

Universidade Federal do Piauí
Centro de Ciências da Natureza
Programa de Pós-Graduação em Ciência da Computação

Classificação de Células Cervicais Cancerígenas com CNNs e ViTs em Ensemble para Auxílio ao Diagnóstico Médico

Marcelo Victor Sousa Lima

Teresina - PI, Setembro de 2025

Marcelo Victor Sousa Lima

Classificação de Células Cervicais Cancerígenas com CNNs e ViTs em Ensemble para Auxílio ao Diagnóstico Médico

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFPI (área de concentração: Computação Visual), como parte dos requisitos necessários para a obtenção do Título de Mestre em Ciência da Computação.

Universidade Federal do Piauí – UFPI

Centro de Ciências da Natureza

Programa de Pós-Graduação em Ciência da Computação

Orientador: Laurindo de Sousa Britto Neto

Teresina - PI

Setembro de 2025

FICHA CATALOGRÁFICA
Universidade Federal do Piauí
Biblioteca Comunitária Jornalista Calos Castello Branco
Serviço de Processos Técnicos

L732c Lima, Marcelo Victor Sousa.
Classificação de Células Cervicais Cancerígenas com CNNs e ViTs em Ensemble para Auxílio ao Diagnóstico Médico / Marcelo Victor Sousa Lima. – 2025.
86 f. : il. color.

Dissertação (Mestrado) – Universidade Federal do Piauí, Centro de Ciências da Natureza, Programa de Pós-Graduação em Ciência da Computação, Teresina, 2025.
"Orientador: Prof. Laurindo de Sousa Britto Neto."

1. Câncer do colo do útero. 2. Classificação de células cervicais. 3. Fusão de características. 4. Fusão de características. 5. Visão computacional. I. Britto Neto, Laurindo de Sousa. II. Título.

CDD 616.994 66

Bibliotecário: Gésio dos Santos Barros - CRB3/1469

Marcelo Victor Sousa Lima

Classificação de Células Cervicais Cancerígenas com CNNs e ViTs em Ensemble para Auxílio ao Diagnóstico Médico

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFPI (área de concentração: Computação Visual), como parte dos requisitos necessários para a obtenção do Título de Mestre em Ciência da Computação.

Trabalho aprovado. Teresina - PI, 05 de setembro de 2025:

Documento assinado digitalmente
gov.br
LAURINDO DE SOUSA BRITTO NETO
Data: 23/10/2025 17:43:23-0300
Verifique em https://validar.iti.gov.br

Laurindo de Sousa Britto Neto
Orientador

Documento assinado digitalmente
gov.br
RODRIGO DE MELO SOUZA VERAS
Data: 22/10/2025 10:45:35-0300
Verifique em https://validar.iti.gov.br

Professor
Rodrigo de Melo Souza Veras

Documento assinado digitalmente
gov.br
VINICIUS PONTE MACHADO
Data: 22/10/2025 10:36:49-0300
Verifique em https://validar.iti.gov.br

Professor
Vinicius Ponte Machado

Documento assinado digitalmente
gov.br
ANDREA GOMES CAMPOS
Data: 22/10/2025 10:31:23-0300
Verifique em https://validar.iti.gov.br

Professora
Andrea Gomes Campos Bianchi

Teresina - PI
Setembro de 2025

Agradeço profundamente à minha esposa Maria das Graças, cuja paciência, apoio incondicional e amor constante foram fundamentais para que eu seguisse em frente, mesmo nos momentos mais desafiadores.

À minha filha Maria Valentia, dedico este trabalho, pois acredito que o estudo é a chave que liberta, e desejo que ela sempre tenha a coragem de buscar seu conhecimento, de construir seu caminho e de conquistar seus sonhos

Agradecimentos

Agradeço à Deus.

Agradeço à minha família, por todo o apoio e amor.

Agradeço ao meu orientador, Prof. Dr. Laurindo de Sousa Britto Neto, por todos os conselhos, pela paciência e ajuda nesse período.

À Fundação de Amparo à Pesquisa do Piauí (FAPEPI) pelo apoio financeiro para realização deste trabalho de pesquisa.

"Estude muito o que mais lhe interessa da maneira mais indisciplinada, irreverente e original possível."

(Richard Feynman)

Resumo

O câncer do colo do útero permanece um grande desafio global de saúde pública, especialmente em países de baixa e média renda, nos quais triagens ineficazes impactam índices de morbidade e mortalidade. Esta dissertação investiga abordagens computacionais inovadoras para a classificação automática de células cervicais, explorando o aprendizado em ensemble que combina arquiteturas de redes neurais profundas de última geração: EfficientViT, EVA-02 e EdgeNeXt. Foram conduzidos experimentos extensivos utilizando dois conjuntos de dados públicos, Herlev (917 imagens) e SIPaKMeD (4.049 imagens), em cenários de classificação binária e multiclasse. A metodologia proposta utiliza técnicas de aumento de dados, aprendizado por transferência e procedimentos robustos de validação cruzada com 5 folds, além de validação holdout para comparação com a literatura. A análise estatística, utilizando o teste de *Wilcoxon*, confirma a superioridade e relevância estatística do ensemble em relação aos modelos individuais. O ensemble atingiu desempenho estado da arte, com acurácias de 98,35% (binário) e 83,40% (7 classes) no Herlev, e 99,73% (binário), 98,96% (3 classes) e 98,08% (5 classes) no SIPaKMeD. Estes resultados evidenciam a eficácia da fusão de modelos com vieses indutivos distintos e destacam o potencial de redução do esforço do especialista e da análise manual de células na prática clínica. Em suma, este trabalho contribui metodologicamente com uma abordagem de ensemble inovadora que integra arquiteturas convolutionais e baseadas em ViT, ao fornecer benchmarks validados e resultados em dados públicos, abrindo caminho para soluções de triagem apoiadas por IA mais robustas e acessíveis na prevenção do câncer cervical.

Palavras-chaves: Câncer do colo do útero. Classificação de células cervicais. Fusão de características. Visão computacional.

Abstract

Cervical cancer remains a major global public health challenge, particularly in low- and middle-income countries, where ineffective screening impacts morbidity and mortality rates. This dissertation investigates novel computational approaches for the automatic classification of cervical cells by leveraging ensemble learning that combines state-of-the-art deep neural network architectures: EfficientViT, EVA-02, and EdgeNeXt. Extensive experiments were conducted using two public datasets, Herlev (917 images) and SIPaKMeD (4,049 images), across binary and multiclass classification scenarios. The proposed methodology utilizes data augmentation techniques, transfer learning, and robust 5-fold cross-validation procedures, along with holdout validation for literature comparison. Statistical analysis employing the Wilcoxon signed-rank test confirms the superiority and statistical relevance of the ensemble over individual models. The ensemble achieved state-of-the-art performance, with accuracies of 98.35% (binary) and 83.40% (7 classes) on Herlev, and 99.73% (binary), 98.96% (3 classes), and 98.08% (5 classes) on SIPaKMeD. These results demonstrate the efficacy of model fusion with distinct inductive biases and highlight the potential for reducing expert workload and manual cell analysis in clinical practice. In summary, this work contributes both methodologically with an innovative ensemble approach that integrates convolutional and transformer based architectures, by providing validated benchmarks and public dataset results, paving the way for more robust and accessible AI-based screening solutions in cervical cancer prevention. **Keywords**: Cervical cancer. Cervical cell classification. Feature fusion. Computer vision.

Lista de ilustrações

Figura 1 – Ilustração do exame de citologia utilizando a técnica do Papanicolau convencional. 2
Figura 2 – Quantidades de trabalhos obtidos por base bibliográfica. 8
Figura 3 – Quantidade de trabalhos aceitos, rejeitados e duplicados na fase de seleção. 9
Figura 4 – Quantidade de trabalhos aceitos, rejeitados e duplicados na fase de extração. 10
Figura 5 – Células normais e anormais, as três primeiras figuras representam as células normais enquanto as demais são as células anormais seguindo a Tabela 6. Fonte: (JANTZEN et al., 2005) 30
Figura 6 – Células do conjunto SIPaKMeD. As figuras 'a' e 'b' representam células normais. Enquanto as figuras 'c', 'd' e 'e' representam células anormais. Fonte: (PLISSITI et al., 2018) 31
Figura 7 – Representação de um neurônio artificial (Perceptron), em que $(x_1, x_2, ..., x_n)$ são os sinais de entrada do neurônio, $(w_{k1}, w_{k2}, ..., w_{kn})$ são os pesos associados aos sinais de entrada, b_k é o viés do neurônio. $\phi(\cdot)$ é a função de ativação que processa os sinais e y_k é o sinal de saída. Fonte: (SIMON, 2009) 32
Figura 8 – Representação de um MLP. Fonte: (SIMON, 2009) 33
Figura 9 – Representação da arquitetura LeNet5 aplicada ao caractere A manuscrito. Fonte: (LECUN et al., 1998) 35
Figura 10 – A representação da arquitetura *Transformer*. Fonte: (VASWANI et al., 2017) 36
Figura 11 – A representação da arquitetura *Vision Transformer*. Fonte: (DOSOVITSKIY et al., 2021) 38
Figura 12 – A representação da arquitetura *EfficientVit*. Fonte: (CAI et al., 2022) . 39
Figura 13 – A representação da arquitetura ViT tradicional (a) e, a representação da *Transform Vision* (TrV) (b). TrV é construído a partir do ViT, adicionando melhorias na arquitetura. Fonte: (FANG et al., 2024c) . . 41
Figura 14 – Arquitetura do *EdgeNeXt*. Fonte: (MAAZ et al., 2022) 43
Figura 15 – Fluxo dos experimentos realizados. Fonte: Elaborado pelo autor. 49
Figura 16 – Matriz de confusão da abordagem proposta para o conjunto Herlev classificado de modo binário (células anormais vs normais), o conjunto de teste foi composto de 92 imagens. Fonte: Elaboração do autor. . . . 54

Figura 17 – Matriz de confusão do modelo EfficientViT para o conjunto Herlev classificado de modo binário (células anormais vs normais), o conjunto de teste foi composto de 92 imagens. Fonte: Elaboração do autor. . . . 55
Figura 18 – *Loss* de treinamento e validação do modelo EfficientViT para o conjunto Herlev contendo duas classes. Fonte: Elaborado pelo autor. 55
Figura 19 – Matriz de confusão do modelo EVA para o conjunto Herlev classificado de modo binário (células anormais vs normais), o conjunto de teste foi composto de 92 imagens. Fonte: Elaboração do autor. 55
Figura 20 – *Loss* de treinamento e validação do modelo EVA para o conjunto Herlev contendo duas classes. Fonte: Elaborado pelo autor. 56
Figura 21 – Matriz de confusão do modelo EdgeNeXt para o conjunto Herlev classificado de modo binário (células anormais vs normais), o conjunto de teste foi composto de 92 imagens. Fonte: Elaboração do autor. 56
Figura 22 – *Loss* de treinamento e validação do modelo EdgeNeXt para o conjunto Herlev contendo duas classes. Fonte: Elaborado pelo autor. 56
Figura 23 – Matriz de confusão da abordagem proposta para o conjunto Herlev classificado com as sete classes, o conjunto de teste foi composto de 92 imagens. Fonte: Elaboração do autor. 58
Figura 24 – Matriz de confusão do modelo EfficientViT para o conjunto Herlev classificado com as sete classes o conjunto de teste foi composto de 92 imagens. Fonte: Elaboração do autor. 58
Figura 25 – *Loss* de treinamento e validação do modelo EfficientViT para o conjunto Herlev contendo sete classes. Fonte: Elaborado pelo autor. 59
Figura 26 – Matriz de confusão do modelo EVA para o conjunto Herlev classificado de modo binário com as sete classes, o conjunto de teste foi composto de 92 imagens. Fonte: Elaboração do autor. 59
Figura 27 – *Loss* de treinamento e validação do modelo EVA para o conjunto Herlev contendo sete classes. Fonte: Elaborado pelo autor. 60
Figura 28 – Matriz de confusão do modelo EdgeNeXt para o conjunto Herlev classificado de modo binário com as sete classes, o conjunto de teste foi composto de 92 imagens. Fonte: Elaboração do autor. 60
Figura 29 – *Loss* de treinamento e validação do modelo EdgeNeXt para o conjunto Herlev contendo sete classes. Fonte: Elaborado pelo autor. 61
Figura 30 – Matriz de confusão da abordagem proposta para o conjunto *SIPaKMeD* classificado com as duas classes, o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor. 63
Figura 31 – Matriz de confusão do modelo EfficientViT para o conjunto *SIPaKMeD* classificado com as duas classes o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor. 63

Figura 32 – *Loss* de treinamento e validação do modelo EfficientViT para o conjunto *SIPaKMeD* contendo duas classes. Fonte: Elaborado pelo autor. 64
Figura 33 – Matriz de confusão do modelo EVA para o conjunto *SIPaKMeD* classificado com as duas classes, o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor. 64
Figura 34 – *Loss* de treinamento e validação do modelo EVA para o conjunto *SIPaKMeD* classificado com as duas classes, o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor. 65
Figura 35 – Matriz de confusão do modelo EdgeNeXt para o conjunto *SIPaKMeD* classificado com as duas classes, o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor. 65
Figura 36 – *Loss* de treinamento e validação do modelo EdgeNeXtpara o conjunto *SIPaKMeD* classificado com as duas classes, o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor. 66
Figura 37 – Matriz de confusão da abordagem proposta para o conjunto *SIPaKMeD* classificado com as três classes, o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor. 67
Figura 38 – Matriz de confusão do modelo EfficientViT para o conjunto *SIPaKMeD* classificado com as três classes o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor. 67
Figura 39 – *Loss* de treinamento e validação do modelo EfficientViT para o conjunto *SIPaKMeD* contendo três classes. Fonte: Elaborado pelo autor. 68
Figura 40 – Matriz de confusão do modelo EVA para o conjunto *SIPaKMeD* classificado com as três classes, o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor. 68
Figura 41 – *Loss* de treinamento e validação do modelo EVA para o conjunto *SIPaKMeD* classificado com as três classes, o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor. 69
Figura 42 – Matriz de confusão do modelo EdgeNeXt para o conjunto *SIPaKMeD* classificado com as três classes, o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor. 69
Figura 43 – *Loss* de treinamento e validação do modelo EdgeNeXtpara o conjunto *SIPaKMeD* classificado com as três classes, o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor. 70
Figura 44 – Matriz de confusão da abordagem proposta para o conjunto *SIPaKMeD* classificado com as cinco classes, o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor. 71

Figura 45 – Matriz de confusão do modelo EfficientViT para o conjunto *SIPaKMeD* classificado com as cinco classes o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor. 71
Figura 46 – *Loss* de treinamento e validação do modelo EfficientViT para o conjunto *SIPaKMeD* contendo cinco classes. Fonte: Elaborado pelo autor. 72
Figura 47 – Matriz de confusão do modelo EVA para o conjunto *SIPaKMeD* classificado com as cinco classes, o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor. 72
Figura 48 – *Loss* de treinamento e validação do modelo EVA para o conjunto *SIPaKMeD* classificado com as três classes, o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor. 73
Figura 49 – Matriz de confusão do modelo EdgeNeXt para o conjunto *SIPaKMeD* classificado com as cinco classes, o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor. 73
Figura 50 – *Loss* de treinamento e validação do modelo EdgeNeXtpara o conjunto *SIPaKMeD* classificado com as cinco classes, o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor. 74

Lista de tabelas

Tabela 1 – Distribuição de publicações por fonte 7
Tabela 2 – Quantidade de trabalhos coletados para revisão sistemática de acordo com sua categoria proposta. 12
Tabela 3 – Sumarização da quantidade de modelos utilizadas na literatura e os conjuntos de dados utilizados por eles. 13
Tabela 4 – Análise quantitativa dos conjuntos de imagens utilizadas pelos trabalhos. 23
Tabela 5 – Sumarização dos resultados de acurácia obtidos pelos artigos da revisão sistemática. Os resultados que não especificam o número de classes referem-se a casos onde essa informação não foi disponibilizada no artigo original. 24
Tabela 5 – Sumarização dos resultados de acurácia obtidos pelos artigos da revisão sistemática. Os resultados que não especificam o número de classes referem-se a casos onde essa informação não foi disponibilizada no artigo original. 25
Tabela 5 – Sumarização dos resultados de acurácia obtidos pelos artigos da revisão sistemática. Os resultados que não especificam o número de classes referem-se a casos onde essa informação não foi disponibilizada no artigo original. 26
Tabela 5 – Sumarização dos resultados de acurácia obtidos pelos artigos da revisão sistemática. Os resultados que não especificam o número de classes referem-se a casos onde essa informação não foi disponibilizada no artigo original. 27
Tabela 5 – Sumarização dos resultados de acurácia obtidos pelos artigos da revisão sistemática. Os resultados que não especificam o número de classes referem-se a casos onde essa informação não foi disponibilizada no artigo original. 28
Tabela 6 – Classificação do conjunto de dados *Herlev* com 917 células (242 normais e 675 anormais). 29
Tabela 7 – Conjunto de imagens SIPaKMeD com suas respectivas categorias. . . . 30
Tabela 8 – Número de imagens geradas no aumento de dados para cada cenário. . 50
Tabela 9 – Hiperparâmetros utilizados para o treinamento e parâmetros dos modelos utilizados para o treinamento. 51
Tabela 10 – Desempenho no conjunto *Herlev* com 2 classes para validação cruzada. 53
Tabela 11 – Desempenho no conjunto *Herlev* com 7 classes para validação cruzada.. 53
Tabela 12 – Desempenho dos modelos no conjunto Herlev com 2 classes utilizando a validação holdout. 54

Tabela 13 – Desempenho dos modelos no conjunto Herlev com 7 classes utilizando a validação holdout. 57
Tabela 14 – Desempenho no conjunto *SIPaKMeD* para 2 classes com validação cruzada. 62
Tabela 15 – Desempenho no conjunto *SIPaKMeD* para 3 classes com validação cruzada. 62
Tabela 16 – Desempenho no conjunto *SIPaKMeD* para 5 classes com validação cruzada. 62
Tabela 17 – Desempenho no conjunto *SIPaKMeD* para 2 classes utilizando holdout. 62
Tabela 18 – Desempenho no conjunto *SIPaKMeD* para 3 classes utilizando holdout. 66
Tabela 19 – Desempenho no conjunto SIPaKMeD para 5 classes utilizando holdout. 68
Tabela 20 – Comparação da abordagem proposta com os principais estudos da literatura em cada cenário do experimento realizado. As siglas represetam: acurácia (acc), precisão (prec), conjunto de dados (conjunto), divisão (divisão de dados), validação cruzada com *k-fold (k-fold)*. 77
Tabela 21 – P-valores obtidos pelo WSR, com nível de confiança de 99%, comparando a abordagem proposta com os melhores modelos em cada cenário de experimentos da validação cruzada. 77

Lista de abreviaturas e siglas

Acc	Acurácia
AICCS	*Artificial Intelligence Cervical Cancer Screening*
AUC	*Area Under the Curve*
BCNN	*Bilinear Convolutional Neural Network*
BERT	*Bidirectional Encoder Representations from Transformers*
CACCD-GOADL	*Computer Aided Cervical Cancer Diagnosis using Gazelle Optimizer Algorithm with Deep Learning*
CAD	*Computer-Aided Diagnosis*
CG	*Classifier Generalization*
c-MLP	*cascaded Multilayer Perceptron*
CLAHE	*Contrast Limited Adaptive Histogram Equalization*
CLIP	*Contrastive Language-Image Pre-Training*
CNN	*Convolutional Neural Network*
DDPM	*Denoising Diffusion Probabilistic Model*
DIFF	*Deep Integrated Feature Fusion*
DL	*Deep Learning*
DR	*Data Removal*
DS	*Data Synthesis*
DSA	*Dual-Stream Self-Attention*
DT	*Decision Tree*
DW	*Data Weighting*
ELM	*Extreme Learning Machine*
FC	*Fully-Connected*
FDA	*Food and Drug Administration*

FFN	*Feed-Forward Network*
FL	*Federated Learning*
FLOPs	*Floating Point Operations Per Second*
FPR	Taxa de Falsos Positivos
GLU	*Gated Linear Unit*
GRN	*Global Response Normalization*
GRU	*Gated Recurrent Unit*
HDFF	*Hybrid Deep Feature Fusion*
HEM	*Hard Example Mining*
HPV	*Human Papillomavirus*
IATS	*Integrated AttentionTransformer System*
IID	*Independent and Identically Distributed*
IoT	*Internet over Things*
IS	*Importance Sampling*
JSO	*Jellyfish Search Optimizer*
KNN	*K-Nearest Neighbors*
LBC	*Liquid Based Cytology*
LBP	*Local Binary Pattern*
LDA	*Linear Discriminant Analysis*
LSTM	*Long Short-Term Memory*
MIM	*Masked Image Modeling*
ML	*Machine Learning*
MLP	*Multilayer Perceptrons*
MSA	*Multi-head Self-Attention*
NB	*Naive Bayes*
NLP	*Natural Language Processing*

OGL	*Origin Grouping Loss*
OMS	Organização Mundial da Saúde
PAN	*Path Aggregation Network*
PCA	*Principal Component Analysis*
Prec	Precisão
PSO	*Particle Swarm Optimization*
RBF	*Radial Basis Function*
ReLU	*Rectified Linear Unit*
RF	*Random Forest*
RNA	Rede Neural Artificial
ROC	*Receiver Operating Characteristic*
ROPE	*Rotary Position Embedding*
RP	*Random Projection*
SDTA	*Split Depth-wise Transpose Attention*
SELM	*Stacked Extreme Learning Machine*
SiLU	*Sigmoid Linear Unit*
SMOTE	*Synthetic Minority Over-sampling Technique*
SPPF	*Spatial Pyramid Pooling Fusion*
SVC	*Support Vector Classifier*
SVM	*Support Vector Machines*
T2T	*Token-to-Token*
TPR	Taxa de Verdadeiros Positivos
TrV	*Transform Vision*
USA	*United States of America*
VGG	*Visual Geometry Group*
ViT	*Vision Transformer*

WSI	*Whole Slide Analysis*
XAI	*Explainable Artificial Intelligence*
YOLO	*You Only Look Once*

Sumário

1 INTRODUÇÃO 1
1.1 Objetivo 3
1.2 Contribuições Científicas 3
1.3 Organização da Dissertação 4

2 REVISÃO DA LITERATURA 5
2.1 Metodologia de Revisão da Literatura 5
2.1.1 Definição das Questões de Pesquisa 6
2.1.2 Definição das Palavras-Chave 6
2.1.3 Processo de Busca 7
2.1.4 Triagem dos Artigos 7
2.1.5 Processo de Seleção dos Estudos 8
2.1.6 Extração de Dados e Sumarização 9
2.2 Resultados da Revisão da Literatura 10
2.2.1 Abordagens CNNs 13
2.2.2 Abordagem ViT 14
2.2.3 Abordagens Híbridas 15
2.2.4 Abordagens *Ensemble* 21
2.2.5 Abordagens Alternativas 22
2.3 Discussão da Revisão de Literatura 23
2.3.1 Conjuntos de Dados 23
2.3.2 Sumarização dos Resultados 24

3 MATERIAIS E MÉTODOS 29
3.1 Conjuntos de Dados 29
3.1.1 Herlev 29
3.1.2 SIPaKMeD 30
3.2 Classificadores 31
3.2.1 Redes Neurais Artificias 32
3.2.2 Redes Neurais Convolucionais 33
3.2.3 Arquitetura *Transformer* 35
3.2.4 *Vision Transformer* 37
3.3 Modelos utilizados no *ensemble* 39
3.3.1 *EfficientViT* 39
3.3.2 *EVA-02* 40
3.3.3 *EdgeNeXt* 42

3.4 Métricas para avaliação . . . 44
3.5 Teste de Wilcoxon para Amostras Pareadas . . . 46
3.5.1 Objetivo e Hipóteses . . . 46
3.5.2 Condições de Aplicação . . . 46
3.5.3 Procedimento do Teste . . . 46
3.5.4 Exemplo Ilustrativo . . . 47
3.5.5 Aplicação em Classificadores de Imagens Citológicas . . . 47

4 ABORDAGEM PROPOSTA . . . 48
4.1 Etapas dos Experimentos Realizados . . . 49
4.2 Aumento de Dados . . . 50
4.3 Experimentos e Hiperparâmetros dos Modelos . . . 50

5 RESULTADOS E DISCUSSÃO . . . 52
5.1 Resultados no Conjunto de Dados *Herlev* . . . 52
5.1.1 *Herlev* - 2 classes . . . 53
5.1.2 *Herlev* - 7 classes . . . 57
5.2 Resultados no Conjunto de Dados *SIPaKMeD* . . . 61
5.2.1 *SIPaKMeD* - 2 classes . . . 62
5.2.2 *SIPaKMeD* - 3 classes . . . 66
5.2.3 *SIPaKMeD* - 5 classes . . . 66
5.3 Discussão dos Resultados . . . 70
5.3.1 Desempenho e Robustez da Abordagem Proposta . . . 70
5.3.2 Análise Específica: Conjunto *Herlev* . . . 74
5.3.3 Análise Específica: Conjunto *SIPaKMeD* . . . 75
5.3.4 Comparação Entre os Conjuntos e Implicações Clínicas . . . 75
5.3.5 Comparativo com a Abordagem Proposta e os Resultados da Literatura . . . 76
5.3.6 Validação Estatística e Considerações Finais . . . 76

6 CONCLUSÕES E TRABALHOS FUTUROS . . . 78
6.1 Conclusões e Trabalhos Futuros . . . 78
6.2 Avaliação do Cumprimento dos Objetivos Propostos . . . 79

REFERÊNCIAS . . . 80

1 Introdução

O câncer cervical figura como o quarto tipo de câncer mais prevalente entre mulheres (FERLAY et al., 2024a). De acordo com a Organização Mundial da Saúde (OMS), em 2022, aproximadamente 660 mil novos casos foram diagnosticados globalmente, resultando em cerca de 350 mil óbitos no mesmo período (FERLAY et al., 2024a). Desses falecimentos, 94% ocorreram em países de baixa e média renda, evidenciando que, nessas regiões, a baixa qualidade dos sistemas de saúde e a ausência de programas de rastreamento impactam negativamente o tratamento da doença (JIANG et al., 2023). Para o Brasil, a OMS estima que em 2022 existiram 18.700 casos e há uma previsão de aumento de mil novos casos para 2025 (FERLAY et al., 2024b). Essa patologia está fortemente associada ao Papilomavírus Humano (*Human Papillomavirus* – HPV) (BEDELL et al., 2020), especialmente aos tipos de vírus de alto risco, como o HPV-16 e o HPV-18, que provocam alterações significativas nos núcleos das células cervicais.

O rastreamento por citologia cervical constitui o método mais amplamente empregado para a detecção de lesões pré-neoplásicas do colo do útero. Nesse contexto, distinguem-se duas modalidades principais: o esfregaço convencional de Papanicolaou e a citologia em base líquida (Liquid-Based Cytology – LBC) (OLIVEIRA et al., 2020). No método convencional, as células da zona de transformação são coletadas com espátula e escovinha e imediatamente fixadas em lâmina. Após a coloração pelo método de Papanicolaou, a análise microscópica é inicialmente realizada por um citotecnologista e, nos casos suspeitos, revisada por um patologista. O diagnóstico final baseia-se na classificação das células em normais, anormais ou categorias intermediárias, utilizando-se critérios padronizados de morfologia celular. A Figura 1 ilustra esse processo.

A identificação de alterações nas células cervicais é um fator crucial para prevenir o desenvolvimento de células anormais em tumores malignos. O período necessário para que a fase pré-cancerosa evolua pode variar de 5 a 20 anos (JIANG et al., 2023). Assim, um sistema eficaz para a identificação das fases iniciais dessa patologia poderia direcionar intervenções terapêuticas específicas, visando reduzir a taxa de mortalidade.

Um dos desafios da triagem convencional é que, como a análise é feita de forma manual pelos especialistas, então o processo é dispendioso e lento. Uma das primeiras propostas para a análise automatizada de células anormais foi o *Cytoanalyzer*, desenvolvido em 1956 (TOLLES; BOSTROM, 1956). No entanto, devido às diversas limitações tecnológicas da época e à ocorrência frequente de erros, essa tecnologia não obteve êxito ao substituir o processo manual de análise das células cervicais. Desde então, diversos sistemas automatizados foram desenvolvidos com o objetivo de alcançar essa meta. Um

sistema de última geração, aprovado pela agência reguladora norte-americana *Food and Drug Administration* (FDA), é o *Genius Digital Diagnostics System* (IKENBERG et al., 2023).

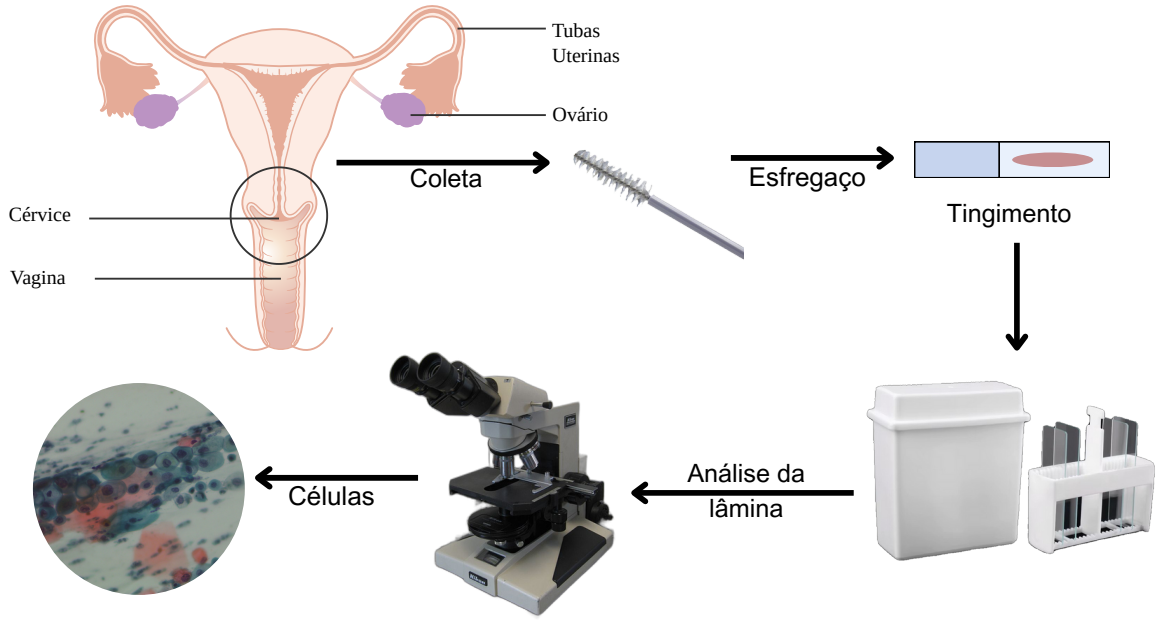


Figura 1 – Ilustração do exame de citologia utilizando a técnica do Papanicolau convencional.

Atualmente, com o avanço da inteligência artificial (IA) e o crescente poder computacional para processamento de dados, são utilizadas abordagens baseadas em visão computacional e aprendizado de máquina para o reconhecimento e classificação de células cervicais (FANG et al., 2024b). As redes neurais convolucionais (*Convolutional Neural Network* – CNN) são uma categoria de redes neurais profundas (*Deep Learning* – DL) que se destacaram por sua capacidade de reconhecer e classificar objetos, pessoas, animais, extraindo características de modo automático (KRIZHEVSKY; SUTSKEVER; HINTON, 2012) utilizando a convolução (filtros) e o *pooling* para extração de características das imagens globais.

