learning* ou "Aprendizado Profundo" foi proposto, consistindo em uma rede neural com várias camadas; e por fim o quarto estágio, o período de maturação (2010) no qual o evento mais recente seria a publicação da Google da nova arquitetura *Transformer* de redes neurais com os mecanismos de atenção ou autoatenção em 2017, arquitetura essa que deu origem ao *ChatGPT* entre outros, mudando o atual paradigma da IA (Vaswani A, et al., 2017).

O reconhecimento de imagens é a tecnologia de processamento e análise de imagens por meio de computadores. É uma tecnologia importante no campo da IA de treinamento Supervisionado, que utiliza exames de imagens com diagnóstico médico ou laboratorial para treinar o modelo, o qual se baseia no Aprendizado Profundo. Segundo Masood A, et al. (2018), o processo de reconhecimento pode ser dividido em cinco etapas, sendo elas: (1) processamento de entrada, (2) pré-processamento de imagens, (3) extração de imagens, (4) construção de classificadores e (5) produção de saída ().

Atualmente, a tecnologia de reconhecimento de imagens pode ser aplicada em muitas áreas médicas e existem muitos equipamentos de exames que já a utilizam. O câncer cervical, por exemplo, é uma das quatro principais causas de morte em mulheres, e é causado principalmente por infecção pelo papilomavírus humano (HPV). As pacientes não apresentam sintomas óbvios nos estágios iniciais. Embora muitos tratamentos estejam disponíveis para pacientes com câncer cervical, como cirurgia, radioterapia e quimioterapia, o prognóstico das pacientes depende muito do diagnóstico precoce do câncer. O Aprendizado Profundo, no reconhecimento de lesões em imagens do colo do útero, poderia ajudar os médicos no diagnóstico precoce do câncer cervical com uma taxa de precisão de aproximadamente 90%. Além disso, sistemas de reconhecimento de imagens poderiam auxiliar médicos com uma segunda opinião em situações mais complexas, e assim produzir diagnósticos mais precisos (Lin H, et al., 2021; Liu J, et al., 2018).

O Aprendizado Profundo desempenha um papel importante na aplicação da tecnologia de reconhecimento de imagens para identificar lesões. Kooi T, et al. (2017) utilizaram redes neurais convolucionais (*Convolutional Neural Networks - CNNs*) em um estudo para identificar lesões malignas na mama. Este método teve um desempenho superior ao sistema mais avançado em detecção assistida por computador, uma vez que a precisão de reconhecimento da CNN foi maior (Liu F, et al., 2018).

O reconhecimento de imagens para diagnóstico com o auxílio da IA só se torna possível por meio da imagem médica. A imagem médica é o processo de representação visual da estrutura e função de diferentes tecidos e órgãos do corpo humano para fins clínicos e para o estudo detalhado da anatomia e fisiologia normais e anormais do corpo na ciência médica. As técnicas de imagem médica são usadas para mostrar estruturas internas sob a pele e os ossos, bem como para diagnosticar anormalidades e tratar doenças. A imagem médica transformou-se em uma ciência na área da saúde. É uma parte importante da imagem biológica e inclui a radiologia, que utiliza tecnologias de imagem como radiografia de raios X, tomografia computadorizada por raios X (CT), endoscopia, ressonância magnética (RM), espectroscopia de ressonância magnética (ERM), tomografia por emissão de pósitrons (PET), termografia, fotografia médica, imagem elétrica de fonte (ESI), mamografia digital, imagem tátil, imagem magnética de fonte (MSI), imagem óptica médica, tomografia computadorizada por emissão de fóton único (SPECT) e tomografia computadorizada por impedância elétrica (EIT) (Hussain S, et al., 2022).