Uma abordagem alternativa e mais recente, que se originou da área de Processamento de Linguagem Natural, é a arquitetura Transformer (VASWANI et al., 2017), que se utiliza do mecanismo de atenção. Os *Vision Transformers* (ViTs) foram introduzidos por pesquisadores do Google (DOSOVITSKIY et al., 2021) como uma adaptação dos Transformers para o contexto de imagens, demonstrando um desempenho superior ao das Redes Neurais Convolucionais (CNNs) em grandes conjuntos de dados. De acordo com (DEININGER et al., 2022), os ViTs superaram as CNNs em diversas tarefas de classificação por serem livres de vieses induzidos pela convolução, o que permite ao modelo capturar características globais e relações complexas nos dados. Consequentemente, técnicas basea-

das em ViTs e CNNs têm sido empregadas na área da saúde (YADAV; JADHAV, 2019) para tarefas de classificação, segmentação e detecção de imagens.

1.1 Objetivo

O presente estudo tem como objetivo principal desenvolver uma abordagem em visão computacional voltada à classificação de células cervicais, buscando alcançar resultados superiores ou comparáveis aos métodos mais avançados descritos na literatura, contribuindo, assim, para o aprimoramento de sistemas baseados em inteligência artificial voltados ao rastreamento automatizado do câncer. Para atingir esse objetivo, foram estabelecidas metas específicas, que serão detalhadas a seguir.

Objetivos Específicos

Os objetivos específicos deste estudo são os seguintes:

1. Analisar o estado da arte em classificação de células cervicais, com foco em abordagens de aprendizado profundo (CNNs e ViTs), para identificar desafios persistentes como a diferenciação de subtipos celulares e o desbalanceamento de dados.

2. Realizar o tratamento e a preparação dos conjuntos de dados públicos Herlev e SIPaKMeD , aplicando um protocolo de aumento de dados para mitigar o *overfitting* e o desbalanceamento entre as classes em cenários de classificação binária e multiclasse.

3. Conduzir uma validação experimental sistemática da abordagem proposta, utilizando um protocolo de validação cruzada pela sua robustez e, para vias comparativas com a literatura a validação por *Holdout*.

4. Executar uma análise comparativa do desempenho do modelo que será proposto em relação a outros trabalhos de referência do estado da arte. Adicionalmente, será utilizada a análise estatística com o teste *Wilcoxon Signed-Rank* para uma comparação mais acertiva em relação aos resultados que serão obtidos.

1.2 Contribuições Científicas

As contribuições científicas desta dissertação podem ser sumarizadas em dois eixos principais:

- **Contribuição Metodológica:** O desenvolvimento de uma nova abordagem de *ensemble learning* para a classificação de células cervicais, fundamentada na combinação de três arquiteturas de última geração: EfficientViT, EVA-02 e EdgeNeXt. A

contribuição reside na fusão de modelos com vieses indutivos distintos, aliando a eficiência de redes com base convolucional à capacidade de modelagem de contexto global dos ViTs.

- **Contribuição Empírica:** A obtenção de resultados de estado da arte, com validação estatística, em diversos cenários de classificação nos conjuntos de dados públicos Herlev e SIPaKMeD. Obtendo-se o estado da arte para o conjunto de dados SIPaKMeD para as métricas de acurácia, precisão, *recall* e f1.

O principal fruto desta dissertação foi o artigo publicado no Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS):

- LIMA, Marcelo Victor; BRITTO, Maria Helena Mesquita; BRITTO NETO, Laurindo. Classificação de Células Cervicais Cancerígenas com CNNs e ViTs em Ensemble para Auxílio ao Diagnóstico Médico. In: SIMPÓSIO BRASILEIRO DE COMPUTAÇÃO APLICADA À SAÚDE (SBCAS), 25. 2025, Porto Alegre/RS. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2025 . p. 497-508. ISSN 2763-8952. DOI: <https://doi.org/10.5753/sbcas.2025.7467>. (LIMA; BRITTO; NETO, 2025)

1.3 Organização da Dissertação

Esta dissertação está estruturada em seis capítulos. O Capítulo 1 apresenta uma introdução ao tema, contextualizando sua relevância e os desafios na análise automatizada de citologia cervical. O Capítulo 2 realiza uma revisão sistemática da literatura, identificando as lacunas que motivam este estudo, especialmente no que tange à combinação de diferentes arquiteturas de aprendizado profundo.

O Capítulo 3 detalha os materiais e métodos, incluindo a descrição dos conjuntos de dados públicos Herlev e SIPaKMeD, as estratégias de pré-processamento e aumento de dados, e as métricas de avaliação adotadas. O Capítulo 4 apresenta a principal contribuição metodológica deste trabalho: a proposição de um modelo *ensemble* inovador que combina as arquiteturas EfficientViT, EVA-02 e EdgeNeXt, projetado para a classificação robusta de células cervicais.

Os resultados desta abordagem são apresentados e analisados no Capítulo 5, em que se demonstra o desempenho de estado da arte alcançado pela proposta, validado por um rigoroso protocolo de validação cruzada *5-fold*. Esse capítulo também detalha a análise comparativa com outros modelos e a comprovação estatística da superioridade do *ensemble*. Por fim, o Capítulo 6 expõe as conclusões da pesquisa, sintetizando as contribuições e propondo perspectivas para investigações futuras.

2 Revisão da Literatura

A inteligência artificial (IA) tornou-se uma ferramenta indispensável em diversas áreas, com notáveis avanços no setor da saúde. No contexto da oncologia, especificamente no diagnóstico do câncer cervical, tecnologias de IA têm demonstrado um potencial significativo para aprimorar a precisão e a eficiência dos diagnósticos. A visão computacional, alicerçada por técnicas de aprendizado de máquina, emerge como uma abordagem promissora, permitindo a análise automatizada de imagens citológicas para identificar alterações pré-cancerígenas com maior confiabilidade.

Este capítulo apresenta uma revisão sistemática da literatura de 2024 sobre as principais metodologias de visão computacional para a classificação de imagens de células cervicais, complementando revisões anteriores que cobriram o período até 2023. A análise foca estritamente em abordagens de classificação, excluindo trabalhos que abordam apenas detecção de regiões de interesse ou segmentação. A metodologia de busca e seleção, baseada no protocolo PRISMA, identificou 47 artigos relevantes após a triagem em bases como IEEE, ACM, Scopus, Web of Science e PubMed.

Na Seção 2.1 será apresentada a metodologia aplicada para a concepção da revisão. A Seção 2.2 apresenta os resultados referentes à revisão e, por último, na Seção 2.3 foram discutidos os resultados. As abordagens identificadas foram categorizadas em: Redes Neurais Convolucionais (CNNs), Vision Transformers (ViTs), Híbridas, Ensemble e Alternativas. A análise que se segue explora as tendências e os resultados mais relevantes dentro de cada categoria.

2.1 Metodologia de Revisão da Literatura

Para o desenvolvimento desta revisão sistemática, utilizou-se a ferramenta StArt (*State of the Art through Systematic Review*) (FABBRI et al., 2012) baseada na metodologia PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) (MOHER et al., 2010). A metodologia foi dividida em cinco partes:

1. **Planejamento:** Definição das questões de pesquisa e das palavras de busca com base na temática do trabalho, além da definição dos critérios de seleção;

2. **Execução:** Busca das palavra-chaves nas bases bibliográficas escolhidas;

3. **Seleção:** Seleção a partir da leitura das partes principais do trabalho, respondendo aos critérios de inclusão e exclusão;

4. **Extração:** Leitura completa dos trabalhos selecionados, respondendo às questões de pesquisa e aplicando novamente os critérios de inclusão e exclusão;
5. **Sumarização:** Criação de uma tabela comparativa entre os trabalhos selecionados e seus resultados.

2.1.1 Definição das Questões de Pesquisa

A revisão sistemática objetiva responder à questão principal a seguir: "Quais abordagens mais recentes em visão computacional e inteligência artificial são utilizadas para classificação das células cervicais?". Tomando por base a questão principal, elaborou-se outras perguntas que foram utilizadas para cada artigo selecionado na etapa seguinte.

1. Qual é a abordagem proposta?
2. Quais foram os experimentos realizados?
    a) Quais conjuntos de dados foram utilizadas?
    b) Qual pré-processamento foi realizado no conjunto de dados?
    c) Como o conjunto de dados foi dividido?
    d) Quais metodologias e métricas de comparação foram usadas?
    e) Quais abordagens foram comparadas?
3. Quais são as abordagens que obtiveram os melhores resultados nos conjuntos de dados utilizados?

Este estudo teve como objetivo identificar os trabalhos mais promissores no campo da triagem e diagnóstico automatizado de células cervicais, analisando suas vantagens, desvantagens e os conjuntos de dados empregados. A pesquisa busca fornecer uma base sólida para pesquisadores iniciantes, facilitando a escolha de metodologias que apresentem as melhores métricas e que sejam o estado da arte, promovendo o acesso a soluções eficientes e acessíveis. Embora o foco principal esteja nos conjuntos de dados e nas abordagens metodológicas adotadas, o estudo também se propôs a destacar os resultados mais expressivos, evidenciando os avanços mais relevantes nessa área de pesquisa.

2.1.2 Definição das Palavras-Chave

Neste estudo, foram pesquisadas as bases bibliográficas IEEE[1], ACM[2], SCOPUS[3], Web of Science[4] e PubMed[5]. As principais *strings* de busca utilizadas foram "*Cervical*

[1] https://ieeexplore.ieee.org/ Acesso em: 01/11/2024
[2] https://dl.acm.org/search/advanced Acesso em: 04/11/2024
[3] https://www.scopus.com/ Acesso em: 01/11/2024
[4] https://webofscience.com/ Acesso em: 04/11/2024
[5] https://pubmed.ncbi.nlm.nih.gov/advanced/ Acesso em: 04/11/2024

cancer screening", "*Computer Vision*", "*Classification*", "*Deep Learning*" e "*Detection*" e "*WSI Analysis*". As bases IEEE, ACM, SCOPUS e Web of Science foram selecionadas devido à sua relevância e influência na área de ciências da computação, enquanto a base PubMed foi escolhida por sua importância no campo da saúde. A ferramenta StArt possui compatibilidade com as referências dessas bases bibliográficas, de modo que a ferramenta foi utilizada no processo de triagem para encontrar artigos duplicados.

2.1.3 Processo de Busca

O processo de busca foi estrategicamente definido para complementar as revisões sistemáticas existentes (JIANG et al., 2023; FANG et al., 2024b; KHARE et al., 2024; VARGAS-CARDONA et al., 2024), concentrando-se nos trabalhos publicados no ano de 2024. Para garantir a seleção de artigos de máxima relevância, foi construída uma *string* de busca abrangente, fundamentada na intersecção (operador `AND`) de três pilares conceituais:

1. **Domínio do Problema:** Palavras-chave que definem o contexto clínico, como *"Cervical cancer screening"*, *"Pap smear"* e *"Cervical cytology"*.
2. **Solução Tecnológica:** Termos que descrevem as técnicas de IA, como *"Deep Learning"*, *"CNN"* e *"Vision Transformer"*.
3. **Objetivo da Análise:** Palavras que especificam a tarefa computacional, como *"Classification"*, *"Diagnosis"* e *"WSI Analysis"*.

A combinação lógica destes três pilares permitiu filtrar eficientemente os estudos que não se alinhavam ao foco principal da pesquisa. O modelo genérico da *string* de busca foi então adaptado à sintaxe específica de cada base de dados consultada. A distribuição quantitativa dos artigos encontrados nesta fase inicial é apresentada na Tabela 1 e ilustrada na Figura 2.

Tabela 1 – Distribuição de publicações por fonte

Base bibliográfica	**Quantidade**	**Percentual (%)**
IEEE	75	23,60
ACM	53	16,70
SCOPUS	104	32,70
Web of Science	45	14,10
PubMed	41	12,90

2.1.4 Triagem dos Artigos

A próxima etapa é a seleção dos artigos de acordo com os critérios de inclusão e exclusão definidos em conformidade com as questões da pesquisa.

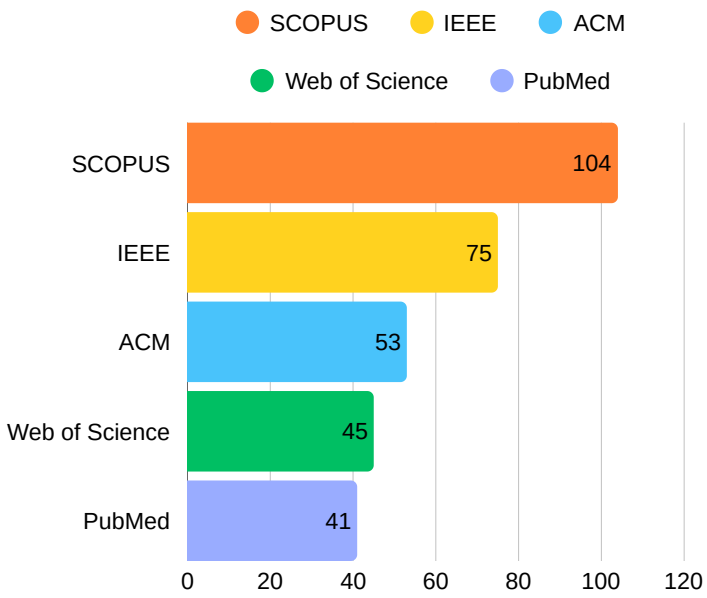


Figura 2 – Quantidades de trabalhos obtidos por base bibliográfica.

Critérios de inclusão

- Ser uma abordagem sobre classificação de células cervicais;
- Ser uma revisão sistemática sobre o tema;
- Ser uma abordagem que envolve detecção e classificação.

Critérios de exclusão

- Não ser um trabalho relacionado às células cervicais;
- Ser um trabalho apenas da área da saúde;
- Não utilizar abordagens relacionadas à visão computacional;
- Não ser possível acessar o artigo por meio público ou institucional;
- Ser uma abordagem relacionada somente com a segmentação;
- Ser um trabalho que não relata os algoritmos usados para classificação.

2.1.5 Processo de Seleção dos Estudos

O processo de seleção dos estudos ocorreu por meio da leitura dos títulos e dos *abstracts* de todos os trabalhos, aplicando-se os critérios de inclusão e exclusão. Para alguns trabalhos, não foi possível responder adequadamente às questões estabelecidas, exigindo, portanto, um processo de leitura parcial dos trabalhos para a realização da triagem.

De acordo com a Figura 3, a fase de seleção resultou na identificação de 97 artigos duplicados (30,5%), que foram excluídos. Após a aplicação dos critérios de exclusão, rejeitaram-se 146 trabalhos (45,9%), enquanto 75 artigos (23,6%) foram aceitos para a etapa de extração.



Figura 3 – Quantidade de trabalhos aceitos, rejeitados e duplicados na fase de seleção.

2.1.6 Extração de Dados e Sumarização

Durante a etapa de extração, os critérios de inclusão e exclusão continuaram a ser rigorosamente aplicados. Adicionalmente, as referências de cada estudo foram inspecionadas com o objetivo de identificar trabalhos adicionais potencialmente relevantes para a revisão.

Após a leitura completa e extração de dados dos trabalhos, na base de dados Web of Science foram aceitos apenas dois artigos: (JIANG et al., 2024) e (CHITRA; SUBBULEKSHMI, 2024). A base PubMed obteve apenas três trabalhos aceitos: (WU et al., 2024), (YI et al., 2024) e (YANG et al., 2024a). Na ACM foram aceitos quatro artigos: (PACAL, 2024),(ABD-ALHALEM et al., 2024), (DEVARAJ et al., 2024) e (SHARMA; KUMAR; SHARMA, 2024). Na Scopus, foram aceitos vinte trabalhos. Enquanto na IEEE, foram aceitos dezoito artigos.

A Figura 4 ilustra a quantidade de trabalhos aceitos e rejeitados nesta etapa do processo. Observa-se que 60,00% dos estudos provenientes da triagem anterior foram aprovados para a próxima fase.

A etapa subsequente consistiu na sumarização dos trabalhos selecionados, na qual foram sintetizadas as principais questões abordadas nos artigos aceitos, com o objetivo de responder às perguntas de pesquisa previamente estabelecidas. Tais questões encontram-se detalhadas na Seção 2.1.1.

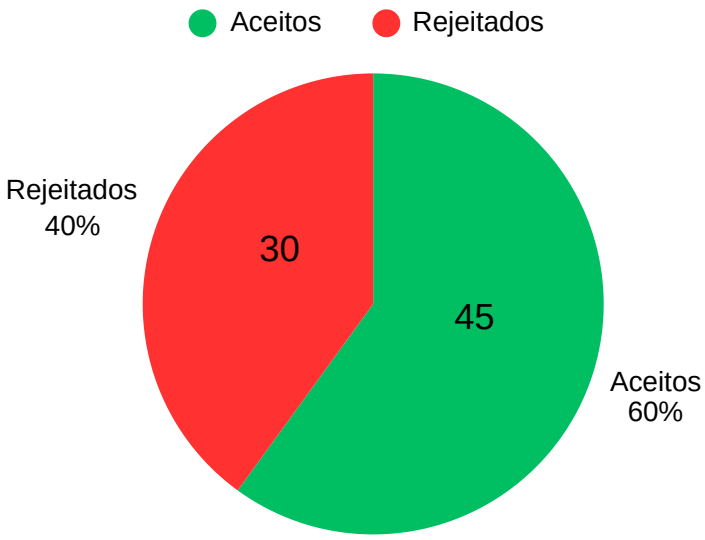


Figura 4 – Quantidade de trabalhos aceitos, rejeitados e duplicados na fase de extração.

2.2 Resultados da Revisão da Literatura

A primeira revisão sistemática encontrada foi a de Jiang *et al.* (JIANG et al., 2023), em que foi realizada uma análise de 2016 até 2022 sobre a classificação, detecção, segmentação e análise da lâmina inteira (*Whole Slide Analysis* – WSI). Considerando que a revisão foi publicada em 2023, não houve necessidade de fazer uma nova revisão sistemática naquela época. Em termos de classificação das células cervicais, a revisão de Jiang *et al.* relatou trabalhos envolvendo transferência de aprendizado (*Transfer Learning*), *ensemble* de modelos e abordagens baseadas em fusão híbrida de características (*Hybrid feature fusion*).

Foram encontradas seis revisões sistemáticas, publicadas no ano de 2024, por meio das bases bibliográficas. A revisão de Fang *et al.* (FANG et al., 2024b) analisou os artigos de 2017 até 2023 que abordam segmentação e classificação das células cervicais que apresentaram algumas arquiteturas de CNNs e de ViTs. Essa revisão também relatou abordagens para classificação como: transferência de aprendizagem, métodos baseados em *ensemble*, métodos baseados em classificadores únicos e em métodos baseados em fusão de características (*feature fusion*). O estudo destacou a necessidade de melhorar a generalização dos modelos e lidar com o desequilíbrio de classes nos conjuntos de dados.

Os autores Khare *et al.* (KHARE et al., 2024) apresentaram uma revisão sistemática que empregou inteligência artificial para detecção automatizada de câncer cervical, analisando diversas modalidades como citologia, colposcopia, dados sociodemográficos e imagens de tomografia computadorizada, ressonância magnética e tomografia por emissão de pósitrons. A revisão fez uma análise dos trabalhos de 2013 até 2023, identificou lacunas em pesquisas anteriores e destacou a necessidade de conjuntos de dados mais diversos e

acessíveis, além de maior foco em explicabilidade de modelos e quantificação de incertezas para aumentar a confiança dos clínicos. Finalmente, o artigo sugeriu que o aprendizado federado e o meta-aprendizado devem ser explorados para superar desafios relacionados à privacidade de dados e adaptabilidade.

A revisão sistemática de Wubineh, Rusiecki e Halawa (WUBINEH; RUSIECKI; HALAWA, 2024b) analisou estudos publicados entre 2016 e 2023 que utilizaram técnicas de aprendizado profundo para segmentar e classificar imagens de esfregaços de Papanicolaou na detecção de câncer cervical. A revisão investiga os conjuntos de dados mais usados, algoritmos de segmentação (como U-Net e Mask R-CNN) e algoritmos de classificação como Máquinas de Vetores de Suporte (*Support Vector Machines* – SVM) e CNN, destacando a alta acurácia obtida (acima de 90% em vários casos). Os autores discutiram as limitações dos métodos atuais e sugeriram direções futuras para pesquisa, como a inclusão de imagens histopatológicas e o uso de conjuntos de dados maiores e mais diversos.

O artigo científico de Vargas-Cardona *et al.* (VARGAS-CARDONA et al., 2024) revisou a literatura sobre o uso de IA, especificamente aprendizado de máquina e *deep learning*, para o rastreio e diagnóstico de câncer cervical a partir de 2009 a 2022. A análise de 32 estudos mostrou que algoritmos como CNN, ResNet, SVM e YOLO+EfficientNetB0 apresentaram os melhores resultados de diagnóstico, com acurácia acima de 97% em alguns casos. No entanto, a revisão destacou a heterogeneidade dos estudos e a necessidade de pesquisas futuras para validar esses resultados em contextos clínicos e populações diversas, especialmente em países de baixo recurso. A conclusão é que, embora promissoras, as abordagens envolvendo visão computacional e inteligência artificial ainda precisam de mais desenvolvimento e validação antes de serem aplicadas no rastreio do câncer cervical.

O artigo de Mathew e Cheriyan (MATHEW; CHERIYAN, 2024) revisou o uso de inteligência artificial para classificação e análise de fatores de risco do câncer cervical. Diversas técnicas de aprendizado de máquina são comparadas em sua capacidade de classificar imagens de exames citológicos e outras modalidades de imagem para detectar diferentes estágios da doença. O estudo também examinou fatores de risco, incluindo infecção por HPV, tabagismo e idade, e analisa métodos para identificar esses fatores utilizando algoritmos de aprendizado de máquina. A precisão dos modelos de classificação foi avaliada, e os resultados demonstraram o potencial da IA para auxiliar no diagnóstico e prevenção do câncer cervical. Finalmente, o estudo apresentou uma tabela comparando diferentes métodos de classificação em termos de precisão e características.

O artigo de Wu *et al.* (WU et al., 2024) explorou o potencial da inteligência artificial (IA) para revolucionar o rastreio e o diagnóstico do câncer do colo do útero, especialmente em países em desenvolvimento com infraestrutura de saúde limitada. Os autores descreveram como a IA pode ser utilizada para prever o risco de cancro do colo do útero usando dados clínicos, analisar imagens citológicas para identificar células anormais, e interpretar

imagens de colposcopia para auxiliar no diagnóstico e na orientação de biópsias. O artigo destacou a capacidade da IA de aumentar a precisão, a eficiência e o acesso aos cuidados de saúde, ao mesmo tempo em que reconhece os desafios relacionados à disponibilidade de dados, interpretabilidade e considerações éticas. Os autores argumentaram que a IA, quando combinada com diretrizes claras e regulamentação, pode desempenhar um papel crucial na eliminação global do câncer do colo do útero.

A revisão sistemática proposta neste capítulo tem como objetivo complementar as revisões anteriores para o ano de 2024. Vale ressaltar que as abordagens encontradas na realização desta revisão foram classificadas em cinco categorias:

1. CNN: As abordagens que utilizaram somente CNNs para classificação das imagens, sejam pré-treinadas ou não.

2. ViT: As abordagens que utilizaram somente *Vision Transformers* ou abordagens derivadas para a classificação das células.

3. Híbrida: Considerou-se as abordagens sendo híbridas aquelas que misturaram algum extrator de característica com algum otimizador ou seletor de característica e um outro tipo de classificador. Um exemplo proposto por Almohimeed *et al.* (ALMOHIMEED et al., 2024) foi a utilização do *Vision Transformer* com o otimizador *Particle Swarm Optimazion* (PSO) com o classificador SVM.

4. *Ensemble*: São abordagens que combinam as previsões de múltiplos modelos (chamados de modelos base ou *weak learners*) para produzir uma previsão mais robusta, precisa e confiável do que qualquer modelo individual isolado.

5. Alternativa: São abordagens que utilizam técnicas e algoritmos menos comuns ou que não se encaixam diretamente nas categorias previamente definidas.

A Tabela 2 apresenta os trabalhos coletados de acordo com suas respectivas categorias. Pode-se observar que há uma alta quantidade de trabalhos que se enquadram na categoria híbrida.

Tabela 2 – Quantidade de trabalhos coletados para revisão sistemática de acordo com sua categoria proposta.

Categoria da abordagem	**Quantidade de artigos**
CNN	10
ViT	1
Híbrida	22
Ensemble	3
Alternativa	3

2.2.1 Abordagens CNNs

A Tabela 3 apresenta a sumarização dos resultados obtidos na revisão sistemática, destacando os modelos que utilizaram exclusivamente redes neurais convolucionais (CNN) para extração e classificação de características.

Tabela 3 – Sumarização da quantidade de modelos utilizadas na literatura e os conjuntos de dados utilizados por eles.

Modelo utilizado	**Número de trabalhos**	**Conjunto de dados**
Xception	2	SIPaKMeD e RepoMedUNM
MobileNetV2	1	SIPaKMeD
ResNet50V2	1	Herlev
ResNet18	1	SIPaKMeD
Inception	1	Privado
DenseNet121	1	Conjunto de dados não disponivel
Cervical Net	1	SIPaKMeD
Demais	1	SIPaKMeD

Maesaroh *et al.* (MAESAROH et al., 2024) empregaram a rede Xception pré-treinada no conjunto RepoMedUNM, obtendo 95% de acurácia para quatro classes e 100% para classificação binária. As imagens foram redimensionadas para 224×224 e aumentadas por inversões e rotações. O conjunto foi dividido em 80% para treinamento, 15% para validação e 5% para teste.

Na mesma linha, Prome e Tahiti (PROME; TAHITI, 2024) utilizaram a MobileNetV2 no conjunto SIPaKMeD, alcançando 95% de acurácia em cinco classes, com precisão de 95,46%, *recall* de 95,20% e F1 de 95,23%. As imagens foram redimensionadas para 224×224 e padronizadas, sendo divididas em 90% para treinamento e 10% para validação.

De forma comparativa, Ramesh *et al.* (RAMESH et al., 2024) avaliaram as redes VGG16, VGG19, Inception e ResNet50 em um conjunto privado, destacando a Inception com 98% de acurácia. O pré-processamento envolveu redimensionamento, cisalhamento e inversão das imagens, com divisão em 80% para treinamento e 20% para validação.

Goswami e Sampathila (GOSWAMI; GOSWAMI; SAMPATHILA, 2024) também exploraram o conjunto SIPaKMeD, comparando ResNet18, ResNet50, GoogleNet e SqueezeNet, destacando a ResNet18 com 98,51% de acurácia. Foram utilizadas 397 imagens para teste e as demais para treinamento e validação, com melhor desempenho observado após 30 épocas de treinamento.

No estudo conduzido por Upadhyay *et al.* (UPADHYAY et al., 2024), foram avaliadas as redes DenseNet121, VGGNet e GoogleNet no SIPaKMeD, considerando diferentes taxas de aprendizado. O VGGNet apresentou o melhor resultado, com acurácia de 98,93%, precisão de 97,40%, *recall* de 96,80% e F1 de 95,30%.

Joynab *et al.* (JOYNAB et al., 2024) propuseram uma abordagem com CNN baseada em aprendizagem federada (FL), comparando-a com modelos clássicos de ML

no SIPaKMeD. A CNN-FL alcançou 94,30% de acurácia e 98,20% de AUC, explorando divisões IID e não IID, além de diferentes estratégias de aumento e pré-processamento.

Shandilya *et al.* (SHANDILYA et al., 2024) utilizaram a Xception no SIPaKMeD, atingindo 95,70% de acurácia. O conjunto foi dividido em 80% para treino, 10% para validação e 10% para teste, com imagens redimensionadas e normalizadas.

Já Anand e Bachhal (ANAND; BACHHAL, 2024) propuseram a *Cervical Net*, uma arquitetura própria que superou o VGG16, alcançando acurácia de 92,00% após 30 épocas no SIPaKMeD. O conjunto foi dividido em 70% para treino, 15% para validação e 15% para teste.

No trabalho de Devaraj *et al.* (DEVARAJ et al., 2024), foram comparados os modelos ResNet50V2, InceptionV3 e Xception no conjunto Herlev (sete classes), destacando a ResNet50V2 com 97,71% de acurácia e F1 médio de 98,57%.

Por fim, Kaur *et al.* (KAUR et al., 2024) analisaram VGGNet, GoogleNet e DenseNet121 em um conjunto de esfregaços cervicais, obtendo 93% de acurácia com DenseNet121, além de maior *recall* (95%) e AUC (0,97).

2.2.2 Abordagem ViT

A abordagem proposta por Yi *et al.* (YI et al., 2024) é um modelo chamado *Multi-scale Window Transformer* (MWT) para reconhecimento de imagens de citopatologia cervical. O MWT foi projetado para integrar simultaneamente características celulares de diferentes escalas usando atenção multi-cabeça de janela multi-escala. Ao contrário de outros ViTs, o design de *multi-scale window multi-head self-attention* (MW-MSA) permite a integração de características de longo alcance, evitando *Self-Attention* (SA) de imagem inteira no ViT ou duas vezes SA de janela local no *Swin Transformer*.

Os autores utilizaram o conjunto de dados obtidos do Hospital de Saúde Materno-Infantil de *Hubei* e do *Tongji Medical College* da Universidade de Ciência e Tecnologia de *Huazhong*, na província de *Hubei*, China, e também o Mendeley LBC (quatro classes). O conjunto de dados foi dividido em 80% para treinamento, 10% para validação e 10% para teste. As imagens de entrada foram redimensionadas para 224×224 pixels. O modelo proposto atingiu os seguintes resultados:

- Duas classes: Acurácia de 96,1% no conjunto de teste interno, 94,3% no conjunto de teste externo Data3 e 88,9% no conjunto de teste externo Data 8.
- Quatro classes: Acurácia de 89,1% no conjunto de teste interno e 79,1% no conjunto de teste externo Data 3.
- Mendeley LBC: Acurácia de 94,7%.

2.2.3 Abordagens Híbridas

As abordagens híbridas representam um avanço significativo no diagnóstico assistido por computador do câncer cervical, buscando superar as limitações de modelos baseados unicamente em uma arquitetura, como CNNs, ViTs ou algoritmos clássicos de aprendizado de máquina. Elas são desenvolvidas para obter características mais discriminativas e ricas, melhorando a capacidade de classificação das células cervicais. Essas abordagens combinam diferentes técnicas ou modelos para aproveitar os pontos fortes de cada um, resultando em desempenho aprimorado.

O artigo de Fang *et al.* (FANG et al., 2024a) propôs um bloco de Fusão Integrada de Características Profundas (*Deep Integrated Feature Fusion* – DIFF) que combina características locais, extraídas por CNNs, e características globais, extraídas por ViTs, para a classificação de imagens citológicas cervicais. O objetivo é obter características mais discriminativas e ricas para melhorar a classificação de células cervicais. Todas as imagens foram redimensionadas para 128×128 pixels. Isso foi feito para padronizar o tamanho das imagens e reduzir a complexidade computacional. Foi utilizado um esquema de validação cruzada *k-fold* com $k = 5$.

A abordagem dos autores, com o uso dos blocos DIFF, demonstrou os melhores resultados em comparação com 20 métodos concorrentes em três conjuntos de dados principais (SIPaKMeD, CRIC e Herlev), superando métodos baseados em CNNs, *Transformers*, e técnicas híbridas. Obteve como resultados acurácia de 97,83% no conjunto de dados Herlev, 100% no conjunto de dados Mendeley LBC e 98,52% no conjunto de dados SIPaKMeD.

O artigo científico de Alohali *et al.* (ALOHALI et al., 2024) apresentou um modelo inovador, denominado Swin-GA-RF, para a classificação do câncer cervical em imagens de esfregaços de Papanicolau. Esse modelo integra a arquitetura *Swin Transformer* com um algoritmo genético para seleção de características e utiliza uma floresta aleatória como classificador, superando modelos baseados em CNNs pré-treinadas em termos de precisão. O conjunto de dados utilizado foi o SIPaKMeD, com aplicação de técnicas de aumento, incluindo *flips*, recortes e redimensionamento. A divisão dos dados foi realizada em 70% para treinamento, 10% para validação e 20% para teste.

A abordagem Swin-GA-RF foi comparada com o *Swin Transformer* e CNNs pré-treinadas como DenseNet121, VGG16, ResNet18 e AlexNet. Para a classificação em cinco classes, as três melhores abordagens foram: Swin-GA-RF (98,808% de acurácia), ResNet18 (97,81% de acurácia) e *Swin Transformer* (97,044% de acurácia). Para a classificação binária, destacaram-se: Swin-GA-RF (99,012% de acurácia), *Swin Transformer* (97,535% de acurácia) e ResNet18 (95,435% de acurácia).

Analogamente ao Swin-GA-RF (ALOHALI et al., 2024), O ViT-PSO-SVM foi

desenvolvido por Almohimeed *et al.* (ALMOHIMEED et al., 2024) que combinou ViT para extração de características, otimização por enxame de partículas (*Particle Swarm Optimization* – PSO) para seleção das mais relevantes e SVM para classificação, utilizando GradCAM para explicabilidade. Testado nos conjuntos SIPaKMeD e Herlev, com pré-processamento padrão e divisão dos dados em em 70% para treinamento, 25% para teste e 5% para validação, foi comparado a CNNs pré-treinadas e modelos híbridos. O ViT-PSO-SVM obteve as maiores acurácias: 97,247% (SIPaKMeD, cinco classes) e 97,778% (Herlev, duas classes).

O artigo de Li *et al.* (LI et al., 2024) propôsa VTCNet, arquitetura que integra CNNs e ViT para classificação de células cervicais, utilizando módulos como como Focus (para amostragem sem perdas), *Spatial Pyramid Pooling Fusion* (SPPF) e convoluções separáveis. Após pré-processamento com redimensionamento, aumentos de dados e normalização, o modelo foi avaliado nos conjuntos SIPaKMeD (validação cruzada 5-fold) e Herlev (80% para treinamento, 10% para validação, e 10% para teste). A VTCNet obteve as maiores acurácias: 97,16% (SIPaKMeD) e 95,95% (Herlev), superando modelos CNNs e híbridos como DIFF e DeepCervix.

M2Net proposto por Yang *et al.* (YANG et al., 2024b) é uma rede de fusão de decisão em duas vias que combina características manuais (ex.: LBP, razão núcleo/citoplasma) e profundas extraídas por uma DarkNet modificada. Avaliada nos conjuntos SIPaKMeD (60% para treino, 20% para validação e 20% para teste) e CCID (70% para treino, 20% para validação e 10% para teste), a M2Net superou métodos como DeepCervix, BFCNet e HCLCNN. Para o conjunto SIPaKMeD, o M2Net atingiu acurácia de 94,79%, precisão de 94,27%, recall de 94,42% e F1 de 94,34%. Para o CCID, o M2Net obteve 87,19% de acurácia, 83,82% de precisão, 82,41% de recall e 83,09% de F1.

O Sistema Integrado de Atenção-Transformador (*Integrated AttentionTransformer System* – IATS) proposto por Khowaja, Zou e Kui (KHOWAJA; ZOU; KUI, 2024) é é uma abordagem que integra ViT, Token-to-Token (T2T) e a rede Squeeze-and-Excitation (SeNet) com ResNet101 e DenseNet169 para análise citológica, utilizando tokenização progressiva e votação ponderada. Após pré-processamento e divisão de 80% para treinamento e 20% para validação, o modelo alcançou 98,44% de acurácia no Mendeley LBC e 95,88% no SIPaKMeD, superando outros métodos e mostrando potencial para aplicação clínica.

A metodologia proposta por Qin *et al.* (QIN et al., 2024) baseia-se em uma abordagem de aprendizado comparativo para classificação de imagens de citologia cervical em lâminas inteiras (WSIs). O método emprega dois modelos pré-treinados para detectar células normais e anormais, cujas características são extraídas por meio de aprendizado contrastivo auto-supervisionado e posteriormente fundidas por codificadores *Transformer*, com classificação final realizada por aprendizado de instâncias múltiplas (MIL). As WSIs, digitalizadas e divididas em segmentos de 512×512 pixels, compõem um conjunto de

dados com 2177 imagens provenientes de cinco centros médicos, sendo divididas em 70% para treinamento e 30% para teste. A abordagem alcançou 88,29% de acurácia, 79,73% de sensitividade, 92,10% de especificidade e 93,19% de AUC, demonstrando desempenho competitivo na tarefa de classificação de WSIs cervicais.

Akbar *et al.* (AKBAR et al., 2023) propuseram uma abordagem de classificação de câncer cervical utilizando imagens citológicas de Papanicolaou, combinando redes neurais convolucionais (ResNet50 e ResNet50V2) para extração de características com máquinas de vetores de suporte (SVMs) para a etapa de classificação. O pré-processamento envolveu o redimensionamento das imagens e sua normalização. O conjunto de dados utilizado foi o SIPaKMeD, que contém imagens de células cervicais normais e patológicas classificadas em cinco categorias de malignidade, sendo dividido em 80% para treinamento, 15% para validação e 5% para teste. Os autores investigaram quatro configurações distintas, variando o congelamento de camadas e o uso de camadas de *pooling*, e obtiveram como melhor resultado uma acurácia de 97,09% com o modelo ResNet50V2 aliado ao classificador SVM, evidenciando o potencial dessa combinação para sistemas de diagnóstico assistido por computador.

A pesquisa de Nour *et al.* (NOUR et al., 2024) introduziu o método CACCD-GOADL (*Computer Aided Cervical Cancer Diagnosis using Gazelle Optimizer Algorithm with Deep Learning*) para detecção e classificação do câncer cervical a partir de imagens de esfregaço de Papanicolau. A abordagem é composta por três etapas principais: extração de características com um modelo MobileNetv3 aprimorado, ajuste de hiperparâmetros utilizando o algoritmo de otimização Gazela (GOA) e classificação com o modelo *Stacked Extreme Learning Machine* (SELM). As melhorias no MobileNetv3 incluíram a remoção de camadas para reduzir latência e custo computacional, além da adoção da função de ativação *swish*. O GOA otimizou os hiperparâmetros do modelo com base em um processo de exploração e fuga inspirado no comportamento das gazelas, enquanto o SELM foi empregado para classificar as imagens em positivas ou negativas para câncer cervical. O conjunto de dados utilizado foi o Herlev, contendo 918 amostras distribuídas em sete classes, com divisão de 80% para treinamento e 20% para teste. A abordagem foi comparada a seis métodos existentes e obteve desempenho superior, com acurácia de 99,38%, precisão de 96,71%, *recall* de 97,45% e F1-score de 97,04%.

A abordagem de Mahajan e Kaur (MAHAJAN; KAUR, 2024) consiste em duas etapas principais: pré-processamento e classificação. Inicialmente, as imagens RGB do conjunto de dados SIPaKMeD foram convertidas para escala de cinza por meio da transformação CIELAB e filtradas com mediana adaptativa para redução de ruído. Em seguida, aplicou-se o método de agrupamento *K-means* para segmentar as imagens em três *clusters* distintos: núcleo, citoplasma e fundo. Para a etapa de classificação, foi utilizado um modelo ResNet50 pré-treinado combinado com a técnica de aprendizado progressivo, na qual

imagens em diferentes resoluções (64×64, 224×224, 512×512 e 1024×1024) foram utilizadas de forma sequencial, com congelamento dos parâmetros aprendidos em estágios anteriores. O conjunto SIPaKMeD foi dividido em 80% para treinamento e 20% para teste. Os melhores resultados foram obtidos com a combinação ResNet50 e *Random Forest* (97,89% de acurácia) e com o modelo dos autores Mahajan e Kaur (97,4% de acurácia), demonstrando a eficácia da técnica de aprendizado progressivo na detecção de câncer cervical.

O estudo de Wang *et al.* (WANG et al., 2024) apresentou o desenvolvimento e a validação de um sistema de rastreamento de câncer cervical (*Artificial Intelligence Cervical Cancer Screening* – AICCS), cuja abordagem metodológica combinou o algoritmo de aprendizado profundo RetinaNet, para detecção de anomalias em nível de *patch*, com o classificador de aprendizado de máquina Random Forest (RF), para a classificação da imagem de lâmina inteira (WSI). O conjunto de dados principal, originário do SYSMH, consistiu em 11.468 WSIs, que foram pré-processadas para exclusão de amostras de baixa qualidade e subsequentemente divididas em conjuntos de treinamento (9.316 imagens) e validação interna (2.152 imagens), além de validação externa com dois conjuntos de dados independentes (600 cada). A performance do sistema foi quantificada por meio das métricas de Área Sob a Curva (AUC) de 94,70%, sensibilidade de 94,60%, especificidade de 89% e acurácia de 89,2%.

O trabalho de Ashfaque *et al.* (ASHFAQUE; BEIJI; KUI, 2024) propôs o *Cervix Visionator ELM*, um sistema automatizado para a classificação de câncer cervical, cuja abordagem metodológica integra a arquitetura EfficientNet para a extração de características e um classificador baseado em *Extreme Learning Machine* (ELM). O modelo foi desenvolvido utilizando o conjunto de dados SIPaKMeD, que passou por um pré-processamento incluindo a redimensionalização das imagens, normalização de camada e técnicas de aumento de dados para mitigar o sobreajuste, com uma divisão de 70% para treinamento, 15% para validação e 15% para teste. A performance do sistema foi quantificada por meio das métricas de acurácia de 98,89%, precisão de 99,42%, *recall* de 97,87% e F1-score de 98,76%.

A abordagem de Sholik *et al.* (SHOLIK; FATICHAH; AMALIAH, 2024) propôs uma metodologia baseada na fusão de características extraídas por Redes Neurais Convolucionais (CNNs) e *Vision Transformers* (ViTs), seguida pela redução de dimensionalidade com Análise de Componentes Principais (*Principal Component Analysis* – PCA) e Análise Discriminante Linear (*Linear Discriminant Analysis* – LDA). O estudo utilizou três conjuntos de dados públicos (Herlev, Mendeley LBC e SIPaKMeD), com os dados sendo divididos em proporções variadas para treinamento, teste e validação (70%/15%/15%; 80%/10%/10%; 90%/5%/5%). A avaliação da performance foi realizada com múltiplos classificadores (SVM, KNN, MLP e LR), alcançando acurácias de 97,83% no conjunto

Herlev, 100% no Mendeley LBC e 98,52% no SIPaKMeD.

O artigo de Chitra e Subbulekshmi (CHITRA; SUBBULEKSHMI, 2024) propôs duas arquiteturas de aprendizagem profunda para melhorar a classificação do câncer cervical: MASO-DenseNet-121 otimizado e ST-CNN-LSTM. O MASO-DenseNet-121 utiliza o algoritmo MASO para otimizar hiperparâmetros da arquitetura DenseNet 121, enquanto o ST-CNN-LSTM emprega um classificador CNN-LSTM otimizado por Sooty Tern e o modelo KWFLICM para segmentação. A divisão dos dados é feita na proporção de 70% para treinamento e 30% para teste no conjunto Herlev. O algoritmo ST-CNN-LSTM apresentou a maior acurácia, atingindo 99,21% na detecção de câncer cervical. A arquitetura MASO-DenseNet-121 obteve uma acurácia de 98,11%, demonstrando um desempenho um pouco inferior ao ST-CNN-LSTM, mas ainda superior aos métodos tradicionais.

O trabalho de Ahmed *et al.* (AHMED et al., 2024) é composto por quatro fases: pré-processamento, extração e fusão de características com modelos de transferência, redução de características usando PCA, e classificação com camadas densas. A fusão de características é realizada com MobileNet, DenseNet, EfficientNet, Xception, ResNet e ResNet-50. A metodologia foi aplicada ao conjunto de dados SIPaKMeD, com as imagens redimensionadas para 64×64 pixels e uma divisão de 80% para treinamento e 20% para teste. Utilizando uma função de perda de entropia cruzada suavizada, o modelo proposto alcançou uma acurácia de 97%, superando outras configurações avaliadas no estudo.

O trabalho de Wubineh *et al.* (WUBINEH; RUSIECKI; HALAWA, 2024a) empregou uma Rede Gerativa Adversarial Convolucional Profunda com blocos Residuais (RES_DCGAN) para o aumento de dados. A abordagem de classificação utilizou modelos pré-treinados, como ResNet50V2, Xception e DenseNet121, aprimorados com uma camada de autoatenção (*self-attention*). A metodologia foi validada no conjunto de dados SIPaKMeD, com as imagens sendo pré-processadas e divididas em 80% para treinamento e 20% para teste. Os resultados obtidos para o conjunto SIPaKMeD foram: 95% de acurácia na classificação binária usando DenseNet121 com autoatenção, 94% de acurácia na classificação binária com ResNet50V2 e autoatenção e 92% de acurácia na classificação multiclasse usando DenseNet121 com autoatenção. O DenseNet121 com *self-attention*, aliado à técnica de aumento de dados RES_DCGAN obteve o melhor resultado.

O trabalho de Jiang *et al.* (JIANG et al., 2024) propôs o DSA-FFOGNet, um algoritmo que aprimora o Faster R-CNN com o DSA-ResNet50 para foco em áreas relevantes, *Path Aggregation Network* (PAN) para fusão multiescala, um cabeçote de detecção desacoplado e a função de perda *Origin Grouping Loss* (OGL), que aplica penalidades baseadas em conhecimento biomédico. A abordagem foi validada no conjunto de dados SIPaKMeD, que foi submetido a aumento de dados com inversão horizontal e dividido em 80% para treinamento, 10% para validação e 10% para teste. A avaliação de performance, medida pela acurácia, demonstrou que a arquitetura DSA-ResNet50, que

integra o módulo DSA ao ResNet50, alcançou 99,01% na classificação de células cervicais.

O artigo científico de Munirathinam *et al.* (MUNIRATHINAM et al., 2024) apresentou uma pesquisa sobre a categorização de anormalidades cervicais usando CNN e SVM. Os autores utilizaram a rede neural SqueezeNet para extrair características de imagens de esfregaços cervicais do banco de dados SIPaKMeD (multiclasses), posteriormente classificadas por SVM e KNN. O estudo avalia o desempenho do SVM utilizando quatro funções de Kernel diferentes: linear (SVM-L), polinomial (SVM-P), radial (SVM-R) e sigmoide (SVM-S). O KNN é avaliado com três métricas de distância: Manhattan, Euclidiana e Chebyshev. Para divisão de dados, utilizou-se a validação cruzada com 10 *folds*. Os melhores resultados foram: SVM com *kernel* RBF obteve 87% de acurácia,SVM com *kernel* polinomial obteve $86,7\%$ de acurácia e KNN com métrica de Manhattan e função de peso por distância obteve $82,7\%$ de acurácia.

O trabalho de Pacal (PACAL, 2024) introduziu o MaxCerVixT, um modelo leve que combina arquiteturas ViT e CNN, aprimorando o MaxViT com blocos ConvNeXtv2 e módulos MLP baseados em Normalização de Resposta Global (GRN). O MaxCerVixT foi validada nos conjuntos de dados SIPaKMeD e Mendeley LBC, que foram submetidos a técnicas de aumento de dados e divididos em 70% para treinamento, 10% para validação e 20% para teste. A avaliação de performance, medida pela acurácia, demonstrou que o modelo proposto alcançou 99,02% no SIPaKMeD e 99,48% no Mendeley LBC.

O RP-BCNN, desenvolvido por Abd-alhalem *et al.* (ABD-ALHALEM et al., 2024), visa classificar imagens de câncer cervical utilizando uma arquitetura de Rede Neural Convolucional Bilinear (*Bilinear Convolutional Neural Network* – BCNN) aprimorada com Projeção Aleatória (*Random Projection* – RP). O objetivo é aumentar a precisão e reduzir o tempo de processamento na classificação de imagens, combinando a capacidade da BCNN de extrair características complexas com a técnica de redução de dimensionalidade da RP. O conjunto de dados Herlev foi dividido em 70% para treinamento e 30% para teste. Dois experimentos comparativos foram realizados: um para a classificação das sete categorias entre os algoritmos RP-BCNN (95,30% de acurácia) e BCCN (94,43% de acurácia) e outro para classificação binária envolvendo as duas abordagens com 99,83% de acurácia para o RP-BCNN e 99,74% para BCCN.

O trabalho de Yang *et al.* (YANG et al., 2024a) apresentou a FPNC, uma nova rede neural profunda que integra informações de textura e estrutura morfológica por meio de módulos de fusão e purificação. A abordagem foi validada nos conjuntos de dados SIPaKMeD e no recém-introduzido CCID, com as imagens redimensionadas e uma divisão de 70% para treinamento, 10% para validação e 20% para teste. A avaliação de performance, medida pela acurácia, demonstrou que o modelo proposto alcançou 99,38% de acurácia para duas classes e para cinco classes, obteve acurácia de 95,90% no conjunto SIPaKMeD; no conjunto CCID, o modelo alcançou acurácia de 96,23% para classificação binária (normal e

anormal) e, para 15 classes obteve uma acurácia de 89,50%. O FPNC se destacou, por sua capacidade de lidar com redundância de informações e extrair características altamente discriminativas, o que contribui para sua robustez e generalização.

O trabalho de Sharma e Kumar (SHARMA; KUMAR; SHARMA, 2024) descreveu o CervixNet, um modelo de diagnóstico que utiliza o conceito de "digital twin", combinando Internet das Coisas (IoT), aprendizado profundo e de máquina. A abordagem metodológica, validada no conjunto de dados SIPaKMeD com validação cruzada de 5 folds, incluiu o pré-processamento de imagens com filtro de média e a técnica CLAHE para realce de contraste. A avaliação de performance, medida pela acurácia, demonstrou que o CervixNet com um classificador SVM alcançou 98,91% de acurácia.

2.2.4 Abordagens *Ensemble*

Os trabalhos encontrados na literatura que empregaram técnicas de *ensemble* alcançaram acurácia superior a 90%, evidenciando o potencial e a relevância dessa abordagem para a tarefa de classificação. Em razão desse destaque, a presente dissertação optou por adotar a metodologia de *ensemble* em sua proposta. A seguir, são apresentados e discutidos os trabalhos identificados na literatura que utilizaram essa técnica.

O trabalho de Khanarsa e Kitsiranuwat (KHANARSA; KITSIRANUWAT, 2024) focou na comparação de uma técnica de *ensemble*, utilizando modelos como o ResNet152V2 com pontuação máxima de probabilidade, contra modelos (VGG19, NASNetLarge e EfficientNetB7) de forma isolada. A abordagem foi validada no conjunto de dados SIPaKMeD, que passou por um pré-processamento de extração manual de células individuais e foi dividido em 64% para treinamento, 16% para validação e 20% para teste. A avaliação de performance, medida pela acurácia, demonstrou que a técnica de *ensemble* com ResNet152V2 alcançou 97%, superando significativamente os modelos individuais.

O trabalho de Liu *et al.* (LIU et al., 2024) apresentou um modelo de *ensemble* que combina três redes convolucionais (DenseNet-169, VGG-19, Xception) e um Swin *Transformer*. A abordagem metodológica empregou *transfer learning* e otimização bayesiana para ponderar o *ensemble*, e foi validada nos conjuntos de dados SIPaKMeD e Mendeley LBC utilizando validação cruzada de 5 *folds*. A avaliação de performance, medida pela acurácia, demonstrou que o modelo proposto alcançou 95,50% no conjunto SIPaKMeD e 98,65% no conjunto Mendeley LBC.

O trabalho de S. Vaishnodevi *et al.* (S. et al., 2024) investigou a utilização de algoritmos de aprendizado de máquina como AdaBoost, XGBoost, CatBoost e LightGBM, para a predição de câncer cervical. A abordagem metodológica, validada no conjunto de dados SIPaKMeD, incluiu pré-processamento com normalização e codificação de características, além do uso de validação cruzada. A avaliação de performance demonstrou que

o modelo XGBoost apresentou o melhor desempenho geral, alcançando uma acurácia de 99,70%, uma precisão de 96,40%, um *recall* de 97,50% e um F1 de 96,9%. Seguido por AdaBoost que também demonstrou um desempenho sólido, alcançando uma acurácia de 97,88%, uma precisão de 94,94%, um *recall* de 97,63% e um F1 de 96,42%.

2.2.5 Abordagens Alternativas

Diversos estudos exploram métodos avançados para otimizar a classificação de imagens de câncer cervical. As abordagens a seguir destacam-se por suas estratégias inovadoras, que vão desde o foco em exemplos difíceis até o uso de otimizadores bioinspirados e arquiteturas de extração de características.

O trabalho de Song *et al.* (SONG et al., 2024) propôs o algoritmo *worse-case boosting*, que concentra o treinamento em exemplos de difícil classificação, identificados pela norma do gradiente. A abordagem foi validada nos conjuntos de dados SIPaKMeD e LCPSI, com as imagens redimensionadas para 64x64 pixels e utilizando validação cruzada de 5 *folds*. A avaliação de performance, medida pela acurácia, demonstrou que o método alcançou 91,5% no SIPaKMeD e 85,7% no LCPSI.

Em outra frente, Anand *et al.* (ANAND et al., 2024) introduziram o modelo JSODL-CCRC, que combina um otimizador bioinspirado, o *Jellyfish Search Optimizer* (JSO), com técnicas de aprendizado profundo. A arquitetura utiliza a rede InceptionV3 para a extração de características das imagens e uma *Extreme Learning Machine* (ELM) para a etapa de classificação. O pré-processamento consistiu na aplicação de um filtro de média para redução de ruído, e o conjunto de dados foi dividido em 70% para treino e 30% para teste. O modelo JSODL-CCRC demonstrou um desempenho excepcional no conjunto Herlev, atingindo 99,24% de acurácia e superando outras abordagens como EOEL-PCLCCI (99,17%), posicionando-se como uma ferramenta de alta precisão para o diagnóstico automatizado.

Por sua vez, Akhila e Ahilan (AKHILA; AHILAN, 2024) propuseram uma abordagem em duas etapas, focada na extração eficiente de características. Primeiramente, um *autoencoder* é utilizado para aprender representações latentes das imagens de esfregaços cervicais. Em seguida, um perceptron de múltiplas camadas em cascata (c-MLP) realiza a classificação final. Após a aplicação de um filtro mediano adaptativo para remoção de ruído, o método foi avaliado no dataset Herlev (dividido em 71% para treino e 29% para teste). A combinação de autoencoder e c-MLP alcançou 97,63% de acurácia, mostrando-se superior tanto a algoritmos tradicionais de aprendizado de máquina, como SVM (94,69%), quanto a outras abordagens baseadas em redes convolucionais (94,05%).

2.3 Discussão da Revisão de Literatura

Nesta seção, serão apresentados e sumarizados os resultados obtidos a partir da revisão sistemática, com o objetivo de destacar as abordagens mais promissoras e eficazes no campo da visão computacional aplicada ao diagnóstico do câncer cervical. As metodologias analisadas foram agrupadas em categorias representativas das principais tendências técnicas, como CNNs, ViTs, abordagens híbridas, modelos *ensemble* ou abordagens alternativas.

2.3.1 Conjuntos de Dados

A primeira análise em relação aos trabalhos seria a utilização dos conjuntos dados. De acordo com a Tabela 4 houve uma predominância da utilização do conjunto SIPaKMeD (PLISSITI et al., 2018), composto de 4096 imagens em cinco categorias, com 27 trabalhos utilizando-o para treinamento e teste. O segundo mais utilizado é o Herlev (JANTZEN et al., 2005) sendo utilizado em 10 trabalhos. Para mais detalhes sobre o conjunto, veja a Seção 3.1.1.

Tabela 4 – Análise quantitativa dos conjuntos de imagens utilizadas pelos trabalhos.

Conjunto de dados	**Quantidade de trabalhos**
SIPaKMeD	27
Herlev	10
Mendeley LBC	5
RepoMedUNM	1
CRIC	1
Demais conjuntos	8

O Mendeley LBC (HUSSAIN, 2019) é um conjunto composto de 963 imagens divididas em quatro classes de acordo com o sistema *Bethesda* de classificação.

O conjunto de dados RepoMedUNM (RIANA et al., 2021) consiste em 7.437 imagens de células agrupadas e divididas em oito categorias. As imagens foram obtidas por meio do microscópio óptico *OLYMPUS CX33RTFS2* e do microscópio *X52-107BN* com câmeras *Logitech*.

O CRIC (REZENDE et al., 2021) é um conjunto de dados composto por 400 imagens (1.376 × 1.020 pixels) com 11.534 células classificadas. O sistema *Bethesda* é utilizado para dividir as células em seis categorias.

Os demais conjuntos são conjuntos relatados que não estão disponíveis publicamente ou que não foram citados nos trabalhos.

2.3.2 Sumarização dos Resultados

Conforme a Tabela 5, apenas três trabalhos apresentaram acurácia inferior a 90%, sendo eles: (WANG et al., 2024; QIN et al., 2024; MUNIRATHINAM et al., 2024). Para efeito comparativo dos resultados da Tabela 5, iremos considerar a seguinte siglas acurácia (acc), conjunto de dados (conjunto) e abordagem proposta (abordagem), para o propósito de estilização, a coluna resultados será simplificada. Por outro lado, os cinco trabalhos com os melhores desempenhos em termos de acurácia foram:

1. (SHOLIK; FATICHAH; AMALIAH, 2024), com 100% de acurácia no conjunto Mendeley LBC, utilizando como abordagem a integração de CNNs com ViTs.
2. (ABD-ALHALEM et al., 2024), com $99,83\%$ de acurácia para a classificação binária no conjunto Herlev, empregando a abordagem RP-BCNN.
3. (S. et al., 2024), com $99,70\%$ de acurácia no conjunto SIPaKMeD, utilizando como abordagem o método XGBOOST.
4. (NOUR et al., 2024), com $99,38\%$ de acurácia no conjunto Herlev, baseado na abordagem CACCD-GOADL.
5. (YANG et al., 2024a), com $99,38\%$ de acurácia para a classificação binária no conjunto SIPaKMeD, utilizando a abordagem FPNC.

Entre os trabalhos analisados, observa-se que a maioria se enquadra na categoria de abordagens híbridas, enquanto apenas um pertence à categoria *ensemble*. Esse panorama evidencia a predominância, no estado da arte atual, de soluções híbridas, que combinam diferentes metodologias com o intuito de aprimorar a eficiência e a acurácia dos modelos.

Neste trabalho, optou-se pela categoria *ensemble*, descrita no Capítulo 4, como uma estratégia para explorar a utilização de redes neurais mais inovadoras e, assim, potencializar o desempenho do modelo. Para assegurar a robustez e a consistência dos resultados, foi empregada a técnica de validação cruzada com cinco *folds*.

Tabela 5 – Sumarização dos resultados de acurácia obtidos pelos artigos da revisão sistemática. Os resultados que não especificam o número de classes referem-se a casos onde essa informação não foi disponibilizada no artigo original.

Autor(es)	**Abordagem**	**Conjunto(s)**	**Resultados**
Fang *et al.*	DIFF	Sipakmed, Cric e Herlev	Acc.: SIPaKMeD: 96,02%, CRIC: 85,06%, Herlev: 94,55%.

Tabela 5 – Sumarização dos resultados de acurácia obtidos pelos artigos da revisão sistemática. Os resultados que não especificam o número de classes referem-se a casos onde essa informação não foi disponibilizada no artigo original.

Autor(es)	**Abordagem**	**Conjunto(s)**	**Resultados**
Alohali *et al.*	Swin-GA-RF	Sipakmed	Acc.: SIPaKMeD (binário): 99,01%, SIPaKMeD (5 classes): 98,81%.
Almohimeed *et al.*	ViT-PSO-SVM	SIPaKMeD e Herlev	Acc.: Herlev (binário): 97,78%, SIPaKMeD (5 classes): 97,25%, SIPaKMeD (binário): 99,11%.
Li *et al.*	VTCNet	SIPaKMeD e Herlev	Acc.: SIPaKMeD (5 classes): 97,16%, Herlev (7 classes): 95,95%.
Yang *et al.*	M2Net	SIPaKMeD e CCID	Acc.: SIPaKMeD (5 classes): 95,24%, CCID: 87,19%.
Khowaja, Zou e Kui	IATS	SIPaKMeD e Mendeley LBC	Acc.: SIPaKMeD (5 classes): 95,88%, Mendeley LBC (4 classes): 98,44%.
Song *et al.*	Worse-case boosting	SIPaKMeD e LCPSI	Acc.: SIPaKMeD: 91,5%, LCPSI: 85,7%.
Khanarsa e Kitsiranuwa	*Ensemble* ResNet152V2	SIPaKMeD	Acc.: SIPaKMeD: 97%.
Qin *et al.*	Aprendizado contrastivo autosupervisionado	5 centros médicos	Acc.: WSI dataset privado: 88,29%.
Akbar *et al.*	ResNet50V2 com SVM	SIPaKMeD	Acc.: SIPaKMeD: 97,09%.
Nour *et al.*	CACCD-GOADL	Herlev	Acc.: Herlev (7 classes): 98,66%.
Maesaroh *et al.*	Xception	RepoMedUNM	Acc.: RepoMedUNM (4 classes): 95%, RepoMedUNM (binário): 100%.

Tabela 5 – Sumarização dos resultados de acurácia obtidos pelos artigos da revisão sistemática. Os resultados que não especificam o número de classes referem-se a casos onde essa informação não foi disponibilizada no artigo original.

Autor(es)	**Abordagem**	**Conjunto(s)**	**Resultados**
Liu *et al.*	DenseNet-169, VGG-19, Xception e Swin Transformer	SIPaKMeD e Mendeley LBC	Acc.: SIPaKMeD: 95,50%, Mendeley LBC: 98,65%.
Mahajan e Kau	K-means com ResNet50 e aprendizado progressivo	SIPaKMeD	Kappa: SIPaKMeD: 98%.
Wang *et al.*	AICCS	SYSMH	Acc.: SYSMH (WSI): 89,2%.
Prome e Tahiti	MobileNetV2	SIPaKMeD	Acc.: SIPaKMeD (5 classes): 95%.
Ramesh *et al.*	Inception	Privado, disponibilizado por Patho.Ai	Acc.: Dataset privado (Patho.Ai): 98%.
Anand *et al.*	JSODL-CCRC	Herlev	Acc.: Herlev: 99,24%.
Ashfaque, Beiji e Kui	Cervix Visionator ELM	SIPaKMeD	Acc.: SIPaKMeD: 98,89%.
Sholik, Fatichah e Amalia	CNNs com ViTs	Herlev, Mendeley LBC e SIPaKMeD	Acc.: Herlev: 97,83%, Mendeley LBC: 100%, SIPaKMeD: 98,52%.
Akhila e Ahilan	Autoencoder com c-MLP	Herlev	Acc.: Herlev (binário): 97,63%.
Chitra e Subbulekshmi	ST-CNN-LSTM	Herlev	Acc.: Herlev: 99,21%.
Goswami e Sampathil	ResNet18	SIPaKMeD	Acc.: SIPaKMeD (5 classes): 98,51%.