Mas, ainda existem limitações, devido aos diferentes equipamentos utilizados em diversos hospitais, a resolução das imagens adquiridas pode variar bastante, o que pode impactar a eficiência do modelo. Um modelo de Aprendizado Profundo também necessita de uma grande quantidade de imagens e requer muito poder de processamento. Problemas estes que a recente arquitetura do modelo *Transformer* desenvolvida pelo Google promete resolver ao melhorar a eficiência

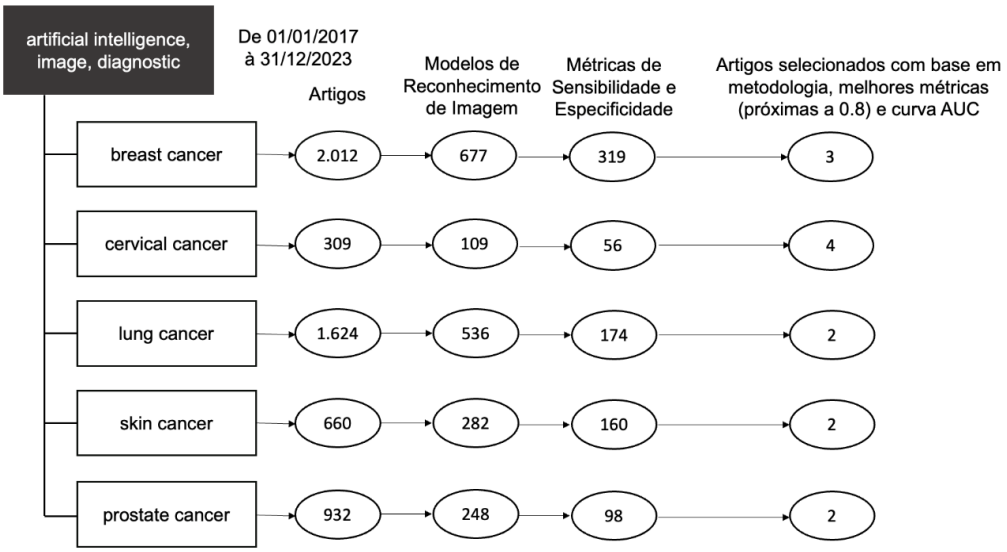
computacional e ao diminuir a necessidade na quantidade de imagens com seu mecanismo de atenção ou autoatenção (Vaswani A, et al., 2017). Este trabalho tem então como objetivo analisar as tendências e avanços recentes na intersecção entre IA e medicina diagnóstica.

MÉTODOS

Este trabalho consiste em uma revisão da aplicação da Inteligência Artificial (IA) no diagnóstico de câncer por meio de exames de imagem. A pesquisa foi conduzida com o objetivo de analisar as tendências e avanços recentes na intersecção entre IA e medicina diagnóstica. A revisão bibliográfica foi realizada em bases de dados eletrônicas, incluindo PubMed, arXiv, IEEE Xplore, e Google Scholar. Foram considerados artigos científicos, revisões sistemáticas e meta-análises publicados entre 2017 e 2023 que abordavam o uso de IA em diagnósticos oncológicos por meio de imagens médicas. A escolha do intervalo de tempo se deve ao fato de que em 2017 foi quando a última competição da ImageNet foi realizada porque julgaram que o estado da arte com redes neurais convolucionais para reconhecimento de imagens já havia sido alcançado, também é o ano em que o artigo do algoritmo *Transformer* foi publicado provocando a mudança de paradigma (Mishkin D, et al., 2017).

Os artigos para análise incluíram estudos que exploraram algoritmos de IA, técnicas de aprendizado de máquina e redes neurais aplicadas a exames de imagem para detecção, classificação e diagnóstico de câncer. Utilizou-se os descritores (palavras-chave): *artificial intelligence* e *image* e *diagnostic* alternando com os seguintes descritores *breast cancer* ou *cervical cancer* ou *lung cancer* ou *skin cancer* ou *prostate cancer* em inglês para a busca, pois as terminologias e tecnologias originais são em inglês. Estudos que não continham as métricas de sensibilidade e especificidade ou não utilizavam modelos de *Deep Learning* ou *Machine Learning* para reconhecimento de imagem foram excluídos do estudo. A inclusão dos estudos se baseou em estudos que demonstraram resultados promissores ou inovações significativas na área. A análise dos dados concentrou-se na avaliação crítica das metodologias utilizadas nos estudos selecionados, destacando as abordagens específicas de IA, conjunto de dados empregado, métricas de sensibilidade e especificidade para desempenho e conclusões relevantes para o diagnóstico de câncer por imagem. A Figura 1 descreve o processo de seleção de artigos e os artigos encontrados.