Tabela 5 – Sumarização dos resultados de acurácia obtidos pelos artigos da revisão sistemática. Os resultados que não especificam o número de classes referem-se a casos onde essa informação não foi disponibilizada no artigo original.

Autor(es)	**Abordagem**	**Conjunto(s)**	**Resultados**
Ahmed *et al.*	Smoothing cross entropy	SIPaKMeD	Acc.: SIPaKMeD (5 classes): 97%.
Wubineh, Rusiecki e Halawa	RES_DCGAN	SIPaKMeD	Acc.: SIPaKMeD (binário): 95%, SIPaKMeD (5 classes): 92%.
Joynab *et al.*	Uso de CNNs dentro de uma arquitetura federada	SIPaKMeD	Acc.: SIPaKMeD (5 classes): 94,36%.
Shandilya *et al.*	Xception	SIPaKMeD	Acc.: SIPaKMeD (5 classes): 95,70%.
Anand e Bachha	Cervical Net	SIPaKMeD	Acc.: SIPaKMeD (5 classes): 92%.
Devaraj *et al.*	ResNet50V2	Herlev	Acc.: Herlev (binário): 97%.
Kaur *et al.*	DenseNet121	Conjunto de dados livre com 8 mil imagens	Acc.: Dataset não especificado: 93%.
Munirathinam *et al.*	SqueezeNet com SVM(RBF)	SIPaKMeD	Acc.: SIPaKMeD (5 classes): 87%.
Pacal	MaxCerVixT	SIPaKMeD e Mendeley LBC	Acc.: SIPaKMeD (5 classes): 99,02%, Mendeley LBC (4 classes): 99,48%.
Abd-alhalem *et al.*	RP-BCNN	Herlev	Acc.: Herlev (binário): 99,83%, Herlev (7 classes): 95,30%.
S. Vaishnodevi *et al.*	XGBoost	SIPaKMeD	Acc.: SIPaKMeD: 99,70%.

Tabela 5 – Sumarização dos resultados de acurácia obtidos pelos artigos da revisão sistemática. Os resultados que não especificam o número de classes referem-se a casos onde essa informação não foi disponibilizada no artigo original.

Autor(es)	**Abordagem**	**Conjunto(s)**	**Resultados**
Yang *et al.*	FPNC	SIPaKMeD e CCID	Acc.: SIPaKMeD (5 classes): 95,9%, SIPaKMeD (binário): 99,38%, CCID (15 classes): 89,5%, CCID (binário): 96,23%.
Yi *et al.*	Multiscale Window Transformer	Mendeley LBC e outro conjunto de dados oriundo de hospitais da China	Acc.: Mendeley LBC (4 classes): 94,70%, Dataset interno (binário): 96,10%, Dataset interno (multiclasse): 89,10%.
Sharma e Kuma	CervixNet com SVM	SIPaKMeD	Acc.: SIPaKMeD (5 classes): 98,91%.

3 Materiais e Métodos

A metodologia empregada para desenvolver a solução proposta nesta dissertação é detalhada neste capítulo. A descrição está organizada em três seções principais: primeiramente, são apresentados os conjuntos de imagens que compõem a base de dados. Em seguida, são descritos os modelos de aprendizado de máquina escolhidos para o estudo. Por fim, são detalhadas as métricas de avaliação utilizadas para quantificar a performance da abordagem, cujos resultados são explorados no Capítulo 4.

3.1 Conjuntos de Dados

Nesta seção, são apresentados os conjuntos de imagens de células cervicais disponíveis publicamente. O primeiro conjunto a ser analisado é o Herlev, por sua relevância e pioneirismo na literatura. E, por último, o SIPaKMeD, um conjunto de dados mais recente e com uma melhor qualidade em relação ao Herlev.

3.1.1 Herlev

O conjunto de dados Herlev, proposto por Jantzen *et al.* (JANTZEN et al., 2005) foi escolhido por ser bastante utilizado na literatura para classificação e segmentação de células cervicais. Possui 917 imagens de esfregaço cervical obtidas por meio da técnica Papanicolau. Sendo dividido em sete classes (três normais e quatro anormais) de acordo com a Tabela 6, a Figura 5 ilustra as células normais e. As resoluções das imagens são 0,201 μm por pixel. É importante notar que há um desbalanceamento nas células, com a quantidade de células anormais sendo quase três vezes maior que a de células normais. Esse aspecto deve ser levado em conta ao desenvolver um modelo de classificação.

Tabela 6 – Classificação do conjunto de dados *Herlev* com 917 células (242 normais e 675 anormais).

Categoria	**Tipo de célula**	**Quantidade**
Normal	*Superficial squamous epithelial*	74
	Intermediate squamous epithelial	70
	Columnar epithelial	98
Total Normal		242
Anormal	*Mild squamous non-keratinizing dysplasia*	182
	Moderate squamous non-keratinizing dysplasia	146
	Severe squamous non-keratinizing dysplasia	197
	Squamous cell carcinoma in situ intermediate	150
Total Anormal		675
Total		917

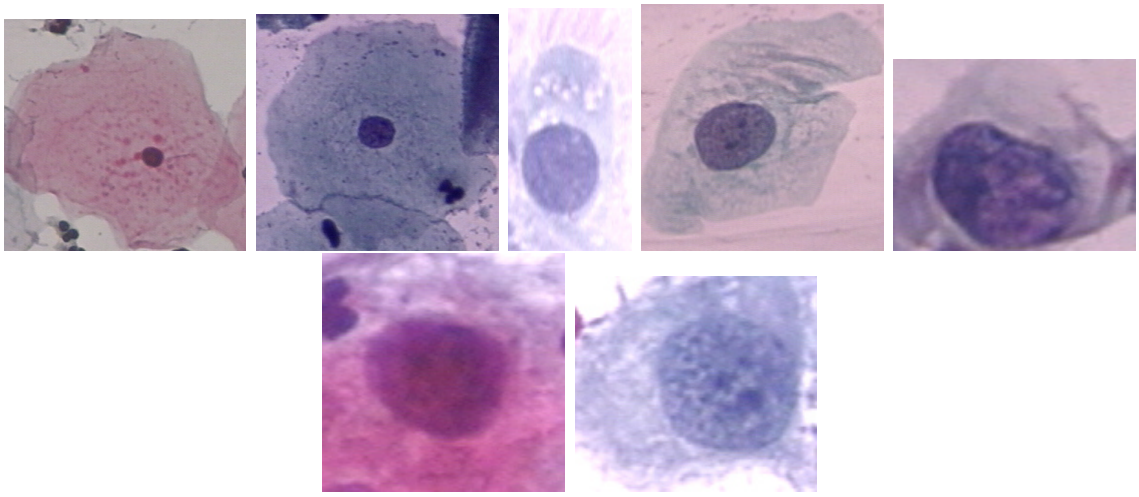


Figura 5 – Células normais e anormais, as três primeiras figuras representam as células normais enquanto as demais são as células anormais seguindo a Tabela 6. Fonte: (JANTZEN et al., 2005)

3.1.2 SIPaKMeD

O conjunto de dados SIPaKMeD de Plissiti *et al.* (PLISSITI et al., 2018) é composto por 4049 imagens de células isoladas, cuidadosamente recortadas a partir de 966 imagens de *clusters* celulares provenientes de lâminas de exame de Papanicolau. As imagens foram obtidas por meio da câmera digital CCD Infinity 1 (Lumenera) ao microscópio óptico (OLYMPUS BX53F). A classificação das células foi realizada por citopatologistas especialistas, que categorizam as amostras em cinco classes diferentes baseadas nas características citomorfológicas. As categorias, os tipos de células e suas respectivas distribuições são descritas pela Tabela 7, a Figura 6 ilustra as células divididas por suas respectivas classes. Os tipos de células podem ser descritas como:

Tabela 7 – Conjunto de imagens SIPaKMeD com suas respectivas categorias.

Categoria	Tipo de célula	Qtd. imagens de cluster celulares	Qtd. imagens de células cortadas
Normal	*Superficial/Intermediate*	126	831
	Parabasal	108	787
Total Normal		234	1618
Anormal	*Koilocytotic*	238	825
	Dyskeratotic	223	813
Total Anormal		461	1638
Benígno	*Metaplastic*	271	793
Total:		**966**	**4.049**

- *Superficial/Intermediate*: células epiteliais escamosas normais, com citoplasma amplo e poligonal, núcleo central pequeno e picnótico, coradas com eosina ou cianofilina.
- *Parabasal*: células epiteliais escamosas imaturas, de menor dimensão, com citoplasma cianofílico e núcleo vesicular grande, morfologicamente semelhantes às células *metaplastic*.
- *Koilocytotic*: células patognomônicas para HPV, com cavidade perinuclear proeminente, núcleo hipercromático e membrana nuclear irregular. Frequentemente binucleadas ou multinucleadas.

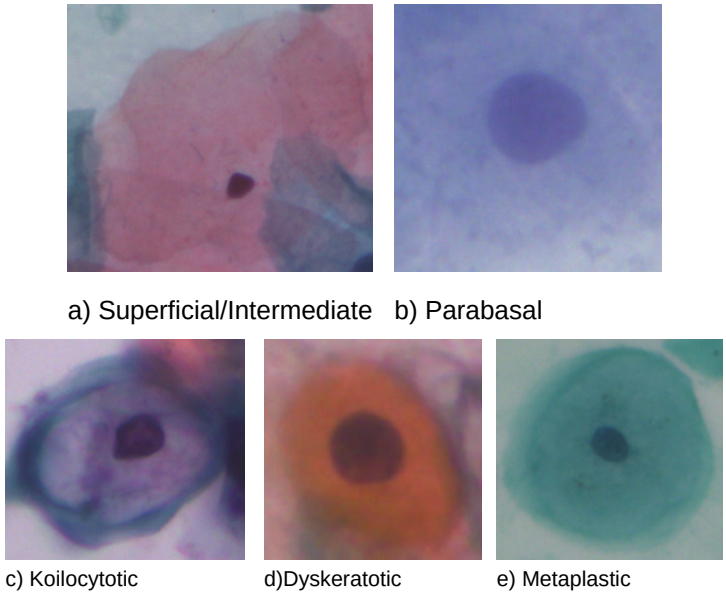


Figura 6 – Células do conjunto SIPaKMeD. As figuras ‘a’ e ‘b’ representam células normais. Enquanto as figuras ‘c’, ‘d’ e ‘e’ representam células anormais. Fonte: (PLISSITI et al., 2018)

- *Dyskeratotic*: células com queratinização anormal, citoplasma laranja brilhante e núcleo vesicular. Associadas à infecção por HPV, mesmo na ausência de células *Koilocytotic*.

- *Metaplastic*: células relacionadas à zona de transformação, com bordas celulares bem definidas, citoplasma arredondado e núcleo excêntrico. Associadas a lesões pré-cancerosas.

3.2 Classificadores

Nesta seção, são apresentados diferentes classificadores de aprendizado de máquina empregados em estudos recentes, incluindo redes neurais multicamadas (MLP), redes neurais convolucionais (CNNs), a arquitetura *Transformer* e sua variação baseada em imagens, o *Vision Transformer* (ViT).

No contexto desta dissertação, propõe-se um classificador baseado em *ensemble*, composto por três modelos: EfficientViT (CAI et al., 2022), EVA02 (FANG et al., 2024c) e EdgeNeXt (MAAZ et al., 2022). Os dois primeiros são modelos fundamentados em ViTs, enquanto o EdgeNeXt combina elementos de CNN e ViT, explorando as vantagens complementares de ambas as arquiteturas.

Os resultados obtidos com esses classificadores, detalhados no Capítulo 5, foram determinantes para a validação e consolidação da abordagem proposta, evidenciando sua robustez e potencial no contexto da classificação automatizada de células cervicais.

3.2.1 Redes Neurais Artificias

Uma rede neural artificial (RNA) é um sistema computacional modelado a partir da estrutura e do funcionamento biológico do cérebro humano, empregado frequentemente na solução de problemas de alta complexidade. Em uma RNA, cada neurônio simulado recebe sinais de neurônios conectados a ele em camadas anteriores. Esses sinais são processados e, somente se a soma ponderada desses sinais ultrapassar um certo limiar, o neurônio artificial emite um sinal para os neurônios subsequentes. Esse mecanismo permite que a rede realize tarefas de aprendizado de máquina, como reconhecimento de padrões e processamento de linguagem natural, com eficiência e adaptabilidade.

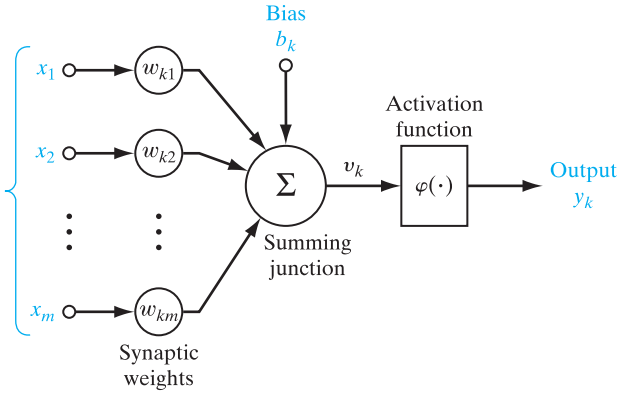


Figura 7 – Representação de um neurônio artificial (Perceptron), em que $(x_1, x_2, ..., x_n)$ são os sinais de entrada do neurônio, $(w_{k1}, w_{k2}, ..., w_{kn})$ são os pesos associados aos sinais de entrada, b_k é o viés do neurônio. $\phi(\cdot)$ é a função de ativação que processa os sinais e y_k é o sinal de saída. Fonte: (SIMON, 2009)

O neurônio artificial, elemento central dos sistemas de redes neurais artificiais, é inspirado na estrutura e funcionamento dos neurônios biológicos. Esse neurônio também pode ser denominado como Perceptron e sua estrutura é representada pela Figura 7. A analogia começa com as sinapses, que em um neurônio artificial são representadas pelos pesos sinápticos, atribuindo importância variável aos diferentes sinais de entrada. A soma ponderada desses sinais, uma combinação linear, corresponde ao potencial de ação no contexto biológico. Por fim, a função de ativação em um neurônio artificial determina se o sinal será transmitido adiante, similar ao limiar de excitação de um neurônio biológico. Esses componentes trabalham em conjunto para permitir que as redes neurais artificiais aprendam e tomem decisões a partir de dados complexos, simulando, de certa forma, o processamento de informações realizado pelo cérebro humano (SIMON, 2009).

Os *Multilayer Perceptrons* (MLPs), ou perceptrons multicamadas, são uma classe avançada de redes neurais artificiais que superam as limitações dos perceptrons simples ao permitir a classificação de dados não linearmente separáveis (SIMON, 2009). Isso é possível graças à sua arquitetura mais complexa, que inclui múltiplas camadas ocultas e o uso de funções de ativação não lineares diferenciáveis, como a função sigmoide ou a ReLU (*Rectified Linear Unit*). Essas funções permitem que os MLPs criem fronteiras de decisão mais flexíveis e lidem com uma variedade maior de padrões de dados. No entanto, essa complexidade traz consigo o desafio da

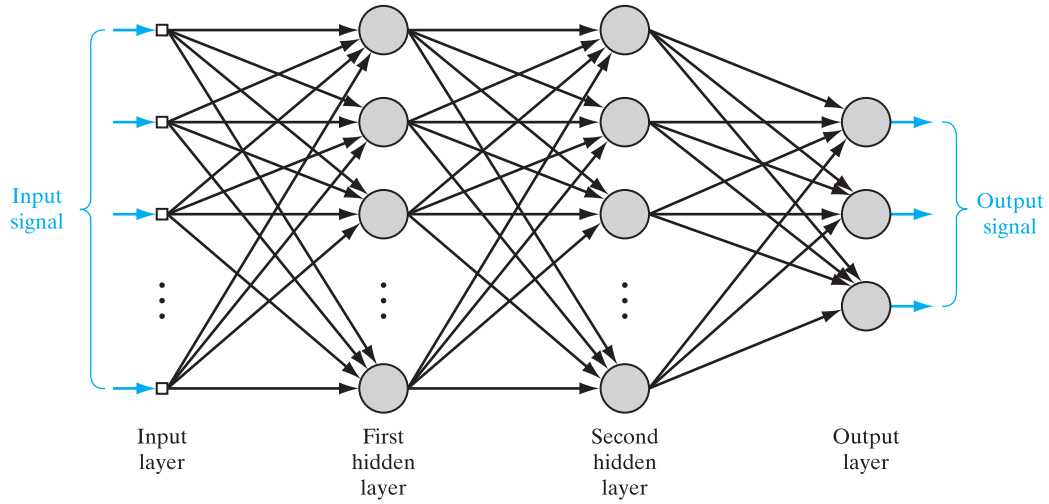


Figura 8 – Representação de um MLP. Fonte: (SIMON, 2009)

opacidade, em que o processo de tomada de decisão da rede não é facilmente interpretável pelos humanos (SIMON, 2009). Isso é frequentemente referido como o problema da "caixa preta" em aprendizado de máquina, em que, ao contrário das árvores de decisão que oferecem uma estrutura clara e facilmente compreensível, os MLPs ocultam o processo interno de como as entradas são transformadas em saídas.

3.2.2 Redes Neurais Convolucionais

As redes neurais convolucionais (*Convolutional Neural Network* – CNN) são redes neurais que utilizam convolução no lugar da multiplicação de matrizes gerais em pelo menos uma de suas camadas (GOODFELLOW; BENGIO; COURVILLE, 2016). As CNNs foram introduzidas por Lecun *et al.* (LECUN et al., 1998) para a classificação de dígitos manuscritos. Diferentemente dos MLPs em que todos os neurônios da camada anterior estão conectados com os neurônios da próxima camada, as CNNs utilizam os chamados mapas de características (*feature maps*), representados na Figura 9 por planos paralelos ou canais.

Camada de Convolução

A camada de convolução é o pilar das Redes Neurais Convolucionais (CNNs), projetada para extrair características hierárquicas de dados com topologia de grade, como as imagens. A sua operação central é a **convolução**, que consiste em aplicar um filtro, também conhecido como *kernel*, sobre os dados de entrada para gerar **mapas de características** (*feature maps*) (SZELISKI, 2022).

Esse conceito é análogo à filtragem em processamento digital de imagens, em que filtros pré-definidos são usados para borramento, detecção de bordas ou realce. A distinção fundamental em uma CNN é que os pesos do *kernel* não são fixos, eles são **parâmetros aprendidos** durante o treinamento. A rede neural otimiza esses pesos para que os *kernels* se tornem detectores de padrões relevantes para a tarefa, como texturas, cores e formas simples nas camadas iniciais, e estruturas mais complexas (como olhos ou rodas) em camadas mais profundas.

A principal inovação da camada de convolução reside em duas propriedades interligadas:

1. Pesos Compartilhados (*Weight Sharing*): Em vez do modelo aprender um conjunto de pesos para cada posição da imagem, o mesmo *kernel* (um único conjunto de parâmetros) é deslizado por toda a imagem para gerar um mapa de características (GOODFELLOW; BENGIO; COURVILLE, 2016). Isso resulta em uma drástica redução no número de parâmetros em comparação com uma camada totalmente conectada, o que torna o modelo: mais eficiente em memória por possuir menos parâmetros para armazenar e mais eficiente estatisticamente, ao possuir menos parâmetros para aprender, portanto evitando o*overfitting*.

2. Invariância à Translação: Uma consequência direta do compartilhamento de pesos é que a rede se torna capaz de detectar a mesma característica independentemente de sua posição na imagem. Se um *kernel* aprende a detectar uma borda vertical, ele a detectará tanto no canto superior esquerdo quanto no centro da imagem. Essa propriedade, conhecida como **equivariância à translação** (a localização da característica no *output* muda conforme a localização no *input*), é fundamental para a generalização em tarefas de visão computacional.

Após a operação de convolução (estágio de filtragem), os mapas de características passam pelo estágio de detecção (*detector stage*) (GOODFELLOW; BENGIO; COURVILLE, 2016). Nesta fase, aplica-se uma função de ativação não-linear elemento a elemento, como a Unidade Linear Retificada (ReLU). A sua fórmula é dada por:

$$\text{ReLU}(z) = \max(0, z)$$

A inclusão da não-linearidade é crucial, pois permite que a rede aprenda relações complexas. Sem ela, uma sequência de camadas de convolução seria matematicamente equivalente a uma única e mais complexa operação linear, limitando severamente a capacidade de aprendizado do modelo.

Os mapas de ativação resultantes contêm as características extraídas e "ativadas", e servem como entrada para a próxima camada, que pode ser uma camada de *pooling* (para subamostragem) ou outra camada de convolução (para aprender características de nível ainda mais alto).

Camada de *Pooling*

A camada de *pooling* tem como função reduzir o tamanho das *feature maps*, minimizando a perda de informações. Para isso, busca preservar suas métricas estatísticas, o que torna o modelo invariante a pequenas translações na entrada (GOODFELLOW; BENGIO; COURVILLE, 2016).

A operação de *max pooling* retorna o valor máximo entre os pixels vizinhos dentro de uma região retangular. Outra operação comumente utilizada é a *average pooling*, que calcula o valor médio dos pixels vizinhos. Após a etapa de *pooling*, as *feature maps* geradas são transformadas

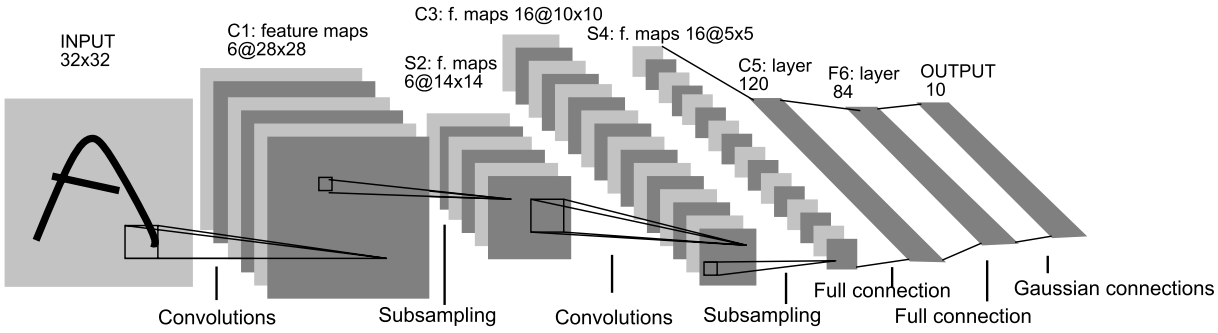


Figura 9 – Representação da arquitetura LeNet5 aplicada ao caractere A manuscrito. Fonte: (LECUN et al., 1998)

em um único vetor de características, que é processado por camadas totalmente conectadas (*fully connected layers*) de uma MLP para realizar a classificação.

3.2.3 Arquitetura *Transformer*

No campo do processamento de linguagem natural (*Natural Language Processing* – NLP) e transformação de sequências, que engloba tarefas como a tradução automática e a modelagem de linguagem, os modelos historicamente dominantes baseavam-se em redes neurais recorrentes (RNNs) complexas, em particular as LSTMs (HOCHREITER; SCHMIDHUBER, 1997) e GRUs (CHUNG et al., 2014). Frequentemente, essas arquiteturas eram organizadas no formato de codificador-decodificador, sendo consideradas o estado da arte para problemas em que uma sequência de entrada é convertida em uma sequência de saída.

No entanto, esses modelos recorrentes apresentavam uma limitação fundamental, sua natureza sequencial impedia a paralelização da computação ao longo das posições da sequência de entrada e saída. Isso significava que os estados ocultos eram gerados um após o outro, tornando o treinamento ineficiente, especialmente para sequências longas, devido a restrições de memória e tempo. Embora alguns trabalhos buscassem reduzir a computação sequencial utilizando convoluções para processamento paralelo, eles ainda enfrentavam desafios ao aprender dependências de longo alcance, onde o número de operações para relacionar posições distantes crescia linearmente ou logaritmicamente com a distância.

Em um avanço significativo para superar essas barreiras, foi proposta uma nova arquitetura de rede neural: o *Transformer* (VASWANI et al., 2017). O *Transformer* se destaca por dispensar completamente a recorrência e as convoluções, baseando-se unicamente em mecanismos de atenção para modelar dependências globais entre a entrada e a saída. Esta abordagem permitiu uma paralelização significativamente maior e a capacidade de aprender dependências entre posições de forma mais eficiente, exigindo um número constante de operações para relacionar quaisquer duas posições, independentemente de sua distância.

O conceito central do *Transformer* é o mecanismo de atenção, que permite ao modelo focar em partes relevantes da sequência de entrada ao gerar a saída, independentemente da distância entre os elementos. Mais especificamente, o *Transformer* utiliza intensivamente a auto-atenção (*self-attention*), que é um mecanismo de atenção que relaciona diferentes posições dentro de uma

única sequência para computar uma representação mais rica da mesma. O *Transformer* foi o primeiro modelo para transformação de sequências que dependia inteiramente da auto-atenção para computar representações de sua entrada e saída, sem o uso de RNNs ou convoluções alinhadas à sequência.

A arquitetura do *Transformer* mantém a estrutura geral de codificador-decodificador. Conforme a Figura 10, o codificador mapeia uma sequência de entrada para uma sequência de representações contínuas, e o decodificador gera a sequência de saída, um elemento por vez, de forma autorregressiva, consumindo os símbolos gerados anteriormente como entrada adicional. Ambos são construídos a partir de pilhas de camadas idênticas que empregam auto-atenção multi-cabeça (*multihead self-attention*) e redes *feed-forward*. A atenção multi-cabeça é um componente crucial, pois permite que o modelo processe informações de diferentes "subespaços" de representação simultaneamente, o que seria inibido por uma única cabeça de atenção.

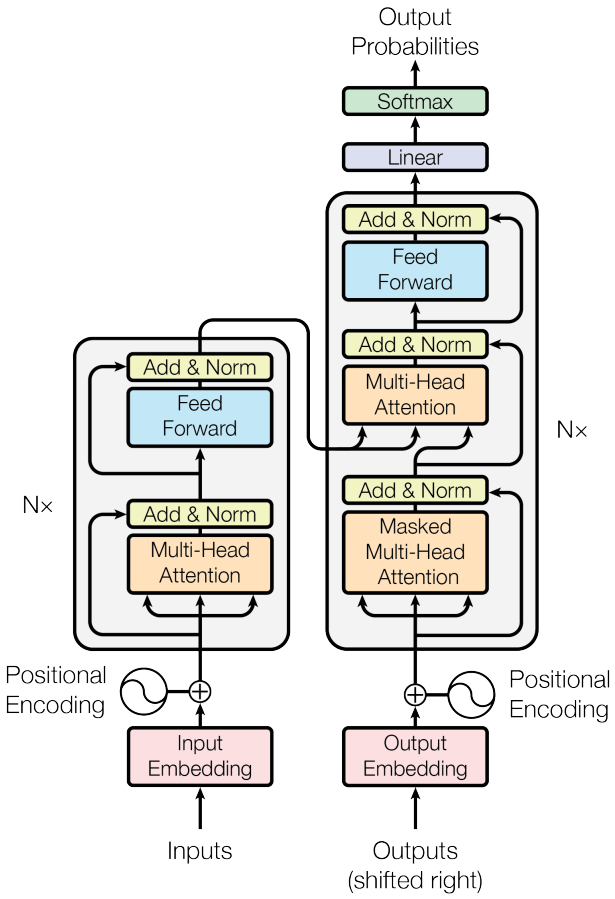


Figura 10 – A representação da arquitetura *Transformer*. Fonte: (VASWANI et al., 2017)

Dado que o *Transformer* não possui recorrência ou convolução, ele precisa de uma forma explícita de capturar a ordem da sequência. Isso é feito pela adição de "codificações posicionais" às *embeddings* de entrada nas camadas inferiores do codificador e decodificador, fornecendo informações sobre a posição relativa ou absoluta dos *tokens*.

Os experimentos com o *Transformer* demonstraram sua superioridade em qualidade e eficiência computacional. Em tarefas de tradução automática como *English-to-German* e *English-*

to-French do WMT 2014, o *Transformer* não apenas alcançou novos estados da arte, superando os melhores resultados existentes em mais de 2.0 BLEU, mas também o fez com um custo de treinamento significativamente menor, exigindo uma fração do tempo e recursos dos modelos anteriores de ponta. Isso solidificou o *Transformer* como uma arquitetura fundamental e um marco no campo do processamento de linguagem natural.

3.2.4 *Vision Transformer*

Nesse contexto, o surgimento dos *Vision Transformers* (ViT) representa um marco significativo, ao demonstrar que a dependência de CNNs não é inerentemente necessária para tarefas de reconhecimento de imagem. Inspirados pelo sucesso dos *Transformers* em NLP, a abordagem do ViT é surpreendentemente direta: aplicar um *Transformer* padrão diretamente a sequências de *patches* de imagem, com o mínimo de modificações possíveis. A premissa fundamental reside em tratar cada *patch* de uma imagem da mesma forma que os *tokens* ou palavras são tratados em uma aplicação de NLP. Essa analogia permite que a arquitetura do *Transformer*, intrinsecamente projetada para processar sequências unidimensionais, seja adaptada para a complexidade das imagens 2D. A Figura 11 ilustra a arquitetura de um *Vision Transformer*.

O processo de preparação da imagem para o *Transformer* começa com a divisão da imagem de entrada em *patches* de tamanho fixo. Por exemplo, uma imagem $x \in \mathbb{R}^{H \times W \times C}$ é remodelada em uma sequência de *patches* achatados $x_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$, em que (H, W) é a resolução da imagem original, C é o número de canais, (P, P) é a resolução de cada *patch* da imagem, e $N = HW/P^2$ é o número resultante de *patches*, que também serve como o comprimento efetivo da sequência de entrada para o *Transformer*. Cada um desses *patches* 2D é então achatado (*flattened*) e projetado linearmente para uma dimensão D constante que é mantida em todas as camadas do *Transformer*, gerando os *embeddings* de *patch*. As fontes indicam que os filtros de *embedding* linear da primeira camada do ViT-L/32 se assemelham a funções básicas plausíveis para uma representação de baixa dimensão da estrutura fina dentro de cada *patch*.

Para preservar a informação posicional, visto que a auto-atenção é global e não possui viés inerente à localização espacial, *embeddings* de posição treináveis são adicionados aos *embeddings* de *patch*. O modelo aprende a codificar a distância dentro da imagem na similaridade dos *embeddings* de posição, com *patches* mais próximos apresentando *embeddings* de posição mais semelhantes, e a estrutura de linha-coluna também se manifesta. É crucial notar que esta ajustagem de resolução e a extração de *patches* são os únicos pontos onde um viés indutivo sobre a estrutura 2D das imagens é manualmente injetado no *Vision Transformer*. Similar ao *token* `[class]` usado no BERT (DEVLIN et al., 2019), um embedding aprendível extra (x_{class}) é pré-adicionado à sequência. O estado final deste *token* na saída do *encoder* do *Transformer* serve como a representação agregada da imagem para a tarefa de classificação, à qual uma cabeça de classificação (MLP) é conectada.

A sequência resultante de vetores composta pelo: *token* de classificação, os *embeddings* de *patch* e os *embeddings* de posição. É então alimentada a um *encoder Transformer* padrão, que consiste em camadas alternadas de atenção multi-cabeça (MSA) e blocos MLP, com normalização

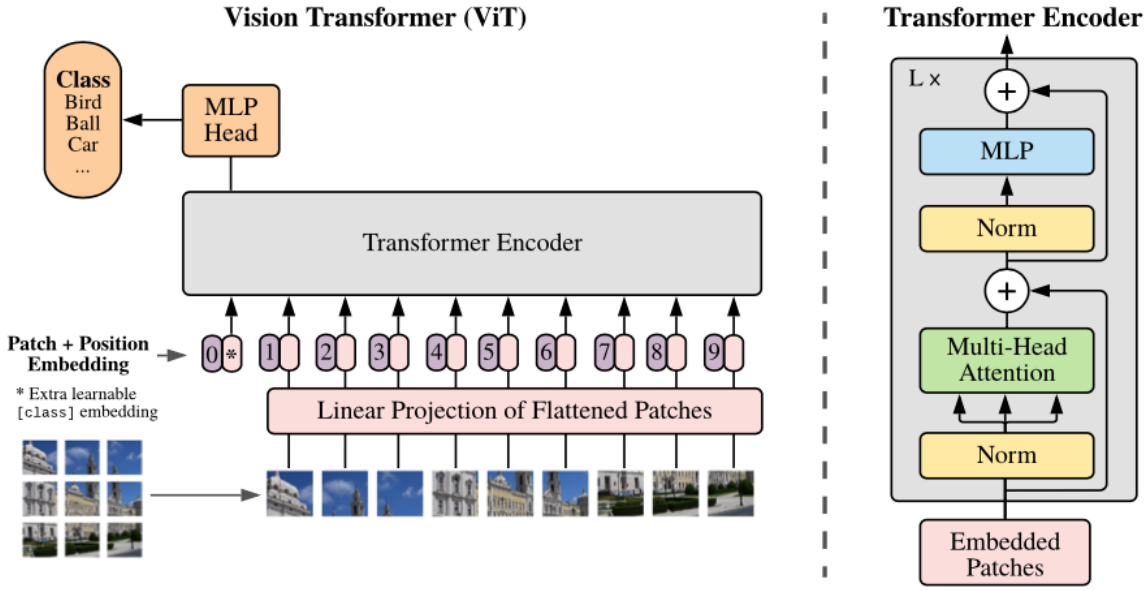


Figura 11 – A representação da arquitetura *Vision Transformer*. Fonte: (DOSOVITSKIY et al., 2021)

de camada e conexões residuais. Uma distinção fundamental do ViT em relação às CNNs é sua significativamente menor quantidade de viés indutivo específico para imagens. Enquanto as CNNs incorporam propriedades como localidade, estrutura de vizinhança bidimensional e equivariância translacional em cada camada, no ViT, apenas as camadas MLP são locais e translacionalmente equivariantes; as camadas de auto-atenção são globalmente conectadas. Isso implica que, à parte da etapa inicial de extração de *patches* e do ajuste de *embeddings* de posição durante o *fine-tuning*, todas as relações espaciais entre os *patches* devem ser aprendidas do zero.

Inicialmente, essa ausência de fortes vieses indutivos resulta em um desempenho modesto quando o ViT é treinado em conjuntos de dados de tamanho médio, como o *ImageNet*, sem forte regularização, ficando abaixo das ResNets de tamanho comparável. No entanto, o cenário muda drasticamente quando os modelos são pré-treinados em conjuntos de dados maiores, variando de 14 milhões a 300 milhões de imagens. Nessas condições, o treinamento em larga escala supera o viés indutivo, permitindo que o ViT atinja ou supere o estado da arte em diversas tarefas de reconhecimento de imagem, enquanto exige substancialmente menos recursos computacionais para o pré-treinamento em comparação com as CNNs de ponta.

A capacidade de auto-atenção permite que o ViT integre informações de toda a imagem desde as camadas mais baixas, com algumas cabeças de atenção já processando a maior parte da imagem em estágios iniciais, e a distância de atenção geralmente aumentando com a profundidade da rede. Em última análise, essa arquitetura simples, porém escalável, demonstra que a interpretação de uma imagem como uma sequência de *patches* e seu processamento por um *encoder Transformer* padrão é uma estratégia notavelmente eficaz, abrindo caminhos promissores para outras tarefas de visão computacional, como detecção e segmentação, e para o avanço de métodos de pré-treinamento auto-supervisionado.

3.3 Modelos utilizados no *ensemble*

Esta seção será explicada os modelos pré-treinados que são utilizados na abordagem proposta: o *EfficientVit*, EVA02, *EdgeNeXt*.

3.3.1 *EfficientViT*

A predição densa, definida como a classificação em nível de pixel de imagens de alta resolução, constitui uma tarefa essencial na visão computacional, com aplicações relevantes em áreas como fotografia computacional, condução autônoma e processamento de imagens médicas. Apesar de seu amplo potencial, a implementação de modelos de predição densa de última geração em dispositivos com restrições de hardware apresenta desafios significativos, sobretudo devido ao elevado custo computacional associado a essas arquiteturas e à limitação dos recursos disponíveis nesses dispositivos. Para alcançar um desempenho satisfatório, tais modelos exigem não apenas imagens de alta resolução, mas também uma forte capacidade de extração de informações contextuais.

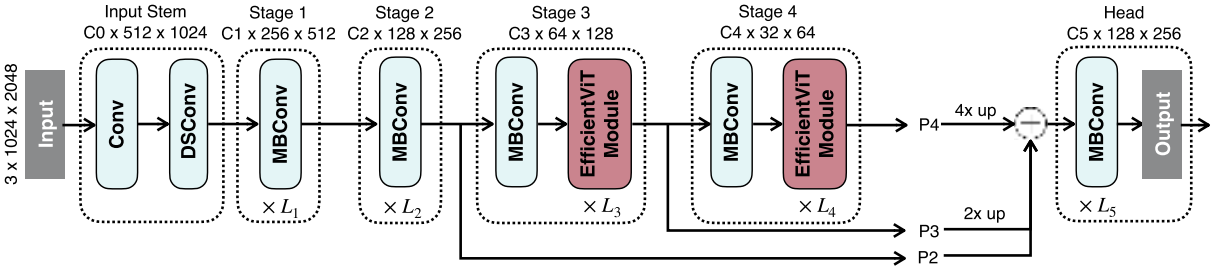


Figura 12 – A representação da arquitetura *EfficientVit*. Fonte: (CAI et al., 2022)

Neste contexto, o trabalho de Cai *et al.* (CAI et al., 2022) apresenta o *EfficientViT*, uma nova família de modelos de visão projetada especificamente para uma previsão densa de alta resolução eficiente.

As características distintivas da atenção linear multi-escala do *EfficientViT* são:

- **Campo Recetivo Global e Aprendizagem Multi-Escala**: Estas são duas características cruciais que os modelos do estado da arte anteriores demonstravam ser importantes para melhorar o desempenho dos modelos de previsão densa de alta resolução. O *EfficientViT* consegue obter estas características com operações leves e eficientes em hardware.
- **Substituição da Atenção Softmax Ineficiente**: Diferente de modelos anteriores como o SegFormer, que utiliza atenção softmax (com complexidade computacional quadrática em relação à resolução de entrada), o EfficientViT propõe substituir a atenção softmax pela atenção linear ReLU. Ao aproveitar a propriedade associativa da multiplicação de matrizes, a atenção linear ReLU reduz a complexidade computacional de quadrática para linear, preservando a funcionalidade e evitando operações ineficientes em hardware como o softmax. Esta otimização resulta numa velocidade significativamente maior em CPUs móveis.

- **Superação das Limitações da Atenção Linear ReLU**: A atenção linear ReLU por si só tem capacidade limitada, especialmente na extração de informações locais e na capacidade de aprendizagem multi-escala. Para mitigar esta limitação, o *EfficientViT* aprimora a atenção linear ReLU com convoluções. Especificamente, são inseridas convoluções de profundidade nas camadas FFN (*Feed-Forward Network*) para melhorar a capacidade de extração de características locais. Além disso, o módulo agrega informações de tokens Q/K/V (*Query/Key/Value*) vizinhos para obter tokens multi-escala usando convoluções de kernel pequeno e separáveis em profundidade, o que melhora a capacidade de aprendizagem multi-escala.

O EfficientViT adota uma arquitetura de backbone-cabeça/encoder-decoder padrão. O módulo de atenção linear multi-escala é integrado nas fases 3 e 4 do backbone, como ilustrado na Figura 12.

Os resultados demonstram que o EfficientViT oferece melhorias notáveis de desempenho e aceleração significativa em diversas plataformas de hardware, incluindo CPU móvel, GPU de ponta e GPU em nuvem. Em suma, o EfficientViT aborda a lacuna crítica entre o alto desempenho de modelos de previsão densa e as limitações de recursos de hardware, tornando os modelos do estado da arte mais práticos e implementáveis para aplicações do mundo real, através do seu design inovador de atenção linear multi-escala.

3.3.2 *EVA-02*

No campo da visão computacional, avanços recentes têm sido marcados pela escalabilidade de modelos de visão e visão-linguagem, nos quais o aumento de parâmetros, dados e poder computacional geralmente resulta em melhor desempenho. Contudo, essa escalada gerou uma lacuna crescente entre os modelos de ponta, que alcançam o estado da arte, e aqueles que são financeiramente acessíveis para a comunidade de pesquisa em geral. O treinamento, ajuste e avaliação de modelos de visão de grande escala demandam recursos computacionais significativos, tornando-se proibitivamente caros e demorados. Nesse contexto, Fang *et al.* (FANG et al., 2024c) apresentam o EVA-02, uma representação visual de nova geração, baseada em *Transformer*, que visa mitigar essa disparidade.

O EVA-02 é pré-treinado para reconstruir características visuais robustas e alinhadas à linguagem por meio da Modelagem de Imagem Mascarada (*Masked Image Modeling* – MIM). Sua estratégia se distingue por utilizar um codificador de visão de um modelo *Contrastive Language-Image Pre-Training* (CLIP) gigante e de acesso público como o gerador das representações-alvo para a tarefa de MIM.

A arquitetura do EVA-02 não se propõe a ser um projeto inteiramente novo, mas sim uma otimização robusta da arquitetura já estabelecida do *Vision Transformer* (ViT) plano. A premissa dos autores é que o potencial dos ViTs de tamanho moderado foi subestimado e que avanços significativos, já consolidados em modelos de linguagem, ainda não haviam sido explorados no contexto de aprendizagem de representação visual. Desse modo, a arquitetura

resultante, denominada pelos autores de *Transform Vision* (TrV), incorpora um conjunto de modificações microarquitetônicas que, em conjunto, aprimoram o desempenho e a estabilidade do treinamento (FANG et al., 2024c), pode-se verificar através da Figura 13.

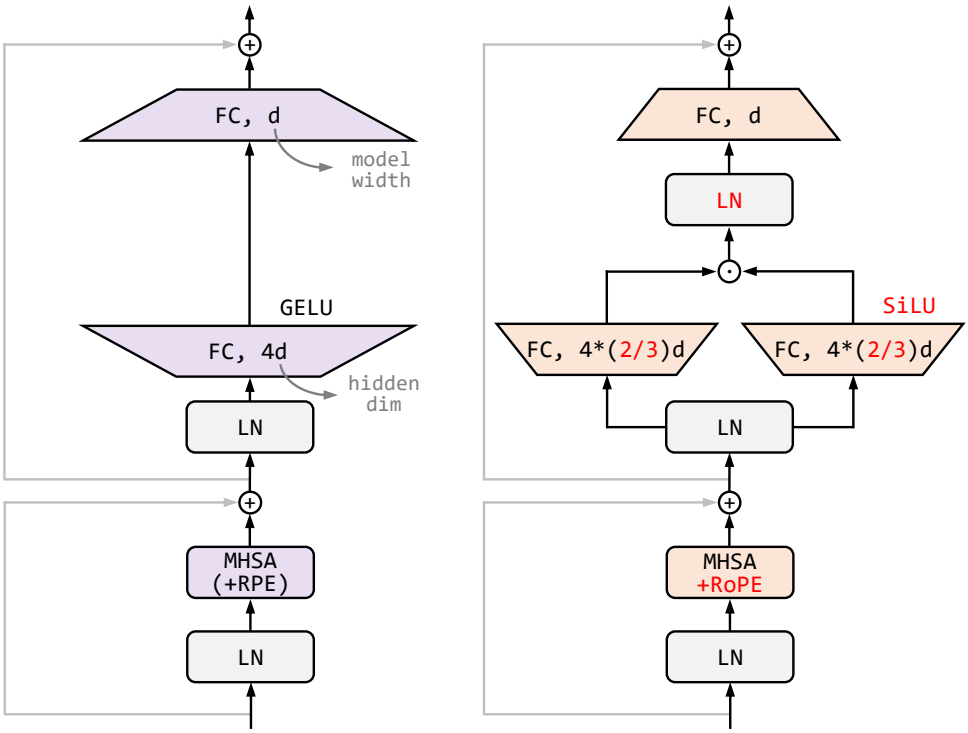


Figura 13 – A representação da arquitetura ViT tradicional (a) e, a representação da *Transform Vision* (TrV) (b). TrV é construído a partir do ViT, adicionando melhorias na arquitetura. Fonte: (FANG et al., 2024c)

A base do EVA-02 é um ViT padrão, que processa uma imagem dividindo-a em uma sequência de *patches* e os utiliza como entrada para uma série de blocos Transformer. As principais inovações do EVA-02 residem nas modificações internas desses blocos:

- **Rede *Feed-Forward* SwiGLU:** A rede *feed-forward* (FFN) tradicional do ViT, que utiliza uma ativação GELU, é substituída por uma variante da *Gated Linear Unit* (GLU) com a função de ativação SiLU (*Sigmoid Linear Unit*), conhecida como SwiGLU. Para manter a contagem de parâmetros e o custo computacional (FLOPs) consistentes com o modelo base, a dimensão oculta da FFN SwiGLU é reduzida em um fator de 2/3. Essa modificação, inspirada em modelos de linguagem de ponta, mostrou-se mais eficaz.

- **Normalização de Camada *sub-LN*:** O esquema de normalização de camada (*Layer Normalization*) é alterado do tradicional *pre-LN* para o *sub-LN*. Essa abordagem, onde a normalização é aplicada dentro do caminho residual, demonstrou um ligeiro ganho de desempenho e maior estabilidade durante o pré-treinamento.

- **Embaralhamento de Posição Rotativo 2D (RoPE):** Para injetar informação posicional, o EVA-02 abandona os embaralhamentos de posição relativos convencionais e adota o

2D Rotary Position Embedding (RoPE). O RoPE codifica a informação de posição absoluta e relativa de forma unificada, aplicando rotações nos vetores de consulta (*query*) e chave (*key*) na camada de atenção. Uma vantagem notável é sua compatibilidade nativa com operadores de atenção de alto desempenho, como o xFormers, ao contrário de métodos que modificam diretamente a matriz de atenção.

- **Inicialização de Pesos Xavier Normal:** Os experimentos indicaram que a eficácia da FFN SwiGLU estava condicionada a uma inicialização de pesos adequada. Por isso, foi adotada a inicialização Xavier normal para todas as matrizes de peso nos blocos TrV, o que se mostrou fundamental para a convergência e o desempenho do modelo.

Os resultados demonstram um desempenho superior em comparação com abordagens anteriores do estado da arte em diversas tarefas de visão, utilizando significativamente menos parâmetros e recursos computacionais. Notavelmente, empregando exclusivamente dados de treinamento de acesso público, o modelo EVA-02, com apenas 304 milhões de parâmetros, alcança uma acurácia de 90,0% no ajuste fino *top-1* no conjunto de validação ImageNet-1K. Adicionalmente, sua variante para tarefas de visão-linguagem, o EVA-02-CLIP, atinge 80,4% de acurácia *top-1* em avaliação *zero-shot* no mesmo conjunto de dados, superando modelos CLIP de código aberto maiores com aproximadamente um sexto dos parâmetros e dos dados de treinamento imagem-texto.

3.3.3 *EdgeNeXt*

A crescente demanda por aplicações de inteligência artificial em dispositivos com recursos computacionais limitados, como smartphones e sistemas embarcados, impulsionou a pesquisa por redes neurais eficientes. As arquiteturas de redes neurais convolucionais, embora predominantes em tarefas de visão computacional, possuem uma limitação inerente ao seu campo receptivo local, o que dificulta a modelagem de contextos globais da imagem. Por outro lado, os *Vision Transformers* (ViTs), que utilizam mecanismos de auto-atenção para capturar relações de longo alcance, geralmente incorrem em um custo computacional quadrático em relação ao tamanho da entrada, tornando-os inadequados para implantação em dispositivos de borda (MAAZ et al., 2022).

Nesse contexto, surgem as arquiteturas híbridas, que buscam unificar as vantagens de ambos os paradigmas. A arquitetura EdgeNeXt, proposta por Maaz *et al.* (MAAZ et al., 2022), representa um avanço significativo nessa direção, ao combinar de forma eficiente os pontos fortes das CNNs e dos *Transformers* para criar um modelo de propósito geral, leve e de alto desempenho para aplicações de visão móvel.

O principal objetivo do EdgeNeXt é superar o compromisso entre acurácia e eficiência, frequentemente observado em modelos estado-da-arte. Para tal, a arquitetura introduz um bloco codificador inovador denominado *Split Depth-wise Transpose Attention* (SDTA). Este componente divide os tensores de entrada em múltiplos grupos de canais e emprega convoluções *depth-wise* em conjunto com um mecanismo de *self-attention* transposta. A atenção é aplicada através da

dimensão dos canais, e não da dimensão espacial, o que reduz a complexidade computacional de quadrática para linear em relação ao número de *tokens* de entrada. Essa abordagem permite aumentar implicitamente o campo receptivo da rede e codificar características em múltiplas escalas de maneira eficiente, sem um aumento substancial no número de parâmetros ou operações de multiplicação-acumulação.

A arquitetura geral do EdgeNeXt,ilustrada na Figura 14, é organizada em quatro estágios hierárquicos que extraem características em diferentes escalas. A rede inicia com uma camada "patchify stem", implementada com uma convolução com *stride* 4, que reduz a resolução espacial da imagem de entrada e aumenta a dimensão dos canais. Os estágios subsequentes utilizam convoluções com *stride* 2 para realizar o *downsampling*. Cada estágio é composto por uma sequência de dois tipos de blocos: o Codificador Convolucional e o inovador codificador *Split Depth-wise Transpose Attention* (SDTA). Uma característica notável é a adoção de um mecanismo de tamanho de kernel adaptativo para os blocos convolucionais, utilizando kernels menores nos estágios iniciais (e.g., 3×3 e 5×5), adequados para características de baixo nível, e kernels progressivamente maiores nos estágios mais profundos (e.g., 7×7 e 9×9), para capturar características semânticas de alto nível de forma mais eficaz (MAAZ et al., 2022).

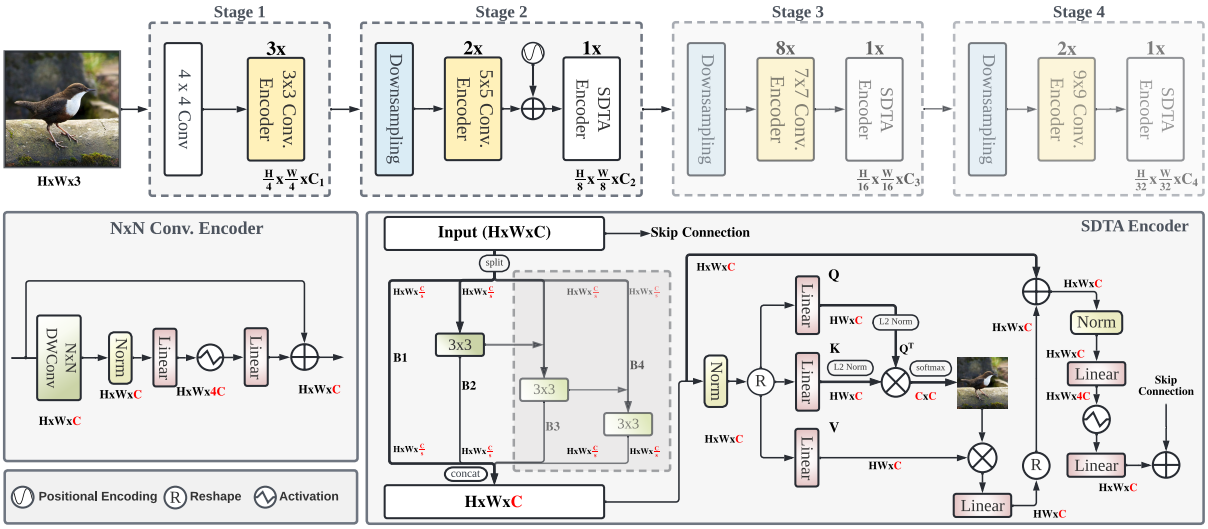


Figura 14 – Arquitetura do *EdgeNeXt*. Fonte: (MAAZ et al., 2022)

A principal inovação da arquitetura reside no bloco codificador SDTA, projetado para aprender eficientemente representações locais e globais. O SDTA opera em duas fases. Primeiramente, inspirado no Res2Net, ele divide o tensor de entrada em múltiplos grupos de canais para processamento em paralelo. Cada grupo, exceto o primeiro, passa por uma convolução *depth-wise* e é somado à saída do grupo anterior, criando uma representação hierárquica e multiescala dentro do mesmo bloco, o que expande o campo receptivo de forma flexível.

A segunda fase do bloco SDTA introduz o mecanismo de atenção global. Para evitar o alto custo computacional da auto-atenção padrão, o EdgeNeXt emprega a atenção transposta. Nessa abordagem, o produto de ponto para o cálculo da atenção não é realizado entre as dimensões espaciais, mas sim entre as dimensões de canais das projeções de Query (Q) e Key (K). Isso resulta em uma matriz de atenção de dimensionalidade $C \times C$ (onde C é o número de canais), reduzindo

a complexidade computacional de quadrática para linear em relação ao número de *tokens* de entrada. Essa operação de atenção, aplicada sobre os canais, permite que o modelo codifique implicitamente o contexto global da imagem com um custo computacional drasticamente menor (MAAZ et al., 2022).

Resultados experimentais em tarefas de classificação, detecção de objetos e segmentação semântica demonstram que o EdgeNeXt estabelece um novo padrão de desempenho para modelos leves, superando arquiteturas como MobileViT com uma redução significativa no custo computacional. Por exemplo, o modelo EdgeNeXt com 1.3 milhão de parâmetros alcança uma acurácia top-1 de 71.2% no ImageNet-1K, superando o MobileViT em 2.2% com 28% menos FLOPs, o que evidencia sua eficácia e adequação para aplicações de visão computacional em tempo real em dispositivos de borda (MAAZ et al., 2022).

3.4 Métricas para avaliação

As métricas de avaliação desempenham um papel crucial no aprendizado de máquina, servindo como ferramentas fundamentais para avaliar o desempenho e a eficácia dos modelos desenvolvidos. São utilizadas para avaliar a qualidade das previsões feitas por um modelo em comparação com os resultados reais, fornecendo uma base objetiva para avaliar a utilidade e a precisão dos modelos.

Existem diversas métricas de avaliação, cada uma adequada para diferentes problemas e objetivos específicos. Em problemas de classificação, as métricas mais utilizadas são a acurácia, a precisão, o *recall* (sensibilidade) e a métrica F1, que combina precisão e *recall* para fornecer uma visão mais equilibrada do desempenho do modelo (FANG et al., 2024b).

A escolha da métrica de avaliação adequada está relacionada ao contexto do problema e aos objetivos específicos do modelo. Em aplicações médicas, pode ser mais relevante minimizar falsos negativos, enquanto, em sistemas de recomendação, maximizar a precisão pode ser relevante.

Acurácia

A acurácia é uma das métricas de avaliação mais utilizadas em problemas de aprendizado de máquina, especialmente em tarefas de classificação. Ela mede a proporção de previsões corretas feitas por um modelo em relação ao total de previsões, oferecendo uma visão geral da eficácia do modelo.

A acurácia é calculada como a soma dos verdadeiros positivos e verdadeiros negativos dividida pelo total de amostras avaliadas (FAWCETT, 2006). Matematicamente, a acurácia é expressa pela fórmula:

$$\text{Acurácia} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}}, \tag{3.1}$$

em que:

- VP (Verdadeiros Positivos) são os casos corretamente previstos como positivos.

- VN (Verdadeiros Negativos) são os casos corretamente previstos como negativos.
- FP (Falsos Positivos) são os casos incorretamente previstos como positivos.
- FN (Falsos Negativos) são os casos incorretamente previstos como negativos.

A simplicidade da acurácia a torna uma métrica atrativa, permitindo uma avaliação rápida e fácil do desempenho do modelo. No entanto, ela pode ser enganosa em cenários com classes desbalanceadas.

Por exemplo, em um conjunto de dados em que 95% dos exemplos pertencem a uma única classe, um modelo que simplesmente prevê essa classe para todas as instâncias alcançará uma alta acurácia de 95%, apesar de não identificar corretamente nenhuma instância da classe minoritária.

Por essa razão, a acurácia deve ser complementada por outras métricas, como precisão, *recall* e F1, para obter uma visão mais completa do desempenho do modelo, especialmente em aplicações em que a distribuição das classes é desigual ou nos casos em que os custos de diferentes tipos de erros variam significativamente.

Precisão

A precisão é uma métrica fundamental em problemas de classificação no aprendizado de máquina, particularmente relevante quando o custo de falsos positivos é alto. A precisão mede a proporção de previsões corretas de uma classe positiva em relação ao total de previsões feitas para essa classe (FACELI et al., 2021).

Em outras palavras, a precisão indica quantas das previsões positivas do modelo são realmente positivas (SOKOLOVA; LAPALME, 2009). Matematicamente, a precisão é expressa pela fórmula:

$$\text{Precisão} = \frac{\text{VP}}{\text{VP} + \text{FP}}, \tag{3.2}$$

Recall

O *recall*, também conhecido como sensibilidade e taxa de verdadeiros positivos (FACELI et al., 2021), é uma métrica essencial em problemas de classificação no aprendizado de máquina, especialmente quando é crucial identificar corretamente todos os exemplos de uma classe positiva (SOKOLOVA; LAPALME, 2009).

O *recall* mede a proporção de exemplos da classe positiva que foram corretamente identificados pelo modelo em relação ao total de exemplos reais dessa classe. Matematicamente, o recall é expresso pela fórmula:

$$\text{Recall} = \frac{\text{VP}}{\text{VP} + \text{FN}}, \tag{3.3}$$

F1-Score

O F1 é uma métrica comum em problemas de classificação no aprendizado de máquina, projetada para fornecer uma medida balanceada entre precisão e *recall*. Essa métrica é especialmente útil quando há um desequilíbrio entre as classes ou quando ambas precisão e *recall* são igualmente importantes para o problema em questão (SOKOLOVA; LAPALME, 2009).

O F1 é calculado como a média harmônica da precisão e do *recall*. Matematicamente, é expresso pela fórmula:

$$F1 = 2 \times \frac{\text{Precisão} \times \text{Recall}}{\text{Precisão} + \text{Recall}}. \tag{3.4}$$

3.5 Teste de Wilcoxon para Amostras Pareadas

O *Teste de Wilcoxon para Amostras Pareadas*, também conhecido como *Signed-Rank Test*, é um teste estatístico não paramétrico utilizado para comparar duas amostras relacionadas ou medições repetidas realizadas sobre o mesmo grupo de sujeitos. Diferentemente do teste t pareado, que assume a normalidade das diferenças entre os pares, o teste de Wilcoxon não impõe tal restrição, sendo apropriado para dados que não seguem uma distribuição normal ou que contêm outliers (BRIDGE; SAWILOWSKY, 1999).

3.5.1 Objetivo e Hipóteses

O objetivo do teste é verificar se a mediana das diferenças entre os pares de observações é significativamente diferente de zero. A hipótese nula (H_0) postula que não há diferença entre os grupos, ou seja, a mediana das diferenças é igual a zero. A hipótese alternativa (H_1) afirma que existe uma diferença significativa, indicando que uma das condições tende a produzir valores sistematicamente maiores ou menores que a outra.

3.5.2 Condições de Aplicação

As principais condições para aplicação do teste de Wilcoxon são:

- Os dados devem ser pareados, ou seja, cada observação de um grupo deve ter uma correspondente no outro grupo.
- As diferenças entre os pares devem ser independentes entre si.
- As diferenças devem ser medidas em escala ordinal, intervalar ou de razão.
- A distribuição das diferenças deve ser aproximadamente simétrica em torno da mediana.

3.5.3 Procedimento do Teste

O teste segue os seguintes passos:

1. Calcular a diferença d_i entre cada par de observações $X_i - Y_i$.
2. Ignorar os pares em que $d_i = 0$.
3. Calcular os valores absolutos $|d_i|$ e ordenar em ordem crescente.
4. Atribuir postos aos $|d_i|$, ajustando para empates quando necessário.
5. Reatribuir os sinais originais aos postos conforme o sinal de d_i.
6. Calcular a soma dos postos positivos (T^+) e negativos (T^-).
7. A estatística de teste W é definida como $W = \min(T^+, T^-)$.

3.5.4 Exemplo Ilustrativo

Considere os seguintes pares de observações dos modelos A e B:

$$\text{Modelo A} = [9,\ 14,\ 7,\ 16,\ 8], \quad \text{Modelo B} = [10,\ 12,\ 9,\ 15,\ 10]$$

1. Diferenças: $d = [1,\ -2,\ 2,\ -1,\ 2]$
2. Valores absolutos: $|d| = [1,\ 2,\ 2,\ 1,\ 2]$
3. Ranks (com ajustes para empates):
   - $|1| \rightarrow$ Rank médio 1.5
   - $|2| \rightarrow$ Rank médio 4.0
4. Aplicando sinais:
   - Positivos: $+1 \rightarrow 1.5,\ +2 \rightarrow 4.0,\ +2 \rightarrow 4.0$
   - Negativos: $-2 \rightarrow 4.0,\ -1 \rightarrow 1.5$
5. Somas: $T^+ = 9.5,\ T^- = 5.5,\ W = \min(9.5, 5.5) = 5.5$

Esse valor é então comparado a uma distribuição crítica (tabelada ou aproximada por normal para grandes amostras) para verificar se W é suficientemente pequeno para rejeitar H_0.

3.5.5 Aplicação em Classificadores de Imagens Citológicas

No contexto desta dissertação, o teste de Wilcoxon foi aplicado para comparar o desempenho da abordagem proposta com modelos de referência em diferentes configurações experimentais. Cada par de valores foi formado pelas métricas obtidas com validação cruzada em cinco *folds*, gerando vetores pareados de acurácia, precisão, *recall* e F1-score para cada modelo avaliado.

O uso desse teste permitiu verificar, com nível de confiança de 99%, que as diferenças observadas entre os modelos são estatisticamente significativas, conferindo maior robustez às conclusões desta pesquisa.

4 Abordagem Proposta

O presente estudo tem como objetivo desenvolver uma abordagem para a classificacão de células cervicais que seja competitiva em relação ao estado da arte. Para isso, foi utilizado o conjunto de imagens Herlev (JANTZEN et al., 2005) e SIPaKMeD (PLISSITI et al., 2018), conforme descrito na Seção 3.1, aliado às ferramentas metodológicas apresentadas no Capítulo 3.