Figura 1 – Artigos revisados e selecionados para a revisão.



A revisão está sujeita a limitações inerentes à disponibilidade de literatura e à constante evolução da tecnologia. Ademais, a interpretação dos resultados também depende da qualidade e metodologia dos estudos revisados.

RESULTADOS

Foram analisados resultados e métricas de modelos de reconhecimento de imagem desenvolvidos para diagnóstico de câncer em áreas específicas, selecionados pela maior incidência de casos e pela utilização de redes neurais de Aprendizado Profundo e das métricas de sensibilidade e especificidade. O Instituto Nacional do Câncer (INCA) lançou recentemente uma estimativa de 704 mil casos de câncer por ano no Brasil até 2025, sendo a neoplasia maligna de maior incidência a de pele não melanoma (31,3%), seguida de tumores malignos de mama (10,5%), próstata (10,2%), cólon e reto (6,5%), pulmão (4,6%) e estômago (3,1%). No entanto, em regiões menos desenvolvidas, de menor IDH, o câncer cervical ocupa o terceiro lugar para mulheres no lugar do câncer colorretal, inclusive estudos já foram conduzidos analisando essa discrepância mostrando que em lugares menos desenvolvidos as mulheres eram mais resistentes a fazer o exame assim como não há muito pouca ou nenhuma infraestrutura (Santos MO, et al., 2023; Arbyn M, et al., 2020).

As métricas de sensibilidade e especificidade foram priorizadas por serem consideradas altamente relevantes na área médica por fornecerem informações cruciais sobre o desempenho de testes diagnósticos, procedimentos de triagem e modelos preditivos em termos de sua capacidade de identificar corretamente as condições de saúde (Junior GBV, et al., 2022).

1. Sensibilidade: Também conhecida como *Recall*. A sensibilidade de um teste médico refere-se à sua capacidade de identificar corretamente os indivíduos que têm a doença (verdadeiros positivos). Um teste com alta sensibilidade é importante para garantir que poucos diagnósticos críticos como câncer sejam perdidos (falsos negativos).

2. Especificidade: A especificidade de um teste é a sua capacidade de identificar corretamente os indivíduos que não têm a doença (verdadeiros negativos). Um teste com alta especificidade é importante para evitar o diagnóstico incorreto de pessoas saudáveis como doentes (falsos positivos). Isso é especialmente importante para evitar tratamentos desnecessários ou invasivos e a ansiedade associada a um diagnóstico incorreto.

3. AUC: *Area Under the Curve*, se refere à precisão do modelo. Um modelo cujas previsões estão 100% erradas tem uma AUC de 0, enquanto um modelo cujas previsões são 100% corretas tem uma AUC de 1. A curva AUC é derivada da curva ROC (*Receiver Operating Characteristic*) que utiliza a taxa de verdadeiro positivo e a taxa de falso positivo.

A Tabela 1 ilustra uma síntese dos melhores resultados encontrados na pesquisa de artigos de modelos de reconhecimento de imagem para diagnóstico de câncer, assim como outras revisões sistemáticas. Os modelos listados foram selecionados com base em 2 métricas principais: sensibilidade e especificidade. Quanto mais próximo de 1 (ou 100%) melhor a assertividade do modelo. Por exemplo, uma sensibilidade de 0,90 indicaria que de cada 10 exames de imagem com diagnóstico positivo para o câncer, o algoritmo erraria 1 diagnóstico ao informar que o paciente não está com câncer, um falso negativo. Já para uma especificidade de 0,90, de cada 10 exames de imagem com diagnóstico negativo, o modelo estaria errando ao diagnosticar o paciente informando que ele está com câncer, sem na realidade ele estar com câncer, um falso positivo.

Tabela 1 – Síntese das melhores métricas encontradas em modelos de reconhecimento de imagem para diagnóstico de câncer em artigos, São Paulo, 2023.