As revisões sistemáticas (JIANG et al., 2023; FANG et al., 2024b; KHARE et al., 2024; WUBINEH; RUSIECKI; HALAWA, 2024b; VARGAS-CARDONA et al., 2024; MATHEW; CHERIYAN, 2024; WU et al., 2024), bem como a análise crítica desenvolvida no Capítulo 2, evidenciam uma tendência crescente no uso de abordagens híbridas, que combinam redes neurais convolucionais (CNNs) com arquiteturas baseadas em *Vision Transformer* (ViT), além da adoção de estratégias baseadas em *ensemble* de modelos, as quais têm se destacado por seu desempenho superior em tarefas de classificação em imagens citológicas.

Neste trabalho, optou-se pelo uso de um *ensemble* de modelos que utilizam redes neurais convolucionais e *Vision Transformer*, empregando redes pré-treinadas e a técnica conhecida como *Transfer Learning*. Foram adotadas três arquiteturas amplamente reconhecidas: *EfficientVit*, EVA02, *EdgeNeXt*. A seleção dessas arquiteturas justifica-se por seus resultados de ponta na ImageNet, suas inovações arquitetônicas e a quantidade relativamente pequena de parâmetros. A hipótese central é que a combinação desses modelos em um ensemble resultará em uma capacidade de classificação superior, o que é posteriormente validado no Capítulo 5.

Os resultados indicam que a abordagem proposta obteve acurácia, precisão, *recall* e *F1-score* médias de $99,73\%$, na tarefa de classificação binária (normal e anormal) conforme a Tabela 14 , utilizando o conjunto de dados SIPaKMeD. Já para o conjunto de dados Herlev, a metodologia proposta obteve acurácia, *recall* e *F1-score* médias de $98,35\%$ e $98,37\%$ para precisão para classificacão binária. Indicando que a abordagem proposta é competitiva em relação ao estado da arte.

A Seção 4.1 apresenta o fluxo detalhado dos experimentos conduzidos, abrangendo as etapas de divisão do conjunto de imagens, aumento de dados, validação interna, otimização de hiperparâmetros e classificação. Para a divisão do conjunto de treinamento e teste, foi empregada a técnica de validação cruzada com cinco *folds*, em que quatro subconjuntos foram utilizados para treinamento com validação interna e um subconjunto foi reservado para teste.

Na etapa de aumento de dados, buscou-se expandir o volume de imagens e promover o balanceamento entre as classes, com o objetivo de mitigar o viés do modelo em relação à classe majoritária durante o processo de aprendizado. Por fim, na etapa de classificação, aplicaram-se as técnicas descritas nas seções metodológicas, cujos desempenhos obtidos estão apresentados e discutidos no Capítulo 5.

4.1 Etapas dos Experimentos Realizados

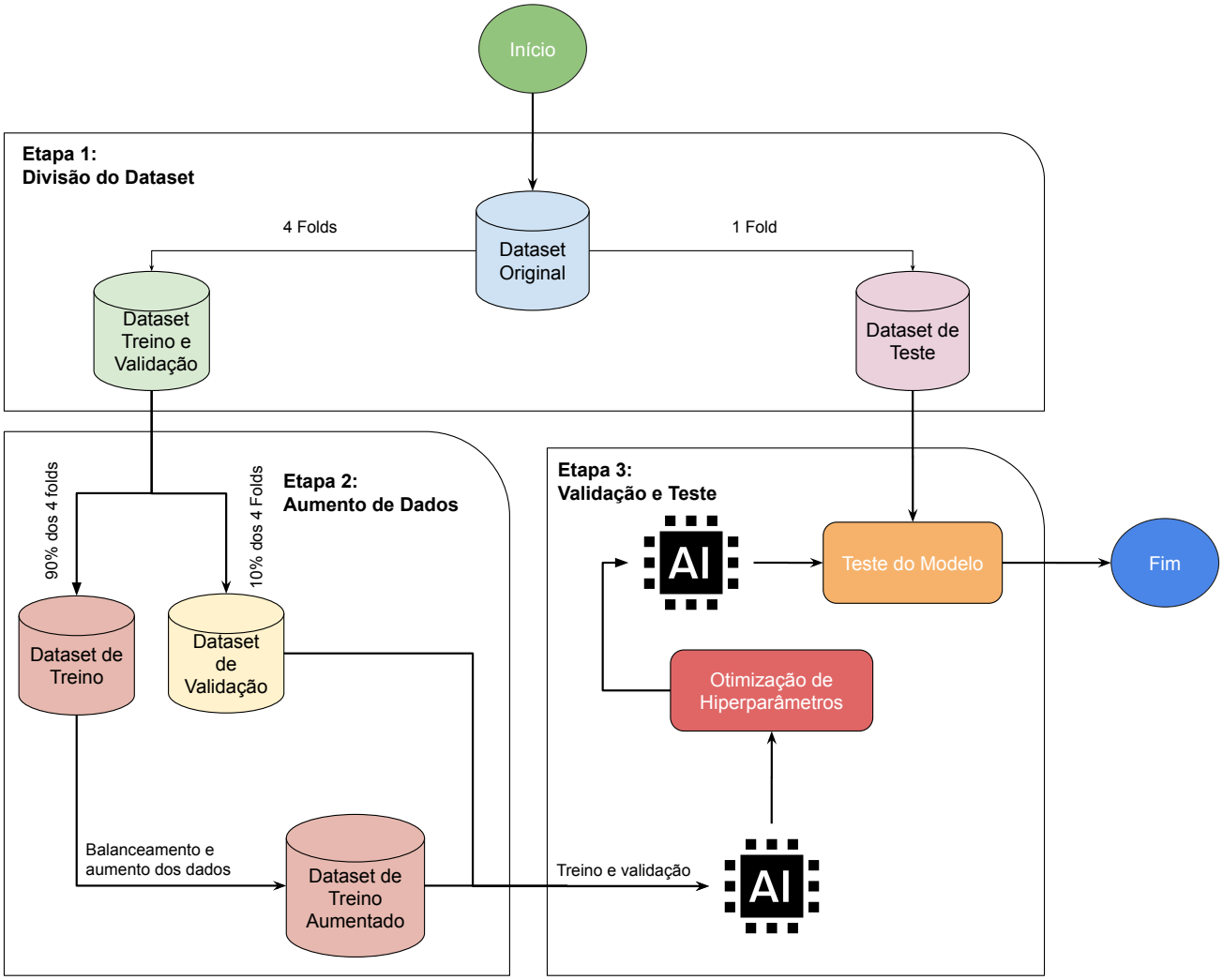


Figura 15 – Fluxo dos experimentos realizados. Fonte: Elaborado pelo autor.

Para a realização dos experimentos, houve três etapas primordiais:

1. **Divisão do Dataset:** A divisão das imagens foi realizada por meio da técnica de validação cruzada com cinco *folds*, contemplando os subconjuntos de treinamento, validação e teste. Quatro *folds* foram empregados no treinamento com validação interna, enquanto o quinto foi reservado para teste. A separação entre os subconjuntos de treinamento e validação foi conduzida pela estratégia *Holdout*, com partição estratificada de 90% para treinamento e 10% para validação, assegurando a preservação da distribuição das classes em ambas as divisões.

2. **Aumento de Dados:** Foram aplicadas transformações sobre as imagens do subconjunto de treinamento, com o objetivo de balancear as classes, reduzir o risco de sobreajuste e mitigar o viés do modelo em relação às classes majoritárias.

3. **Validação e Teste:** Por fim, o modelo foi treinado com os subconjuntos de treinamento e validado com os respectivos dados, sendo posteriormente avaliado por meio da classificação das imagens pertencentes ao subconjunto de teste.

A Figura 15 apresenta o fluxo das três etapas elencadas acima, realizadas nos experimentos da abordagem proposta.

4.2 Aumento de Dados

Com o objetivo de reduzir o desbalanceamento entre as classes e mitigar o *overfitting* dos modelos, foi adotada uma estratégia de aumento de dados durante a fase de pré-processamento. As seguintes técnicas foram aplicadas:

- **Imagens aumentadas:** *flip* horizontal e vertical, rotação aleatória, aplicação de *CLAHE*, filtro gaussiano, aumento de nitidez, além de ajustes de brilho, contraste, saturação e matiz.

A Tabela 8 apresenta a quantidade de imagens geradas por meio do aumento de dados, bem como o total de imagens disponíveis no conjunto de dados para cada experimento após a aplicação dessas técnicas.

Tabela 8 – Número de imagens geradas no aumento de dados para cada cenário.

Conjunto de Dados	Configuração	Imagens Adicionadas	Total de Imagens
Herlev	7 classes	462	1.379
	2 classes	433	1.350
SIPaKMeD	5 classes	106	4.155
	3 classes	865	4.914
	2 classes	813	4.862

4.3 Experimentos e Hiperparâmetros dos Modelos

A abordagem proposta neste trabalho fundamenta-se na utilização combinada de três modelos pré-treinados, previamente adaptados aos conjuntos de imagens citológicas. Os hiperparâmetros empregados no treinamento de cada modelo estão detalhados na Tabela 9. A decisão final do *ensemble* é obtida por meio da média das probabilidades fornecidas por cada modelo.

Para mitigar o sobreajuste, foi implementada uma estratégia de *early stopping* com *patience* de 10 épocas, adotando como critério de parada uma variação mínima de $1 \cdot 10^{-5}$ na função de *loss* de validação. Após o treinamento, os modelos foram combinados em um esquema de *ensemble* para a classificação do conjunto de testes, permitindo a comparação entre os resultados individuais e aqueles obtidos pela combinação dos modelos.

Tabela 9 – Hiperparâmetros utilizados para o treinamento e parâmetros dos modelos utilizados para o treinamento.

Parâmetros e hiperparâmetros	**EfficientViT_b2**	**EdgeNeXt_base**	**EVA-02_small**
Tamanho do lote	32	32	32
Otimizador	AdamW	AdamW	AdamW
Taxa de aprendizado inicial	0.0001	0.0001	0.0001
Épocas máximas	50	50	50
Épocas mínimas	20	20	20
Dimensões da imagem de entrada	(336x336x3)	(336x336x3)	(336x336x3)
Quantidade de parâmetros do modelo	21,8 M	17,9 M	21,7 M
GMACs	1,6	3,8	15,5

Nos experimentos, foram realizadas quatro rodadas de validação cruzada 5-fold (totalizando 20 execuções) para cada modelo, em ambos os conjuntos de dados, contemplando tarefas de classificação binária e multiclasse. Essa abordagem visa garantir robustez estatística, especialmente em conjuntos pequenos como o Herlev (917 imagens). Em cada rodada, os dados foram particionados aleatoriamente, assegurando variação nas imagens selecionadas para os *folds*, sendo quatro utilizados para treinamento e um para teste. Ao final de cada execução, foram computadas as métricas de acurácia, precisão, *recall* e *F1-score*. Nos conjuntos de treinamento, 90% dos exemplos foram utilizados para o treinamento efetivo, enquanto os 10% restantes foram reservados para validação durante o ajuste dos modelos.

Para verificar diferenças estatísticas entre os modelos, foi empregado o teste *Wilcoxon Signed-Rank* (WSR), apropriado para dados pareados e que não exigem distribuição normal. Segundo Bridge e Sawilowsky (BRIDGE; SAWILOWSKY, 1999), testes não paramétricos como o de *Wilcoxon* apresentam vantagens em relação ao teste t de *Student*, especialmente em contextos com dados assimétricos, caudas pesadas ou amostras pequenas, oferecendo maior poder estatístico nessas situações.

5 Resultados e Discussão

Neste capítulo são apresentados e analisados os resultados obtidos a partir dos experimentos conduzidos nos conjuntos de dados *Herlev* e *SIPaKMeD*, com o intuito de avaliar o desempenho da abordagem proposta frente a diferentes configurações de classificação. As experiências foram organizadas de forma a considerar tanto cenários mais complexos, com múltiplas classes, quanto configurações simplificadas de classificação binária, a fim de verificar a robustez e a capacidade discriminativa do modelo em distintas condições experimentais.

No conjunto *Herlev*, foram avaliadas duas configurações distintas: a classificação em sete classes, que representa a configuração mais detalhada das categorias citológicas conforme a taxonomia adotada pelo conjunto, e a classificação binária, que agrupa as classes em categorias mais amplas (normal versus anormal).

Para o conjunto *SIPaKMeD*, os experimentos envolveram três diferentes configurações de classificação: cinco classes (configuração original do conjunto, mantendo todas as categorias citológicas), três classes (com agrupamento de algumas categorias similares) e duas classes (normal versus anormal). Essa diversidade de cenários experimentais visa refletir diferentes perspectivas de aplicação clínica e diagnóstica, permitindo uma avaliação mais abrangente do potencial do modelo proposto.

Os resultados obtidos em cada uma dessas configurações são discutidos nas seções subsequentes, considerando métricas de desempenho como acurácia, precisão, *recall* e F1-score, possibilitando uma análise crítica dos pontos fortes e limitações da abordagem implementada. Na Seção 5.3 são analisados os resultados para cada experimento, discutidos e comparados com a literatura.

5.1 Resultados no Conjunto de Dados *Herlev*

Os experimentos conduzidos com o conjunto de dados *Herlev* foram organizados em duas configurações distintas de classificação, com o objetivo de avaliar a versatilidade e a eficácia da abordagem proposta em diferentes cenários diagnósticos. A primeira configuração foi baseada em uma abordagem de classificação binária, agrupando as sete classes originais em duas categorias principais: normais e anormais. Essa simplificação busca simular um cenário mais próximo da aplicação clínica, em que o foco primário é a triagem de casos suspeitos, permitindo uma análise mais direta da capacidade do modelo em identificar alterações celulares relevantes do ponto de vista diagnóstico.

A segunda configuração consistiu na classificação em sete classes, refletindo a configuração detalhada das categorias citológicas conforme os padrões estabelecidos no conjunto. Essa configuração representa um cenário mais desafiador, uma vez que demanda maior poder discriminativo do modelo para distinguir sutilezas entre as diferentes classes celulares.

As Tabelas 10 e 11 apresentam de forma detalhada os resultados obtidos em ambas as configurações utilizando-se a validação cruzada. Para que seja possível comparar com outras abordagens da literatura, utilizou-se do *holdout* como será mostrada nas próximas subseções.

Nessas tabelas são comparados o desempenho da abordagem proposta e dos modelos utilizados como base para sua construção, considerando métricas como acurácia, precisão, *recall* e F1-score. Os melhores resultados estão destacados, evidenciando a contribuição individual de cada modelo e o impacto positivo da estratégia de *ensemble* na melhoria geral do desempenho do sistema de classificação.

Tabela 10 – Desempenho no conjunto *Herlev* com 2 classes para validação cruzada.

Métrica	**Abordagem Proposta**	***EfficientViT_b2***	***EVA02_small***	***EdgeNeXt_base***
Acurácia	**98,35±0,61%**	97,41±0,74%	96,71± 1,07%	97,91±0,66%
Precisão	**98,37±0,61%**	97,42±0,73%	96,76±1,05%	97,94±0,64%
Recall	**98,35±0,61%**	97,41±0,74%	96,71±1,07%	97,91±0,66%
F1-Score	**98,35±0,61%**	97,41±0,74%	96,71±1,07%	97,91±0,66%

Tabela 11 – Desempenho no conjunto *Herlev* com 7 classes para validação cruzada..

Métrica	**Abordagem Proposta**	***EfficientViT_b2***	***EVA02_small***	***EdgeNeXt_base***
Acurácia	**83,40±1,59%**	80,90±1,76%	78,31±2,29%	80,42±2,05%
Precisão	**83,59±1,60%**	81,17±1,90%	78,94±2,27%	80,54±2,03%
Recall	**83,42±1,58%**	80,91±1,76%	78,33±2,31%	80,43±2,04%
F1-Score	**83,34±1,62%**	80,80±1,87%	78,24±2,21%	80,21±2,07%

Resultados utilizando o *holdout*

Para o treinamento, validação e teste da abordagem proposta o conjunto de dados foi dividido na proporção de 70%, 20% e 10% respectivamente, de modo estratificado. Em seguida, treinou-se cada modelo com uma com um dropout de 50% na última camada para mitigar o *overfitting*. Nas Subseções 5.1.1 e 5.1.2 serão apresentados os gráficos de *loss* do treinamento e validação de cada modelo e suas respectivas matrizes de confusões.

5.1.1 *Herlev* - 2 classes

A Tabela 12 apresenta a comparação dos resultados obtidos pelos modelos individuais e pela abordagem baseada em *ensemble*. Verifica-se que todos os modelos alcançaram desempenho superior a 90% em todas as métricas avaliadas, evidenciando elevada capacidade discriminativa e consistente poder de generalização na tarefa de classificação binária do conjunto *Herlev*.

O modelo EdgeNeXt apresentou o melhor desempenho entre as arquiteturas testadas, seguido pela abordagem proposta (*ensemble*) e pelo modelo EVA. Por outro lado, o Efficient-ViT apresentou o desempenho mais baixo, embora ainda mantendo resultados satisfatórios e consistentes.

A Figura 16 ilustra a matriz de confusão correspondente à abordagem proposta, evidenciando sua capacidade de distinguir adequadamente entre as duas classes avaliadas. As Figuras 17 e 18 apresentam, respectivamente, a matriz de confusão e o gráfico da função de perda do modelo EfficientViT, permitindo observar a estabilidade de seu processo de treinamento. De forma análoga, as Figuras 19 e 20 exibem os resultados do modelo EVA, enquanto as Figuras 21 e 22 ilustram o desempenho do modelo EdgeNeXt, que se destacou entre as arquiteturas avaliadas. Essa análise visual complementa a comparação quantitativa apresentada na tabela, permitindo uma compreensão mais abrangente do comportamento dos modelos.

Tabela 12 – Desempenho dos modelos no conjunto Herlev com 2 classes utilizando a validação holdout.

Métrica	**Abordagem Proposta**	***EfficientViT_b2***	***EVA02_small***	***EdgeNeXt_base***
Acurácia	97,83%	94,57%	97,83%	**98,91%**
Precisão	96,15%	91,82%	96,15%	**98,00%**
Recall	98,53%	94,98%	98,53%	**99,26%**
F1-Score	97,25%	93,22%	97,25%	**98,60%**

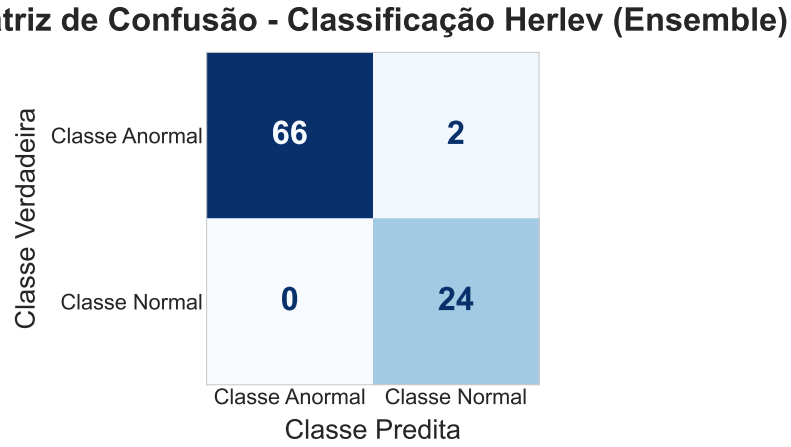


Figura 16 – Matriz de confusão da abordagem proposta para o conjunto Herlev classificado de modo binário (células anormais vs normais), o conjunto de teste foi composto de 92 imagens. Fonte: Elaboração do autor.

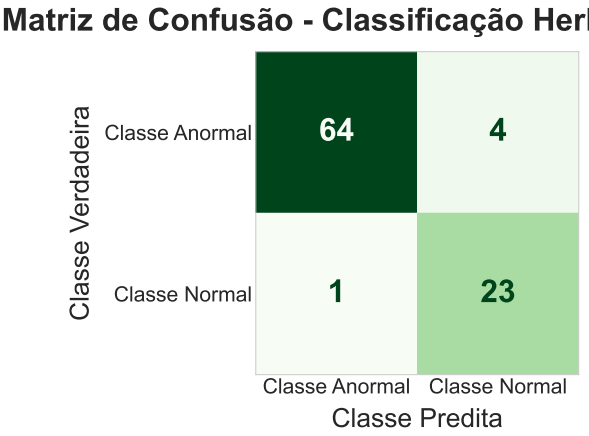


Figura 17 – Matriz de confusão do modelo EfficientViT para o conjunto Herlev classificado de modo binário (células anormais vs normais), o conjunto de teste foi composto de 92 imagens. Fonte: Elaboração do autor.

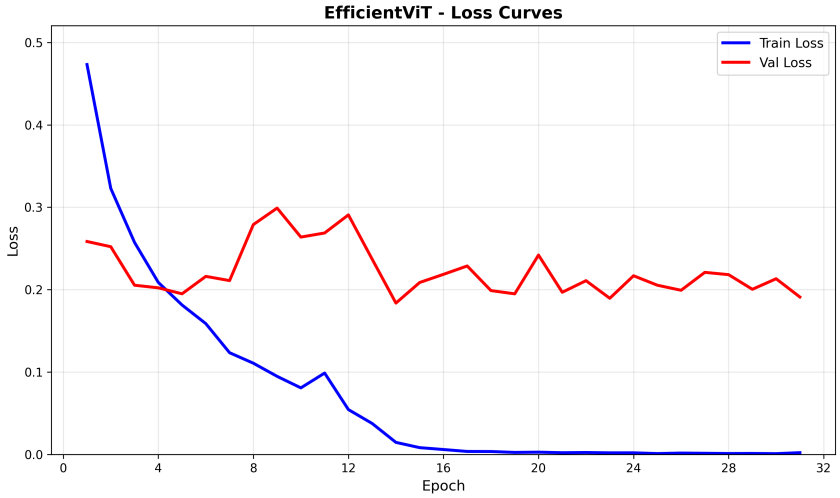


Figura 18 – *Loss* de treinamento e validação do modelo EfficientViT para o conjunto Herlev contendo duas classes. Fonte: Elaborado pelo autor.

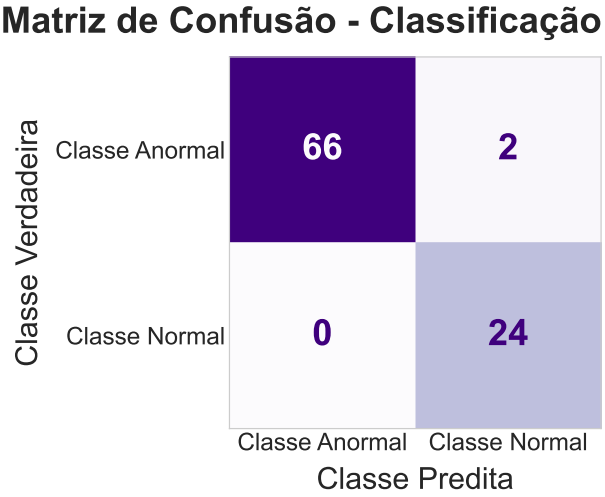


Figura 19 – Matriz de confusão do modelo EVA para o conjunto Herlev classificado de modo binário (células anormais vs normais), o conjunto de teste foi composto de 92 imagens. Fonte: Elaboração do autor.

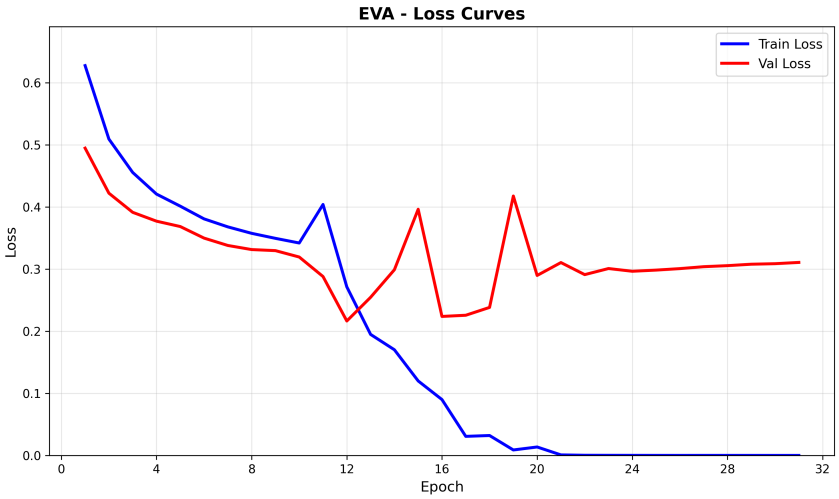


Figura 20 – *Loss* de treinamento e validação do modelo EVA para o conjunto Herlev contendo duas classes. Fonte: Elaborado pelo autor.

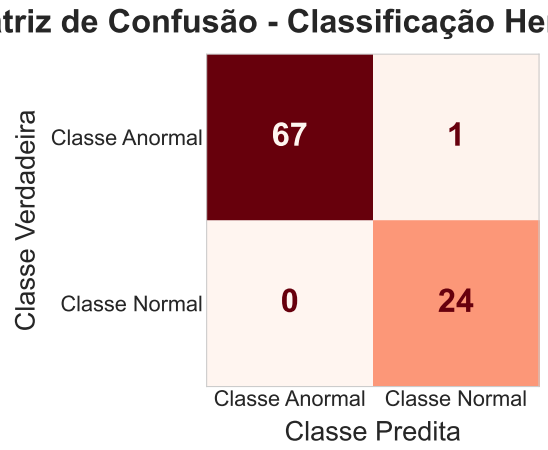


Figura 21 – Matriz de confusão do modelo EdgeNeXt para o conjunto Herlev classificado de modo binário (células anormais vs normais), o conjunto de teste foi composto de 92 imagens. Fonte: Elaboração do autor.

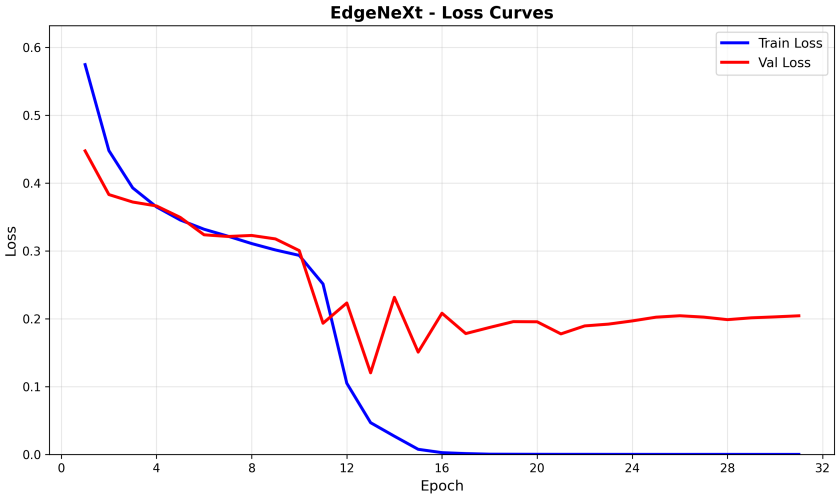


Figura 22 – *Loss* de treinamento e validação do modelo EdgeNeXt para o conjunto Herlev contendo duas classes. Fonte: Elaborado pelo autor.

5.1.2 *Herlev* - 7 classes

A Tabela 13 apresenta a comparação dos resultados obtidos pelos modelos individuais e pela abordagem baseada em *ensemble* na tarefa de classificação multiclasse do conjunto *Herlev*. Observa-se que, diferentemente do cenário binário, o aumento da complexidade da tarefa, envolvendo sete categorias distintas, resultou em uma redução geral do desempenho entre os modelos avaliados.

Ainda assim, a abordagem proposta por meio do *ensemble* apresentou o melhor desempenho entre todas as arquiteturas, alcançando acurácia de aproximadamente 79% e valores superiores a 80% nas demais métricas de avaliação. Os modelos individuais, por outro lado, apresentaram desempenho inferior, com acurácia e métricas complementares abaixo de 80%, indicando maior dificuldade na distinção entre as diferentes classes celulares.

A Figura 23 ilustra a matriz de confusão referente à abordagem proposta, evidenciando que, apesar da complexidade do problema, o modelo conseguiu manter boa consistência nas predições entre as classes majoritárias.

As Figuras 24 e 25 apresentam, respectivamente, a matriz de confusão e o gráfico da função de perda do modelo EfficientViT, enquanto as Figuras 26 e 27 exibem os resultados do modelo EVA. Por fim, as Figuras 28 e 29 ilustram o desempenho do modelo EdgeNeXt, que, embora inferior ao *ensemble*, manteve resultados próximos nas métricas de precisão e F1-score.

Essa análise conjunta evidencia a superioridade da abordagem proposta frente aos modelos individuais e demonstra sua capacidade de generalização mesmo em cenários de maior complexidade de classificação, ainda que haja espaço para aprimoramento na diferenciação de classes com características morfológicas mais sutis.

Tabela 13 – Desempenho dos modelos no conjunto Herlev com 7 classes utilizando a validação holdout.

Métrica	**Abordagem Proposta**	***EfficientViT_b2***	***EVA02_small***	***EdgeNeXt_base***
Acurácia	**79,35%**	70,65%	71,74%	76,09%
Precisão	**84,01%**	76,83%	79,42%	79,53%
Recall	**82,22%**	72,88%	76,90%	79,07%
F1-Score	**82,56%**	73,83%	77,70%	78,59%

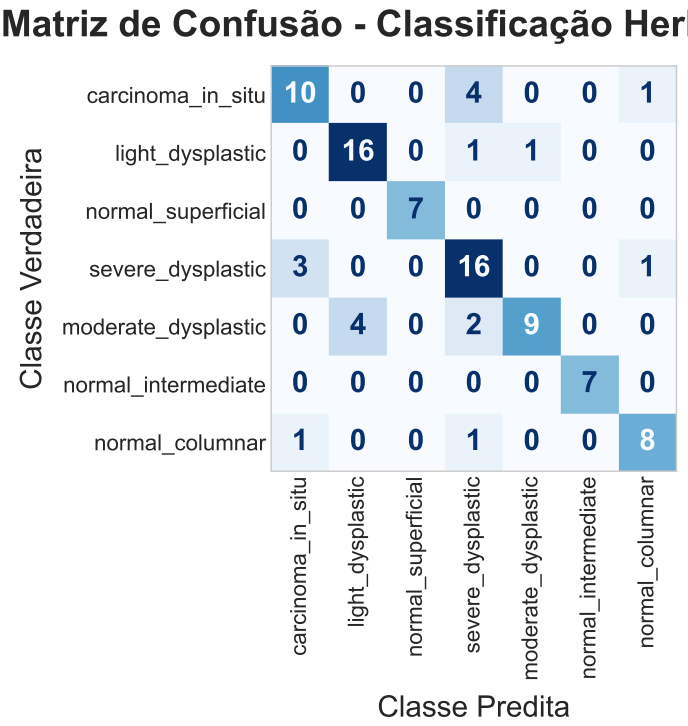


Figura 23 – Matriz de confusão da abordagem proposta para o conjunto Herlev classificado com as sete classes, o conjunto de teste foi composto de 92 imagens. Fonte: Elaboração do autor.