Neoplasia Maligna	Exame de Imagem	Modelo	Sensibilidade	Especificidade	AUC

1	Pulmão	Tomografia Computadorizada	DFCNet	0,837	0,962	n/a
2	Pulmão	Tomografia Computadorizada	Radiomics + DCNN	0,940	0,930	0,950
3	Não Melanoma	Imagem Dermatoscópica	EfficientNetB4	0,667	0,987	0,990
4	Não Melanoma	Imagem Dermatoscópica	CNN	0,84	0,98	0,971
5	Mama	Mamografia	ResNet50	0,839	0,864	0,945
6	Mama	Mamografia	CNN**	0,819	0,966	0,956
7	Mama	Mamografia	CNN**	0,914	0,860	0,959
8	Colo do Útero	Citologia Cervical	VGG-19	0,897	0,892	0,950
9	Colo do Útero	Citologia Cervical	ResNet50	0,992	0,997	0,996
10	Colo do Útero	Citologia Cervical	DPNet	0,907	0,800	0,925
11	Colo do Útero	Citologia Cervical	CNN**	0,957	0,847	0,940
12	Próstata	Ressonância Magnética	MulmoUNet	0,727	0,955	n/a
13	Próstata	Ressonância Magnética	Inception V1	0,815	0,820	0,889

Fonte: Cada linha da tabela é referente a um artigo e um modelo: Masood A, et al., 2018(1); Zhang X, et al., 2023(2); Huang HW, et al., 2021(3); Kuo KM, et al., 2023(4); Lotter W, et al., 2021(5); Salim M, et al., 2020(6); Kim HE, et al., 2020(7); Sornapudi S, et al., 2019(8); Benyes YK, et al., 2022(9); Lin H, et al., 2021(10); Holmström O, et al., 2021(11); Matsuoka Y, et al., 2023(12); Khosravi P, et al., 2021(13); **Modelos Comerciais.

Alguns desses modelos comparam a performance do diagnóstico dado por um radiologista e o mesmo diagnóstico dado por um modelo de IA de forma independente. Os sistemas autônomos têm o potencial de filtrar exames provavelmente normais, diminuindo a carga de trabalho dos radiologistas. Já os sistemas de assistência são utilizados durante a interpretação dos exames.

As métricas de sensibilidade e especificidade podem ser aplicadas também a diagnósticos fornecidos por profissionais da saúde, técnicos e médicos, não apenas a sistemas de computador. Assim como o computador erra o diagnóstico, o profissional de saúde também pode errar. Alguns sinais de tumores em exames de imagem podem muitas vezes passarem despercebidos ao olho humano. Ao mesmo tempo, o modelo de computador para reconhecimento de imagem sempre será tão bom quanto os médicos e profissionais de saúde que alimentam o modelo com exemplos de seus diagnósticos. Para melhorar o sistema, exames adicionais podem ser acrescentados como uma biópsia para confirmar diagnósticos. A partir desse exames adicionais o modelo passaria a procurar

padrões para identificar problemas na imagem que podem ter passados despercebidos ao olho humano.

Tabela 2 – Comparação relativa entre profissionais de saúde e sistemas de IA nas melhores performances de diagnóstico, São Paulo, 2023.

Artigo	Exame de Imagem	Comparação
Kim HE, et al., 2020	Mamografia	**Radiologistas x Sistema de IA:** em média melhora de 16,13% na sensibilidade e 14,04% na especificidade para o sistema de IA autônomo
McKinney SM, et al., 2020	Mamografia	**Radiologistas x Sistema de IA:** nos EUA em média melhora de 9,4% na sensibilidade e 4,7% na especificidade para o sistema de IA autônomo
Lotter W, et al., 2021	Mamografia	**Radiologistas x Sistema de IA:** em média melhora de 14,2% na sensibilidade e 24% na especificidade para o sistema de IA autônomo

Como ilustra a Tabela 2 em muitos casos os sistemas de IA melhoram a performance do diagnóstico de câncer. Kim HE, et al. (2020) relataram uma avaliação independente de um sistema de IA comercial (Lunit INSIGHT MMG) em um estudo de grande escala que incluiu 5 instituições na Coreia do Sul, EUA e Reino Unido, demonstrando alto desempenho diagnóstico. O sistema de IA superou todos os radiologistas quando usado de forma autônoma. Além disso, quando usado simultaneamente, observou-se uma melhoria significativa no desempenho dos radiologistas com a ajuda da IA.

Lotter W, et al. (2021) usaram imagens de diferentes estados dos EUA e um conjunto de dados de um hospital urbano chinês, com resultados promissores, onde o modelo alcançou um aumento absoluto na sensibilidade de 14,2% e na especificidade de 24% em comparação com a sensibilidade média alcançada pelo radiologista.

Salim M, et al. (2020) realizaram uma avaliação externa de três sistemas de IA comerciais para detecção de câncer de mama (não aprovados pela FDA na época e mantidos anônimos), tentando determinar se algum deles poderia ter desempenho igual ou superior ao dos radiologistas em diagnósticos de mamografia. Um dos sistemas mostrou-se superior aos radiologistas que realizaram a primeira leitura, modelo este que foi listado na tabela, com potencial para atuar como um leitor independente no futuro. A combinação desse sistema com os radiologistas da primeira leitura resultou nas avaliações mais precisas e alcançou as pontuações mais altas.

Holmström O, et al. (2021) conduziram um estudo em uma clínica rural no Kenya, um país onde o número de mortes entre mulheres por câncer cervical era o mais elevado, com a proposta de desenvolver um sistema barato e acessível para diagnóstico remoto. Assim como todos os algoritmos de Aprendizado Supervisionado, modelos de CNN necessitam de imagens anotadas para classificar corretamente. No caso do estudo o modelo treinado apresentou a mesma performance dos patologistas locais, os quais foram responsáveis por anotar as imagens para o treinamento do modelo. As interpretações visuais para exames de Papanicolau são conhecidamente subjetivas, por conta disso houve diferenças com as avaliações dos patologistas remotos que analisaram as lâminas digitalizadas. No caso só um exame de biópsia cervical poderia confirmar o diagnóstico. Já os modelos desenvolvidos por Zhang X, et al. (2023) e Matsuoka Y, et al. (2023) contam com imagens com diagnósticos confirmados por biópsia, o que torna suas classificações mais confiáveis.

E pelo número de amostras ser limitado, o estudo também não é conclusivo. Alguns modelos como o modelo desenvolvido por Benyes YK, et al. (2022) podem demonstrar uma performance extremamente enviesada, produzindo bons resultados, mas sendo apenas eficiente para um único conjunto de dados específicos. Isso é comum para modelos treinados com poucas imagens e poucas variações.

DISCUSSÃO

Segundo Nasser M e Yusof UK (2023) em sua revisão sistemática para métodos de *Deep Learning* para diagnóstico de câncer de mama com 98 artigos selecionados de 1267 estudos, a CNN é a técnica de Aprendizado Profundo mais utilizada para diagnóstico de câncer de mama. As CNNs podem ser agrupadas em dois tipos: um modelo baseado em transferência de aprendizado e um modelo treinado do zero. Também segundo a revisão, os modelos mais eficazes são aqueles que utilizaram classificação binária ao invés de multiclasse.

Na maioria dos casos, isso se confirma. Ao comparar os modelos desenvolvidos para diagnóstico de câncer de pulmão de Masood A, et al. (2018) e Zhang X, et al. (2023), por exemplo. O primeiro conduziu uma classificação multiclasse para os 4 estágios de câncer do pulmão utilizando um conjunto de dados fornecido pelo hospital de Shanghai. O segundo tentava distinguir apenas dois problemas, um nódulo de tuberculose ou um câncer de pulmão com uma classificação binária, utilizando um conjunto de PET scans também de um hospital chinês. Para a métrica de sensibilidade menos casos verdadeiramente positivos para câncer foram perdidos no segundo modelo.

Mas, às vezes o modelo CNN também pode ser combinado com outras técnicas, não utilizando somente o reconhecimento de imagem para diagnóstico, mas também utilizando o histórico médico do paciente ou de diagnósticos. Essa combinação geralmente resulta em métricas melhores como no modelo de Lin H, et al. (2021), uma CNN treinada com 19 mil imagens com uma classificação adicional em estratificação de risco baseado em regras definidas com diagnósticos clínicos de citologistas.