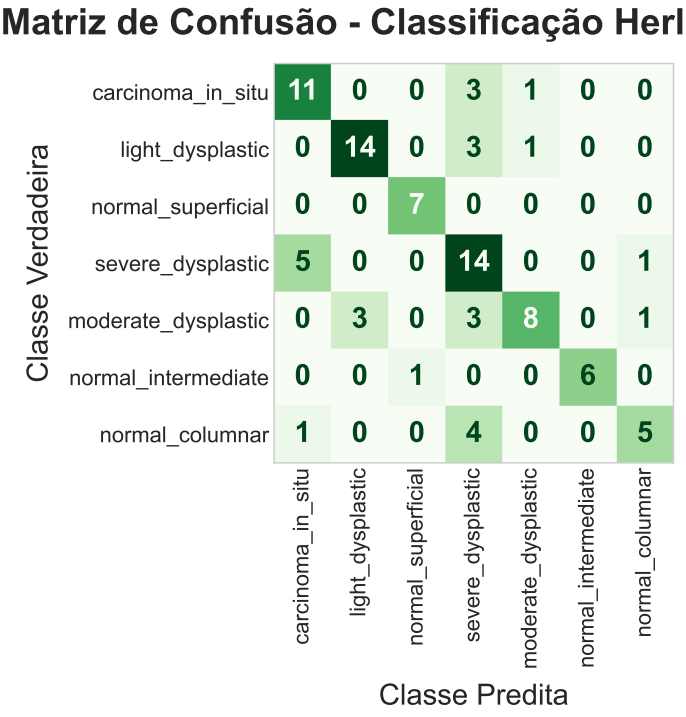


Figura 24 – Matriz de confusão do modelo EfficientViT para o conjunto Herlev classificado com as sete classes o conjunto de teste foi composto de 92 imagens. Fonte: Elaboração do autor.

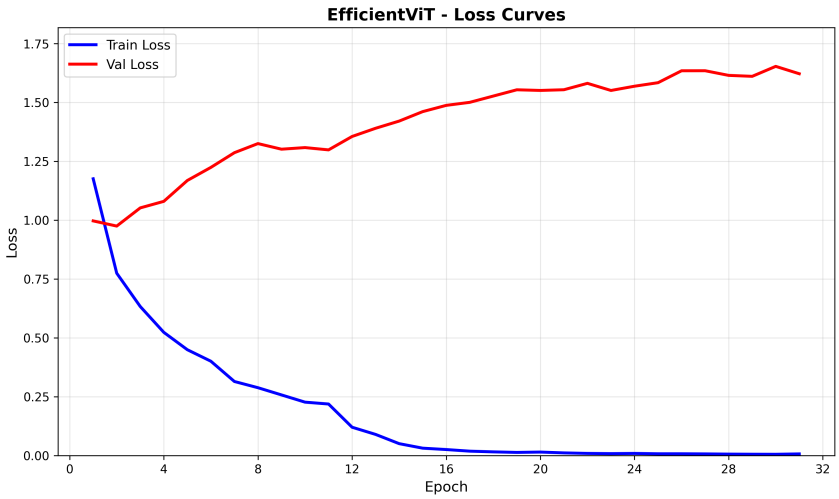


Figura 25 – *Loss* de treinamento e validação do modelo EfficientViT para o conjunto Herlev contendo sete classes. Fonte: Elaborado pelo autor.

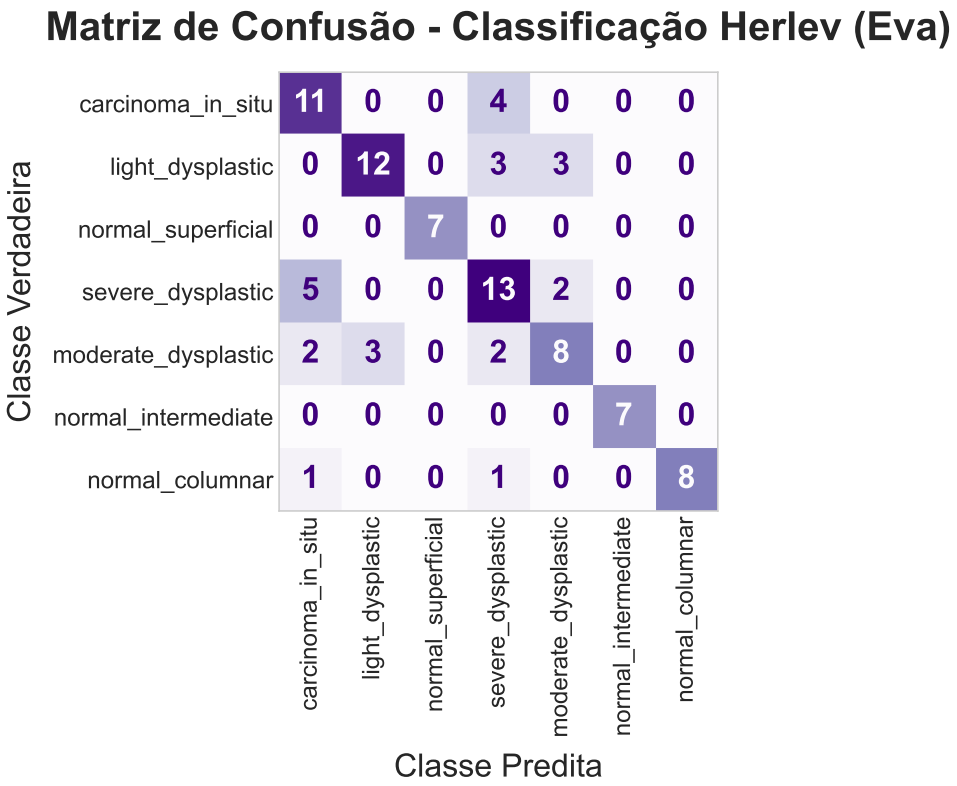


Figura 26 – Matriz de confusão do modelo EVA para o conjunto Herlev classificado de modo binário com as sete classes, o conjunto de teste foi composto de 92 imagens. Fonte: Elaboração do autor.

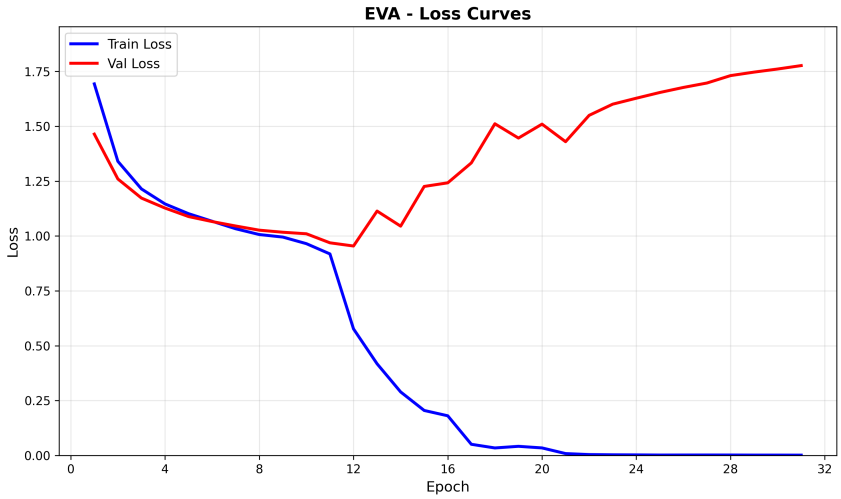


Figura 27 – *Loss* de treinamento e validação do modelo EVA para o conjunto Herlev contendo sete classes. Fonte: Elaborado pelo autor.

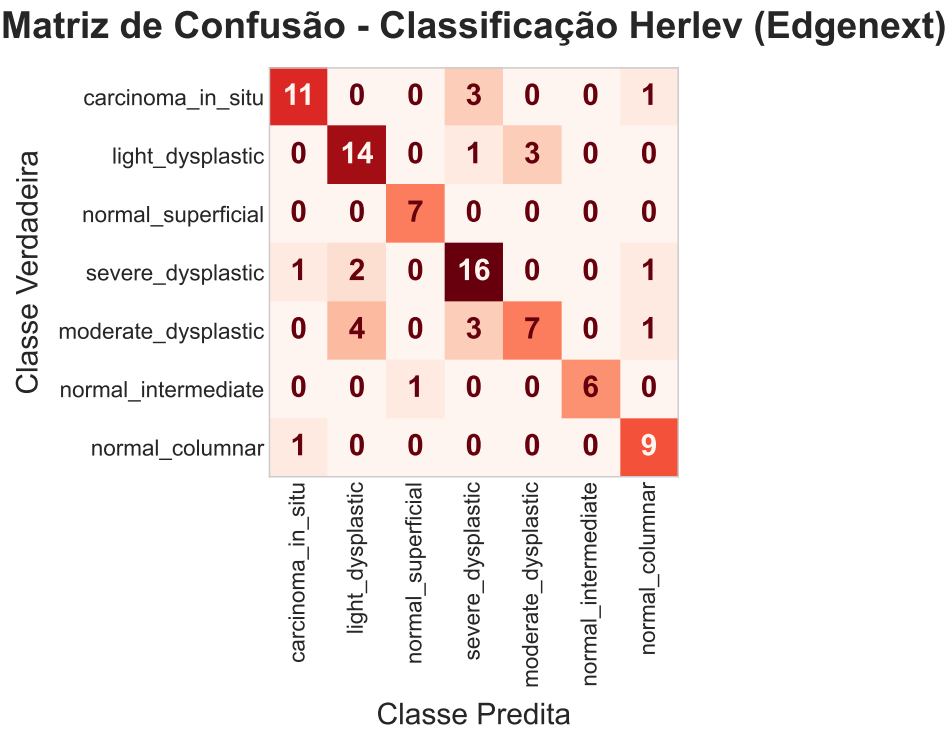


Figura 28 – Matriz de confusão do modelo EdgeNeXt para o conjunto Herlev classificado de modo binário com as sete classes, o conjunto de teste foi composto de 92 imagens. Fonte: Elaboração do autor.

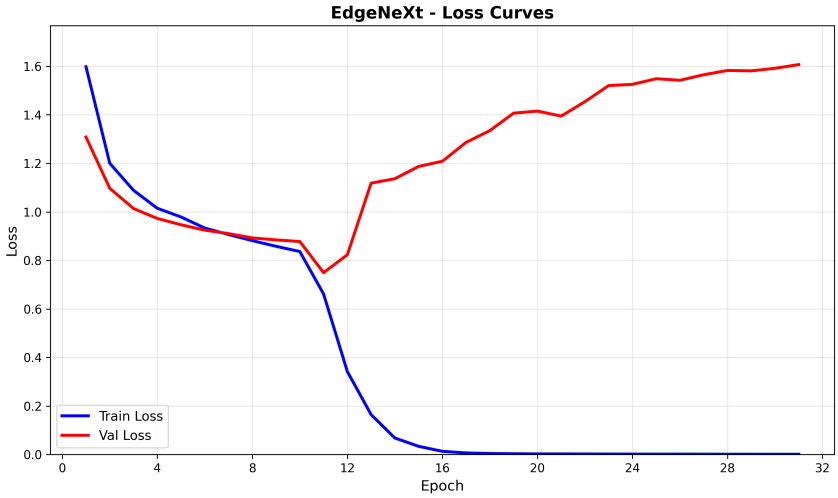


Figura 29 – *Loss* de treinamento e validação do modelo EdgeNeXt para o conjunto Herlev contendo sete classes. Fonte: Elaborado pelo autor.

5.2 Resultados no Conjunto de Dados *SIPaKMeD*

Os experimentos conduzidos com o conjunto de dados *SIPaKMeD* foram estruturados em três diferentes configurações de classificação, com o intuito de avaliar a robustez e a flexibilidade da abordagem proposta frente a distintos graus de complexidade na tarefa de reconhecimento de padrões citológicos.

A primeira configuração compreende a classificação em cinco classes, mantendo a configuração original do conjunto, a qual contempla todas as categorias citológicas previstas no conjunto de dados. Essa configuração visa testar a capacidade do modelo em discriminar um espectro mais amplo de tipos celulares.

A segunda configuração consistiu na reagrupação das cinco categorias em três classes: células normais, anormais e benignas. Essa abordagem intermediária permite investigar o desempenho do classificador em um contexto mais simplificado, mas ainda representativo da variabilidade morfológica presente em exames citopatológicos.

Por fim, a terceira configuração corresponde à classificação binária, em que as imagens são agrupadas em duas categorias: normais e anormais. Essa simplificação é especialmente relevante para aplicações práticas de triagem, nas quais o objetivo é distinguir rapidamente casos suspeitos que demandam investigação mais aprofundada.

Para todas as configurações, foram aplicadas as mesmas métricas de desempenho utilizadas na avaliação do conjunto *Herlev*, incluindo acurácia, precisão, *recall* e F1-score, de modo a garantir consistência nas análises comparativas. As Tabelas 14, 15 e 16 apresentam de forma detalhada os resultados obtidos para cada uma das configurações utilizando a validação cruzada, destacando os desempenhos mais expressivos e evidenciando o impacto das diferentes abordagens sobre a eficácia geral da classificação. Para que seja possível comparar com outras abordagens da literatura, utilizou-se do *holdout* como será mostrada nas próximas subseções.

Tabela 14 – Desempenho no conjunto *SIPaKMeD* para 2 classes com validação cruzada.

Métrica	Abordagem Proposta	***EfficientViT_b2***	***EVA02_small***	***EdgeNeXt_base***
Acurácia	**99,73 ± 0,13%**	99,57 ± 0,26%	99,40 ± 0,25%	99,54 ± 0,19%
Precisão	**99,73 ± 0,13%**	99,57 ± 0,26%	99,40 ± 0,25%	99,54 ± 0,19%
Recall	**99,73 ± 0,13%**	99,57 ± 0,26%	99,40 ± 0,25%	99,54 ± 0,19%
F1-Score	**99,73 ± 0,13%**	99,57 ± 0,26%	99,40 ± 0,25%	99,54 ± 0,19%

Tabela 15 – Desempenho no conjunto *SIPaKMeD* para 3 classes com validação cruzada.

Métrica	Abordagem Proposta	***EfficientViT_b2***	***EVA02_small***	***EdgeNeXt_base***
Acurácia	**98,96 ± 0,19%**	98,67 ± 0,32%	98,20 ± 0,39%	98,45 ± 0,33%
Precisão	**98,97 ± 0,19%**	98,68 ± 0,31%	98,20 ± 0,39%	98,46 ± 0,32%
Recall	**98,96 ± 0,19%**	98,67 ± 0,32%	98,20 ± 0,39%	98,45 ± 0,33%
F1-Score	**98,96 ± 0,19%**	98,67 ± 0,32%	98,20 ± 0,39%	98,45 ± 0,33%

Tabela 16 – Desempenho no conjunto *SIPaKMeD* para 5 classes com validação cruzada.

Métrica	Abordagem Proposta	***EfficientViT_b2***	***EVA02_small***	***EdgeNeXt_base***
Acurácia	**98,08 ± 0,54%**	97,36 ± 0,72%	96,84 ± 0,57%	97,15 ± 0,44%
Precisão	**98,09 ± 0,54%**	97,37 ± 0,73%	96,87 ± 0,57%	97,17 ± 0,44%
Recall	**98,08 ± 0,54%**	97,36 ± 0,72%	96,84 ± 0,57%	97,15 ± 0,45%
F1-Score	**98,08 ± 0,54%**	97,35 ± 0,73%	96,84 ± 0,58%	97,15 ± 0,44%

Resultados utilizando o *holdout*

Para o treinamento, validação e teste da abordagem proposta o conjunto de dados foi dividido na proporção de 70%, 20% e 10% respectivamente de modo estratificado. Em seguida, treinou-se cada modelo com uma com um dropout de 50% na última camada para mitigar o *overfitting*. Nas Subseções 5.2.1, 5.2.2 e 5.2.3 serão apresentados os gráficos de *loss* do treinamento e validação de cada modelo e suas respectivas matrizes de confusões, para duas, três e cinco classes respectivamente.

5.2.1 *SIPaKMeD* - 2 classes

A Tabela 17 apresenta uma comparação dos resultados obtidos pelo *ensemble* e pelos modelos individuais. Todos os modelos apresentaram desempenho satisfatório, com destaque para a abordagem proposta (*ensemble*), que atingiu o estado da arte e superou os modelos individuais em todas as métricas avaliadas.

Tabela 17 – Desempenho no conjunto *SIPaKMeD* para 2 classes utilizando holdout.

Métrica	Abordagem Proposta	***EfficientViT_b2***	***EVA02_small***	***EdgeNeXt_base***
Acurácia	**99,75%**	99,51%	99,01%	99,51%
Precisão	**99,75%**	99,59%	99,07%	99,59%
Recall	**99,69%**	99,38%	98,87%	99,38%
F1-Score	**99,74%**	99,48%	98,97%	99,48%

A Figura 30 ilustra a matriz de confusão da abordagem proposta, evidenciando a alta taxa de acerto entre as classes. As Figuras 31, 33 e 35 apresentam as matrizes de confusão dos modelos EfficientViT, EVA e EdgeNeXt, respectivamente.

Além disso, as Figuras 32, 34 e 36 mostram as curvas de *loss* de treinamento e validação dos respectivos modelos, demonstrando boa estabilidade e ausência de *overfitting*.

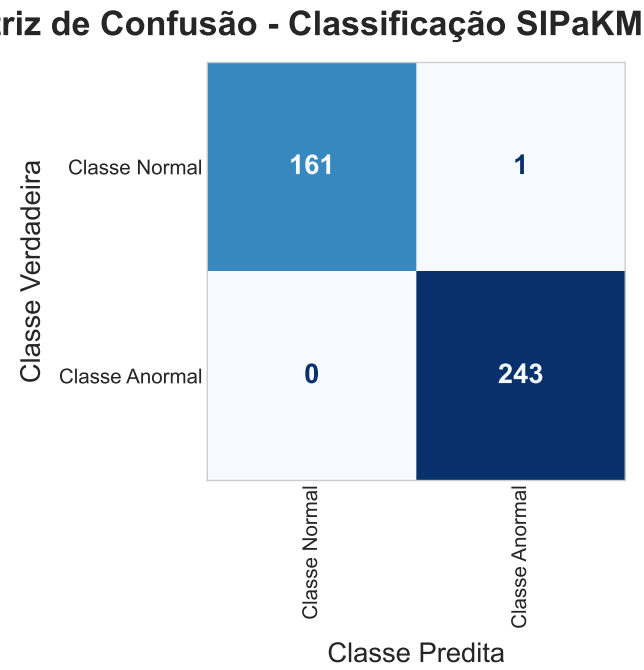


Figura 30 – Matriz de confusão da abordagem proposta para o conjunto *SIPaKMeD* classificado com as duas classes, o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor.

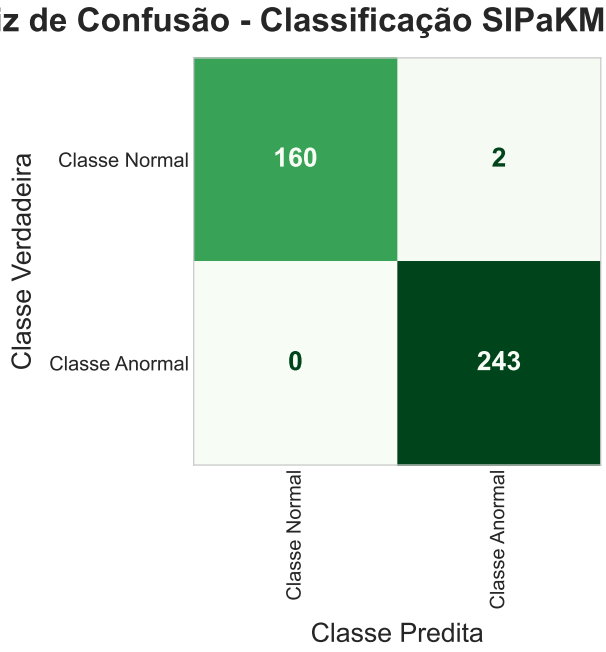


Figura 31 – Matriz de confusão do modelo EfficientViT para o conjunto *SIPaKMeD* classificado com as duas classes o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor.

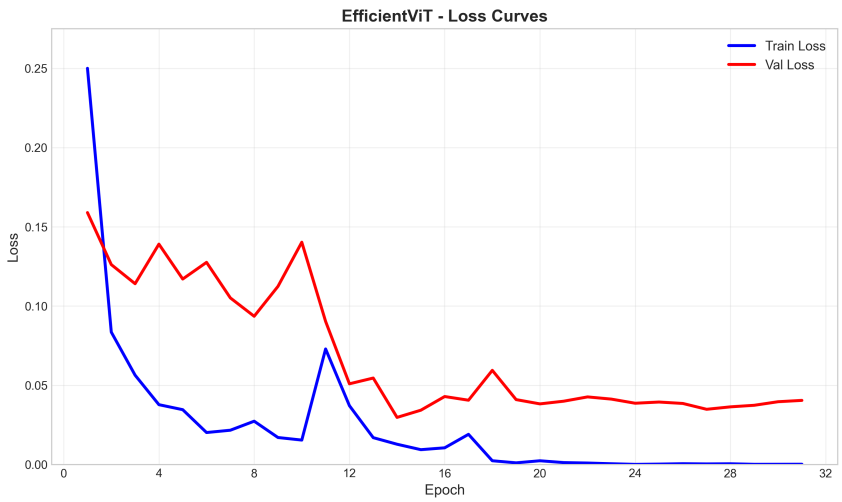


Figura 32 – *Loss* de treinamento e validação do modelo EfficientViT para o conjunto *SIPaKMeD* contendo duas classes. Fonte: Elaborado pelo autor.

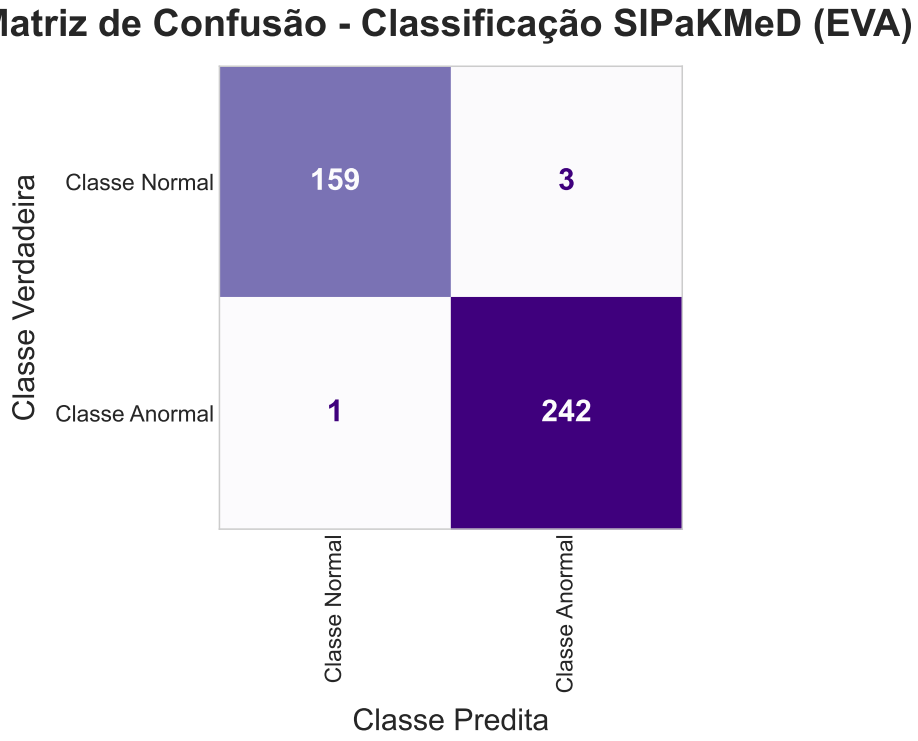


Figura 33 – Matriz de confusão do modelo EVA para o conjunto *SIPaKMeD* classificado com as duas classes, o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor.

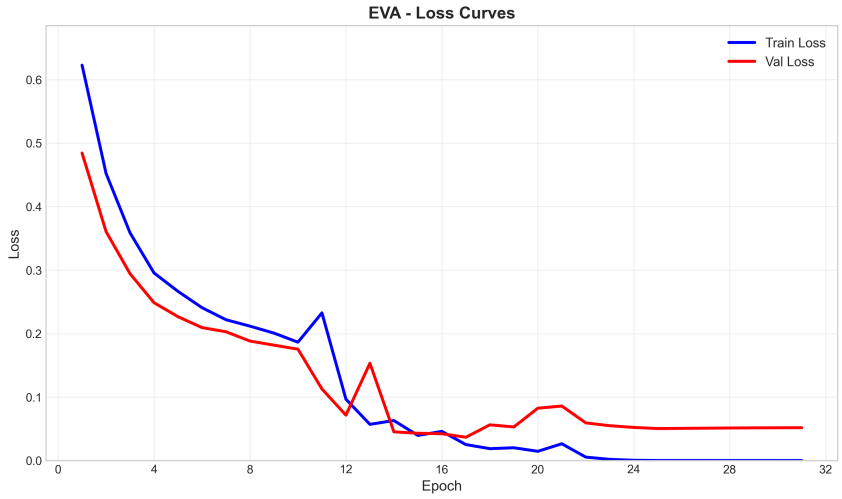


Figura 34 – *Loss* de treinamento e validação do modelo EVA para o conjunto *SIPaKMeD* classificado com as duas classes, o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor.

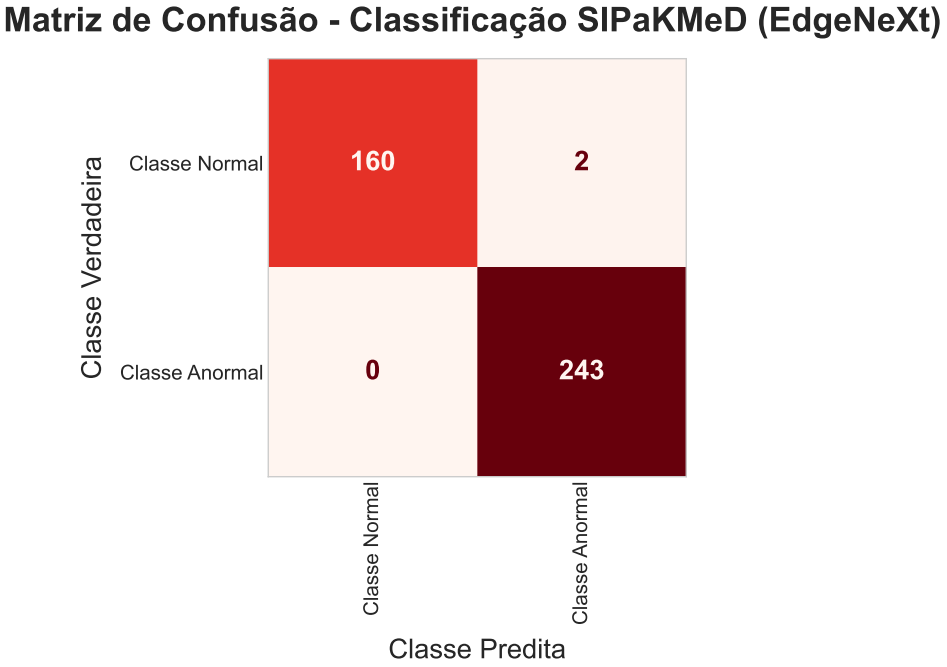


Figura 35 – Matriz de confusão do modelo EdgeNeXt para o conjunto *SIPaKMeD* classificado com as duas classes, o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor.

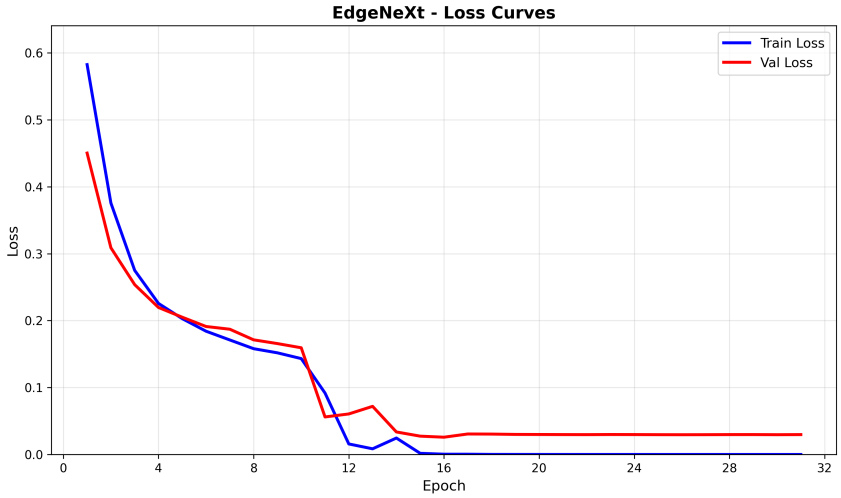


Figura 36 – *Loss* de treinamento e validação do modelo EdgeNeXtpara o conjunto *SIPaKMeD* classificado com as duas classes, o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor.

5.2.2 *SIPaKMeD* - 3 classes

A Tabela 18 apresenta uma comparação dos resultados obtidos pelo *ensemble* e pelos modelos individuais. Todos os modelos apresentaram excelente desempenho, com destaque para o modelo EVA, que atingiu o melhor resultado e superou inclusive o *ensemble*.

Tabela 18 – Desempenho no conjunto *SIPaKMeD* para 3 classes utilizando holdout.

Métrica	**Abordagem Proposta**	***EfficientViT_b2***	***EVA02_small***	***EdgeNeXt_base***
Acurácia	97,53%	97,53%	**98,27%**	96,79%
Precisão	97,11%	97,52%	**98,12%**	95,93%
Recall	97,11%	**97,52%**	97,48%	96,47%
F1-Score	97,09%	97,05%	**97,78%**	96,19%

A Figura 37 apresenta a matriz de confusão da abordagem proposta, evidenciando elevada taxa de acerto nas três classes. As Figuras 38, 40 e 42 exibem as matrizes de confusão dos modelos EfficientViT, EVA e EdgeNeXt, respectivamente.

As curvas de *loss* de treinamento e validação, mostradas nas Figuras 39, 41 e 43, indicam convergência estável e ausência de sobreajuste durante o processo de treinamento.

5.2.3 *SIPaKMeD* - 5 classes

A Tabela 19 apresenta uma comparação dos resultados obtidos pelo *ensemble* e pelos modelos individuais. Todos os modelos apresentaram desempenho satisfatório, com destaque para o modelo EVA, que alcançou os melhores resultados, superando inclusive a abordagem de *ensemble*.

A Figura 44 apresenta a matriz de confusão da abordagem proposta, evidenciando boa capacidade de discriminação entre as cinco classes. As Figuras 45, 47 e 49 mostram as matrizes de confusão dos modelos EfficientViT, EVA e EdgeNeXt, respectivamente. Nota-se que o modelo

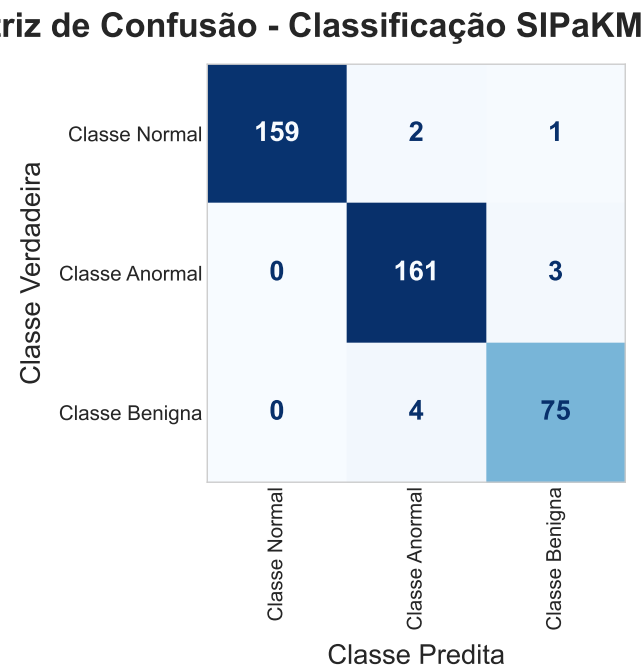


Figura 37 – Matriz de confusão da abordagem proposta para o conjunto *SIPaKMeD* classificado com as três classes, o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor.