Existem muitos modelos comerciais de diagnóstico de câncer já em utilização atualmente, muitos deles já atrelados às próprias máquinas de exames de imagem. Riveira-Martin M, et al. (2023) avaliaram um desses modelos comerciais, o Transpara da empresa ScreenPoint Medical para diagnóstico precoce de câncer de mama. Ao analisar determinados parâmetros chegaram à conclusão de que o sistema possui uma capacidade diagnóstica robusta e que sua precisão é independente do equipamento de mamografia utilizado, qualidade do feixe, rodada de triagem, idade da mulher, espessura da mama e compressão aplicada, densidade definida como PDV (*Percent Dense tissue Volume* – Porcentagem do Volume de tecido Denso) e dose glandular.

Nos exemplos selecionados há modelos clássicos como ResNet50, VGG-19 e InceptionV1 que são comumente utilizados na técnica de transferência de aprendizado, modelos que figuraram no top do ranking da ImageNet com as melhores métricas, mas também há novas arquiteturas que foram desenvolvidas posteriormente. Isso se deve ao fato do conhecimento da arquitetura de modelos de redes neurais estar disponível e pela quantidade crescente de artigos publicados, além da evolução da tecnologia e poder de processamento, que se tornam cada vez mais acessíveis.

As métricas de sensibilidade e especificidade selecionadas na Tabela 1 para avaliar a qualidade dos modelos não possuem nenhuma diferença significativa entre elas ao analisá-las através de um teste estatístico. São métricas muito boas, que indicam um erro muito baixo de anos bem recentes. Porém, a média para a sensibilidade dos modelos selecionados foi de 0,859, já a média de especificidade foi de 0,907, ligeiramente mais alta. Isso provavelmente ocorre porque em geral existem mais exemplos de imagens com diagnósticos saudáveis do que imagens com um diagnóstico de câncer, os modelos então tendem a classificar melhor para pessoas saudáveis. Se

uma pessoa está saudável dificilmente o modelo errará ao diagnosticá-la com câncer, falsos positivos são mais raros. Ao mesmo tempo, a sensibilidade mais baixa implica que pacientes com câncer correm o risco de serem diagnosticados incorretamente.

Muitas limitações foram encontradas nos artigos pela falta de imagens de boa qualidade disponíveis para treinamento dos modelos. A técnica de *Data Augmentation* pode ser empregada para criar variações a partir das imagens disponíveis, mesmo assim a melhora que essa técnica pode garantir aos modelos é limitada. No entanto, com o surgimento da nova tecnologia de IA generativa combinada à arquitetura *Transformer*, novas soluções para esse problema estão sendo criadas. Haveria então a possibilidade de gerar mais imagens com exemplos de câncer para treinar um modelo melhor e assim diminuir os falsos negativos. No projeto de Salvia ML, et al. (2022) por exemplo, se propuseram a desenvolver uma rede adversarial generativa convolucional profunda para gerar imagens hiperespectrais sintéticas de lesões epidérmicas, visando o diagnóstico de câncer de pele, e assim superar o desafio de conjuntos de dados insuficientes para treinar modelos. Obtiveram bons resultados iniciais como uma sensibilidade de 0,861.

É um nova tecnologia que ainda está sendo validada, mas já há muitos estudos com resultados promissores. A partir da tecnologia da IA Generativa surge a possibilidade de se criar dados sintéticos que podem não só ajudar em modelos de reconhecimento de imagem para diagnóstico, mas também no desenvolvimento de novos medicamentos, ou até na telemedicina (Boyd BJ, 2019; Cascella M, et al., 2023).

CONCLUSÃO

A aplicação da Inteligência Artificial (IA) no diagnóstico de câncer por meio de exames de imagem tem demonstrado avanços significativos. Com a utilização de redes neurais de Aprendizado Profundo e métricas de sensibilidade e especificidade revelam-se como ferramentas valiosas para aprimorar a precisão diagnóstica. A evolução da tecnologia de IA, especialmente com o advento da arquitetura *Transformer* e a IA generativa, promete superar desafios relacionados à qualidade e quantidade de dados de treinamento, potencializando a criação de modelos mais eficientes e confiáveis. Estudos recentes indicam que sistemas de IA podem não apenas complementar, mas em alguns casos superar o desempenho de radiologistas e médicos, reduzindo a carga de trabalho e melhorando os resultados clínicos. Apesar das limitações atuais, como a necessidade de grandes conjuntos de dados e poder de processamento, a IA está se consolidando como uma tecnologia revolucionária na medicina diagnóstica, com potencial para transformar o diagnóstico e tratamento do câncer, bem como outras áreas da saúde.

REFERÊNCIAS

1. ARBYN M, et al. Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis. Lancet Glob Health, 2020; 8(2):e191-e203.
2. BENYES YK, et al. A Comparative Analysis of Deep Learning Models for Automated Cross-Preparation Diagnosis of Multi-Cell Liquid Pap Smear Images. Diagnostics, 2022; 12, 1838.
3. BOYD BJ, et al. Successful oral delivery of poorly water-soluble drugs both depends on the intraluminal behavior of drugs and of appropriate advanced drug delivery systems. European Journal of Pharmaceutical Sciences, 2019; 137, 104977.
4. CASCELLA M, et al. Utilizing an artificial intelligence framework (conditional generative adversarial network) to enhance telemedicine strategies for cancer pain management. J Anesth Analg Crit Care, 2023;3(1):19.

5. HOLMSTRÖM O, et al. Point-of-Care Digital Cytology with Artificial Intelligence for Cervical Cancer Screening in a Resource-Limited Setting. JAMA Network Open, 2021; 4(3):e211740.
6. HUANG HW, et al. Development of a light-weight deep learning model for cloud applications and remote diagnosis of skin cancers. The Journal of Dermatology, 2021; 48: 310-316.
7. HUSSAIN S, et al. Modern Diagnostic Imaging Technique Applications and Risk Factors in the Medical Field: A Review. Biomed Res Int, 2022; 5164970.
8. JUNIOR GBV, et al. Métricas utilizadas para avaliar a eficiência de classificadores em algoritmos inteligentes. Centro de Pesquisas Avançadas em Qualidade de Vida, 2022; 14, 2(1).
9. KAPLAN A, HAENLEIN M. Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Business Horizons, 2019; 62(1) 15-25.
10. KHOSRAVI P, et al. A Deep Learning Approach to Diagnostic Classification of Prostate Cancer Using Pathology–Radiology Fusion. Journal of Magnetic Resonance Imaging, 2021; 4:462–471.
11. KIM HE, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. Lancet Digit Heal, 2020; 2(3):e138–48.
12. KOOI T, et al. Large scale deep learning for computer aided detection of mammographic lesions. Med Image Anal, 2017; 35, 303–312.
13. KUO KM, et al. The accuracy of artificial intelligence used for non-melanoma skin cancer diagnoses: a meta-analysis. BMC Medical Informatics and Decision Making, 2023; 23:138.
14. LIN H, et al. Dual-path network with synergistic grouping loss and evidence driven risk stratification for whole slide cervical image analysis. Medical Image Analysis, 2021; 69, 101955.
15. LIU F, et al. Research progress of medical image recognition based on deep learning. Chin J Biomed Eng, 2018; 37(1), 86–94.
16. LIU J, et al. Application and research of deep learning in medical image recognition. J Qingdao Univers (Nat Sci Edit), 2018; 31(1), 69-74,80.
17. LIU R, et al. A review of medical artificial intelligence. Global Health Journal, 2020; 4(2).
18. LOTTER W, et al. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. Nat Med, 2021; 27(2):244–9.
19. MASOOD A, et al. Computer-Assisted Decision Support System in Pulmonary Cancer detection and stage classification on CT images. Journal of Biomedics Informatics, 2018; 79, 117-128.
20. MATSUOKA Y, et al. Deep-learning prostate cancer detection and segmentation on biparametric versus multiparametric magnetic resonance imaging: Added value of dynamic contrast-enhanced imaging. International Journal of Urology, 2023; 1103-1111.
21. MCKINNEY SM, et al. International evaluation of an AI system for breast cancer screening. Nature, 2020; 577(7788):89-94.
22. MISHKIN D, et al. Systematic evaluation of convolution neural networks advances on the Imagenet. Computer Vision and Image Understanding, 2017; 161, 11-19.
23. NASSER M, YUSOF UK. Deep Learning Based Methods for Breast Cancer Diagnosis: A Systematic Review and Future Direction. Diagnostics, 2023; 13,161.
24. PUPPE F. Introduction to knowledge systems. Artificial Intelligence in Medicine, 1997; 9(2), 201-203.
25. RIVEIRA-MARTIN M, et al. Multi-vendor robustness analysis of a commercial artificial intelligence system for breast cancer detection. Journal of Medical Imaging, 2023; 10(5): 051807.
26. SALIM M, et al. External Evaluation of 3 Commercial Artificial Intelligence Algorithms for Independent Assessment of Screening Mammograms. JAMA Oncol, 2020; 6(10):1581–8.
27. SANTOS MO, et al. Estimativa de Incidência de Câncer no Brasil, 2023-2025. Revista Brasileira de Cancerologia, 2023; 69(1):e-213700.
28. SORNAPUDI S, et al. Comparing Deep Learning Models for Multi-cell Classification in Liquid-based Cervical Cytology Image. AMIA Annu Symp Proc, 2019; 820–827.

29. VASWANI A, et al. Attention Is All You Need. Conference on Neural Information Processing Systems. NIPS, 2017.
30. ZHANG X, et al. Deep learning PET/CT-based radiomics integrates clinical data: A feasibility study to distinguish between tuberculosis nodules and lung cancer. Thoracic Cancer, 2023; 1802-1811.

Por fim, o artigo intitulado “Decifrando modelos de classificação de imagens médicas: Explicabilidade como pilar para diagnósticos confiáveis” também foi submetido, refletindo diretamente os achados obtidos ao longo de todas as etapas descritas no trabalho. O artigo aborda os principais resultados alcançados, com ênfase na importância da explicabilidade dos modelos de inteligência artificial no contexto clínico. Essa publicação não apenas valida as conclusões apresentadas na dissertação, mas também amplia o impacto dos resultados, oferecendo diretrizes para futuros estudos e aplicações na área de diagnóstico por imagem. Além disso, reforça a relevância dos esforços empreendidos ao conectar a metodologia adotada à necessidade de modelos confiáveis, robustos e interpretáveis, contribuindo para o avanço do conhecimento em inteligência artificial aplicada à saúde. Submetido para a **Revista Eletrônica de Comunicação Informação & Inovação em Saúde**, ISSN 1981-6278, classificada como Qualis A3 no período de 2017-2020, disponível em: https://www.reciis.icict.fiocruz.br/index.php/reciis/index.

ANEXO A - Registro de software

REPÚBLICA FEDERATIVA DO BRASIL
MINISTÉRIO DO DESENVOLVIMENTO, INDÚSTRIA, COMÉRCIO E SERVIÇOS
INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL
DIRETORIA DE PATENTES, PROGRAMAS DE COMPUTADOR E TOPOGRAFIAS DE CIRCUITOS

Certificado de Registro de Programa de Computador

Processo Nº: **BR512025000290-4**

O Instituto Nacional da Propriedade Industrial expede o presente certificado de registro de programa de computador, válido por 50 anos a partir de 1° de janeiro subsequente à data de 27/01/2025, em conformidade com o §2°, art. 2° da Lei 9.609, de 19 de Fevereiro de 1998.

Título: Modelo de Fusão para classificação de imagens de citologia cervical

Data de publicação: 27/01/2025

Data de criação: 01/12/2024

Titular(es): SILVIA CRISTINA MARTINI; LETICIA BRUM CROFFI

Autor(es): LETICIA BRUM CROFFI

Linguagem: PYTHON

Campo de aplicação: IF-10

Tipo de programa: TC-04

Algoritmo hash: SHA-256

Resumo digital hash: a69a589b7cef68216b319109678fa92875b4208aaca3fe32a1ca2e5ebf4ea892

Expedido em: 04/02/2025

Aprovado por:
Carlos Alexandre Fernandes Silva
Chefe da DIPTO