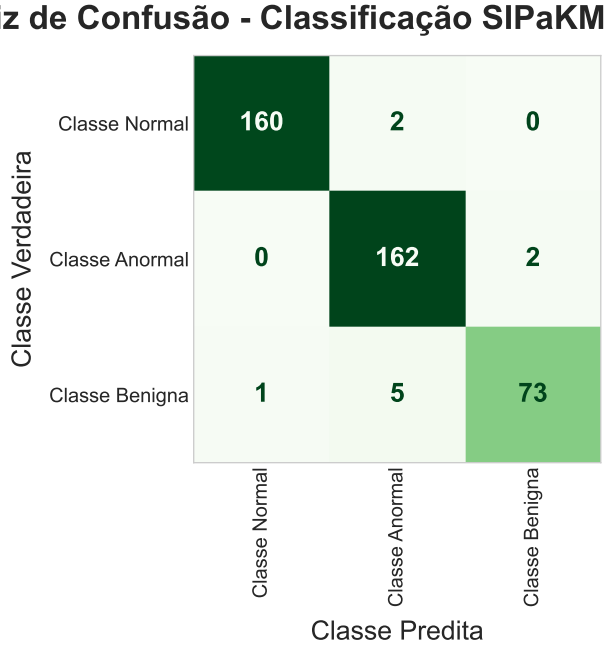


Figura 38 – Matriz de confusão do modelo EfficientViT para o conjunto *SIPaKMeD* classificado com as três classes o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor.

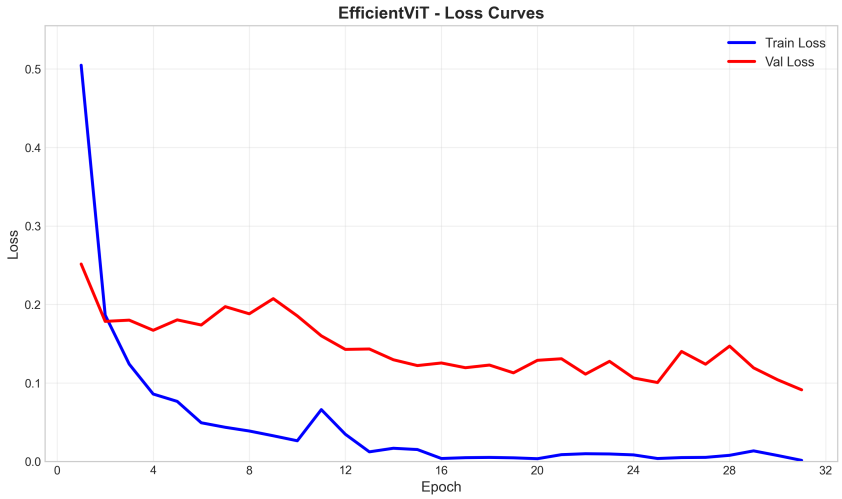


Figura 39 – *Loss* de treinamento e validação do modelo EfficientViT para o conjunto *SIPaKMeD* contendo três classes. Fonte: Elaborado pelo autor.

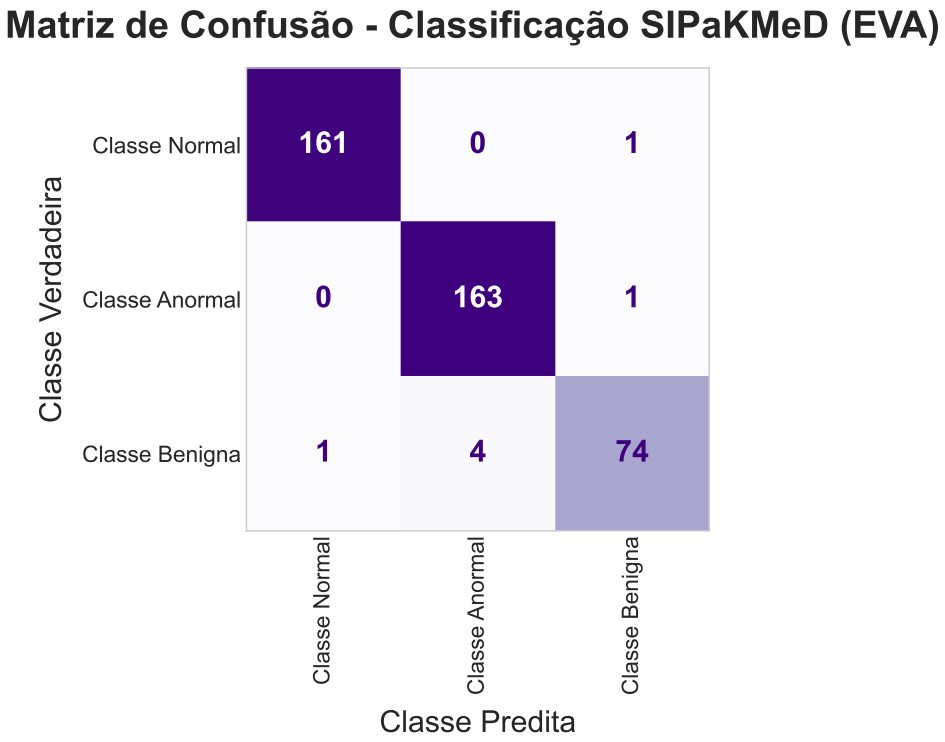


Figura 40 – Matriz de confusão do modelo EVA para o conjunto *SIPaKMeD* classificado com as três classes, o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor.

Tabela 19 – Desempenho no conjunto SIPaKMeD para 5 classes utilizando holdout.

Métrica	Abordagem Proposta	***EfficientViT_b2***	***EVA02_small***	***EdgeNeXt_base***
Acurácia	**97,53%**	96,54%	97,04%	96,79%
Precisão	**97,57%**	96,55%	97,07%	96,84%
Recall	**97,52%**	96,54%	97,06%	96,78%
F1-Score	**97,54%**	96,55%	97,05%	96,79%

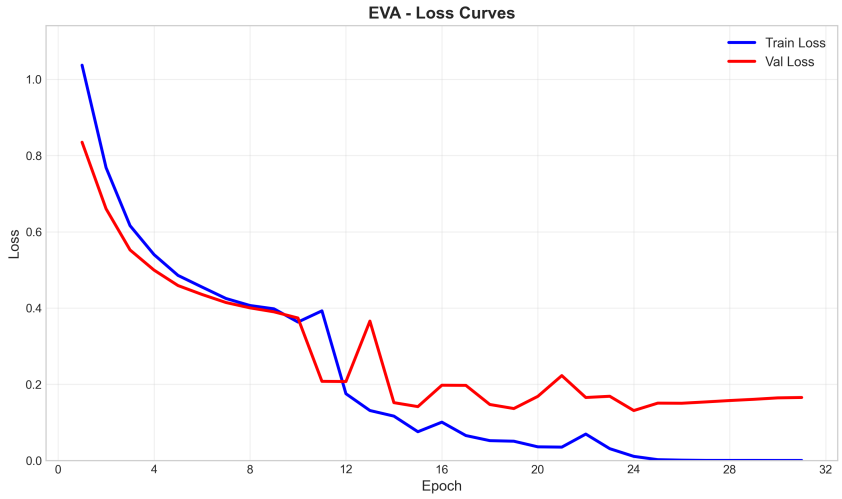


Figura 41 – *Loss* de treinamento e validação do modelo EVA para o conjunto *SIPaKMeD* classificado com as três classes, o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor.

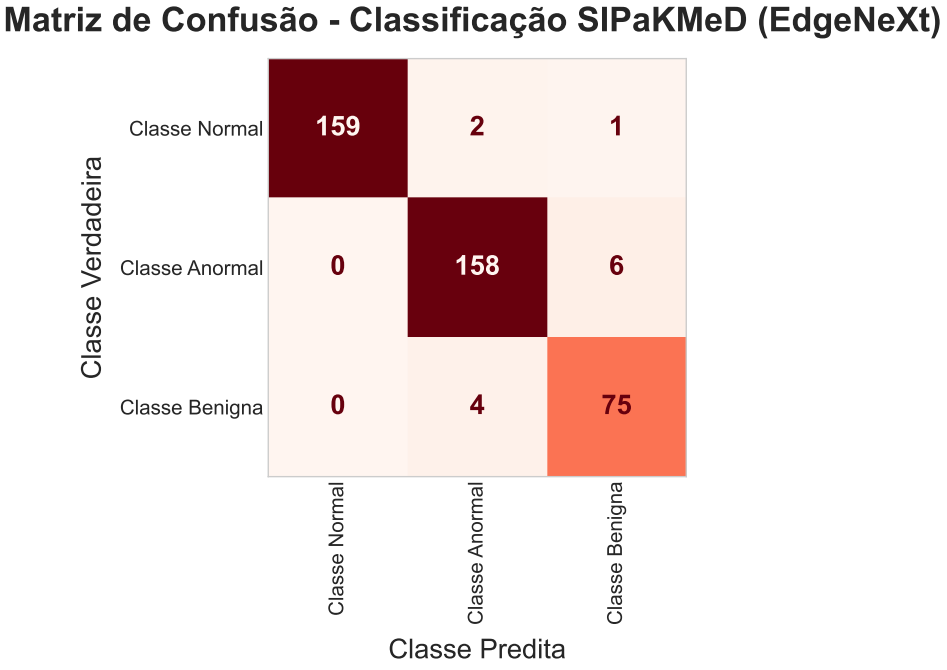


Figura 42 – Matriz de confusão do modelo EdgeNeXt para o conjunto *SIPaKMeD* classificado com as três classes, o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor.

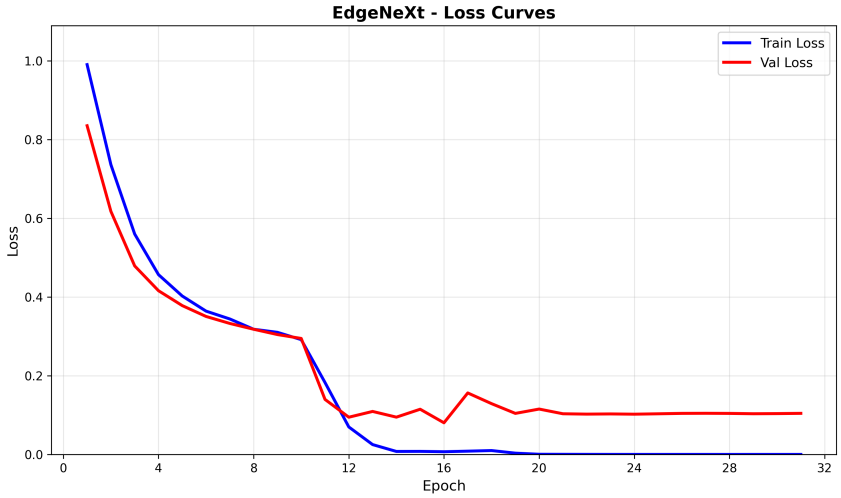


Figura 43 – *Loss* de treinamento e validação do modelo EdgeNeXtpara o conjunto *SIPaKMeD* classificado com as três classes, o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor.

EVA apresentou menor confusão entre classes, principalmente nas categorias intermediárias, refletindo seu melhor desempenho geral.

As curvas de *loss* de treinamento e validação apresentadas nas Figuras 46, 48 e 50 mostram comportamento estável e convergente, sem aumento significativo da perda de validação ao longo das épocas. Esse padrão indica ausência de sobreajuste e boa capacidade de generalização dos modelos, mesmo diante do maior número de classes.

Em particular, o modelo EVA apresentou as curvas de perda mais estáveis e próximas, o que sugere um ajuste adequado entre as etapas de treinamento e validação, compatível com o desempenho observado na Tabela 19. Por outro lado, os modelos EfficientViT e EdgeNeXt apresentaram pequenas oscilações na perda de validação, típicas de variações estatísticas decorrentes do aumento da complexidade do problema, mas sem caracterizar *overfitting*.

5.3 Discussão dos Resultados

A análise comparativa dos resultados obtidos nos conjuntos *Herlev* e *SIPaKMeD* permite avaliar de forma abrangente a robustez e a generalização da abordagem proposta em diferentes cenários de complexidade morfológica. De modo geral, os resultados evidenciam a eficácia do *ensemble* em integrar as vantagens complementares de arquiteturas convolucionais e transformadores visuais, produzindo representações discriminativas mais consistentes e estáveis.

5.3.1 Desempenho e Robustez da Abordagem Proposta

O modelo baseado em *ensemble* demonstrou desempenho superior ou equivalente aos modelos individuais em praticamente todos os experimentos realizados, alcançando resultados de estado da arte em ambos os conjuntos de dados. Nos cenários binários, tanto em *Herlev* quanto

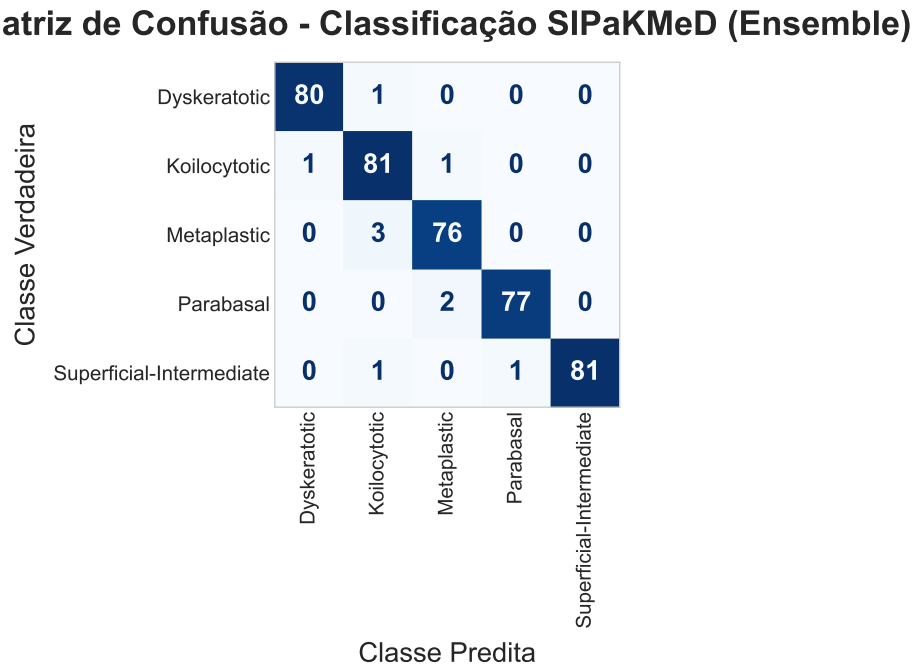


Figura 44 – Matriz de confusão da abordagem proposta para o conjunto *SIPaKMeD* classificado com as cinco classes, o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor.

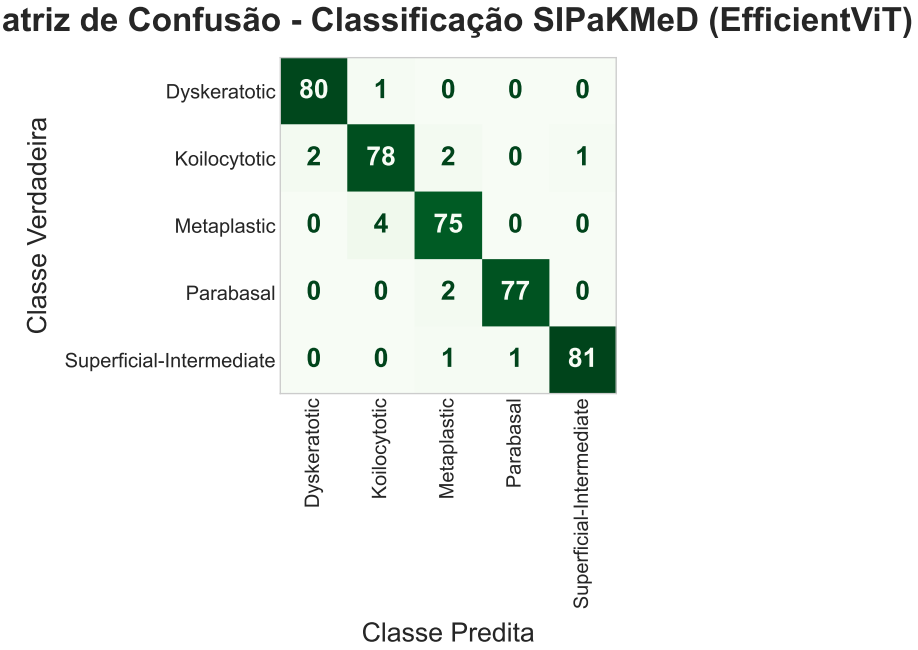


Figura 45 – Matriz de confusão do modelo EfficientViT para o conjunto *SIPaKMeD* classificado com as cinco classes o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor.

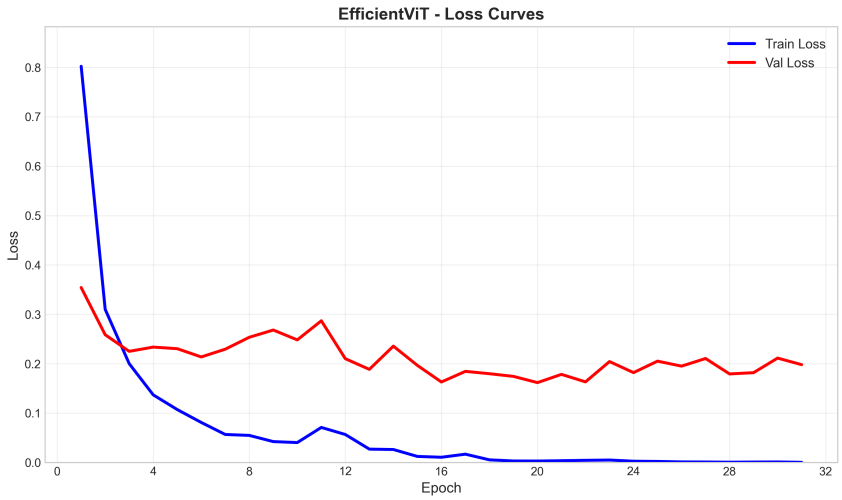


Figura 46 – *Loss* de treinamento e validação do modelo EfficientViT para o conjunto *SIPaKMeD* contendo cinco classes. Fonte: Elaborado pelo autor.

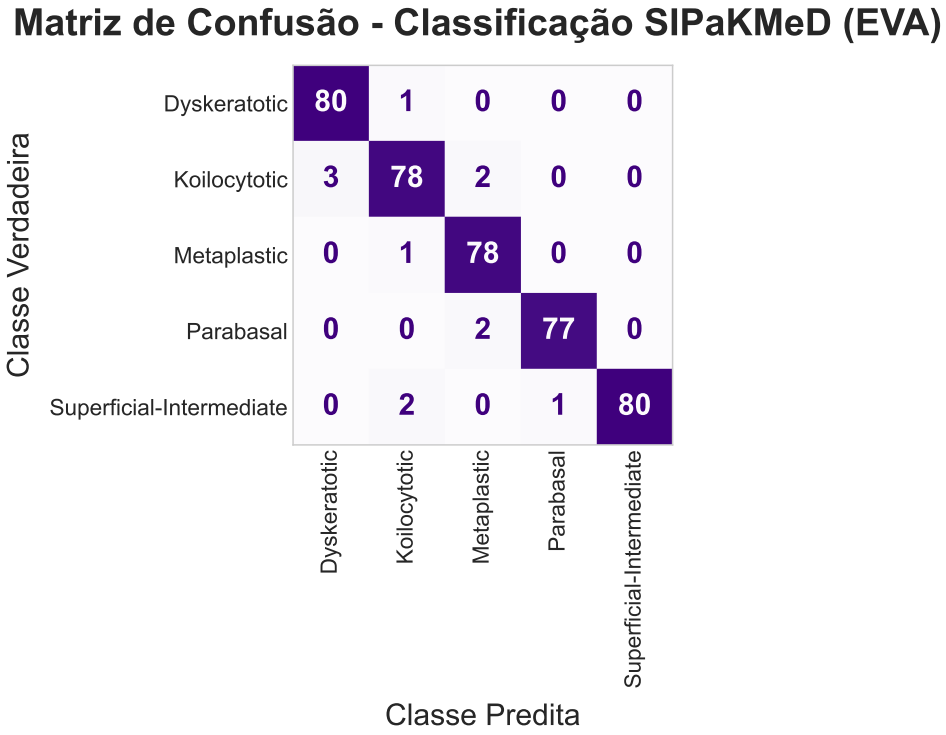


Figura 47 – Matriz de confusão do modelo EVA para o conjunto *SIPaKMeD* classificado com as cinco classes, o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor.

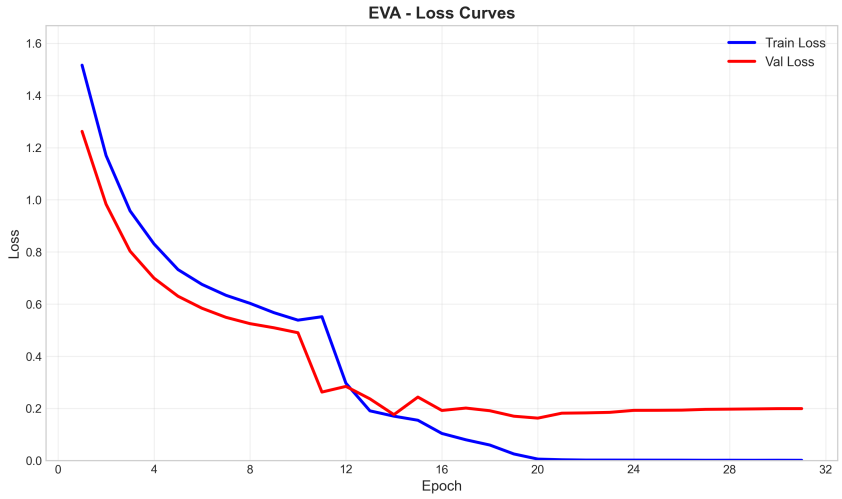


Figura 48 – *Loss* de treinamento e validação do modelo EVA para o conjunto *SIPaKMeD* classificado com as três classes, o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor.

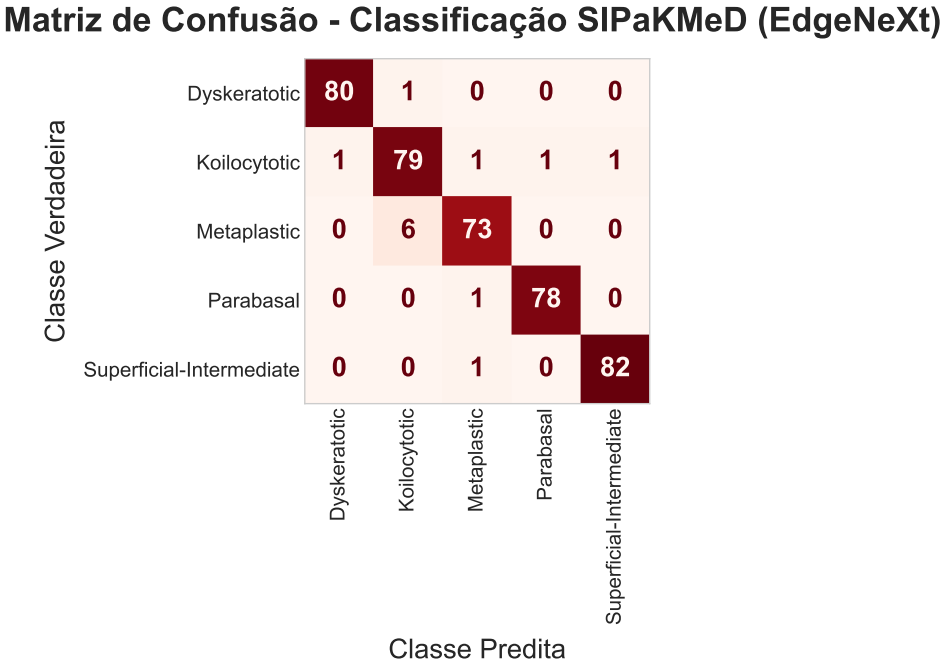


Figura 49 – Matriz de confusão do modelo EdgeNeXt para o conjunto *SIPaKMeD* classificado com as cinco classes, o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor.

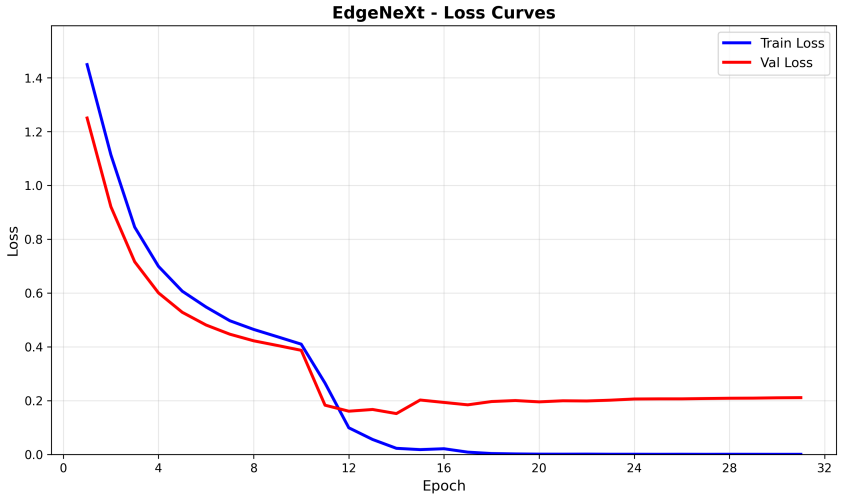


Figura 50 – *Loss* de treinamento e validação do modelo EdgeNeXtpara o conjunto *SIPaKMeD* classificado com as cinco classes, o conjunto de teste foi composto de 405 imagens. Fonte: Elaboração do autor.

em *SIPaKMeD*, as acurácias ultrapassaram 98%, indicando excelente capacidade de separação entre amostras normais e anormais, um aspecto crucial em sistemas de triagem automatizada.

Nas configurações multiclasse, observou-se uma redução esperada no desempenho, devido ao aumento da complexidade da tarefa e à similaridade morfológica entre subclasses citológicas. Ainda assim, o modelo proposto manteve desempenho competitivo, superando ou igualando métodos recentes da literatura, mesmo em condições experimentais mais rigorosas (validação cruzada *5-fold*), o que reforça a consistência estatística e a generalização da proposta.

5.3.2 Análise Específica: Conjunto *Herlev*

Os experimentos com o conjunto *Herlev* mostraram que a abordagem proposta alcançou **98,35%** (validação cruzada) de acurácia na classificação binária, superando trabalhos de referência como o de Sholik et al. (SHOLIK; FATICHAH; AMALIAH, 2024). Já para o *holdout* a acurácia atingida foi de 97,83% para o *ensemble* e 98,91% para o modelo EdgeNeXt que destacou-se de todos os modelos. Para avaliação das classes preditas, as matrizes de confusão (Figuras 16, 17, 19 e 21) evidenciam a baixa taxa de falsos negativos, reforçando a confiabilidade do modelo na detecção de anormalidades. As curvas de *loss* de treinamento e validação (Figuras 18, 20 e 22) indicam uma maior convergência e estabilidade no treinamento em relação aos treinamentos dos modelos no conjunto *Herlev* multiclasse.

Na classificação com sete classes, a acurácia do *ensemble* reduziu para cerca de **83,40%** utilizando-se validação cruzada, o que reflete o aumento da dificuldade em distinguir tipos celulares com diferenças sutis, especialmente entre categorias intermediárias (como células displásicas leves e moderadas). No contexto de Holdout, a queda foi mais brusca, chegando a 79,35% de acurácia para a abordagem proposta e uma queda maior para os modelos individuais que ficaram em uma faixa de resultados em torno de 70%.

As curvas de perda (Figuras 25, 27 e 29) indicam uma maior dificuldade de convergência

dos modelos, devido à dificuldade de diferenciação entre os subtipos celulares anormais que os modelos erraram mais. As matrizes de confusão do *ensemble*, EfficientViT, do EVA e do EdgeNeXt (Figuras 23, 24, 26 e 28) indicam que os modelos diferenciar bem as classes os subtipos normais, porém possuem mais dificuldades em diferenciar os subtipos anormais.

5.3.3 Análise Específica: Conjunto *SIPaKMeD*

O conjunto *SIPaKMeD*, um conjunto que possui imagens de células com uma melhor qualidade em relação ao *Herlev* e mais células, permitiu avaliar a escalabilidade da abordagem. Nas três configurações testadas, a performance foi elevada em todas as métricas, destacando-se o ensemble e, em alguns casos, o modelo EVA isoladamente.

Na configuração binária (duas classes), o modelo atingiu desempenho próximo à perfeição, com acurácia de 99,75% para *holdout* e 99,73% para validação cruzada. As curvas de perda (Figuras 32, 34, 36) indicando convergência suave e ausência de *overfitting*. Esse resultado confirma a eficácia da fusão de representações provenientes de CNNs e ViTs, capazes de capturar tanto padrões locais quanto globais das estruturas celulares. As matrizes de confusão do *ensemble*, EfficientViT, do EVA e do EdgeNeXt (Figuras 30, 31, 33 e 35) indicam que os modelos conseguem diferenciar bem as classes normais das anormais, com uma leve tendência para classes anormais.

Na configuração de três classes, o modelo EVA apresentou o melhor equilíbrio entre as métricas em *holdout*, superando inclusive o *ensemble*. No contexto de validação cruzada, o ensemble apresentou melhores resultados. As matrizes de confusão do *ensemble*, EfficientViT, do EVA e do EdgeNeXt (Figuras 37, 38, 40 e 42) indicam que os modelos conseguem diferenciar bem as classes, porém a classe benígna e anormal são confudidas nas predições dos modelos. Já para as curvas de perda, Figuras 39, 41 e 43 indicam convergência e estabilidade no treinamento do modelo e que não houve *overfitting*.

Por fim, na configuração de cinco classes, que representa a classificação mais complexa do conjunto, observou-se uma ligeira queda na acurácia de 98,08% na validação cruzada e 97,53% em *holdout* na abordagem proposta, ainda superior à maioria dos trabalhos da literatura. As curvas de perda (Figuras 46, 48, e 50) permaneceram estáveis, sem sinais de divergência, indicando que a redução de desempenho está relacionada à alta sobreposição entre classes adjacentes, uma limitação conhecida em classificações citológicas finas. De acordo com as matrizes de confusão do EfficientViT, do EVA e do EdgeNeXt (Figuras 45, 47 e 49) indicam que os modelos conseguem diferenciar bem os subtipos celulares, destacando-se do conjunto *Herlev* que há mais dificuldades.

5.3.4 Comparação Entre os Conjuntos e Implicações Clínicas

A comparação entre os conjuntos *Herlev* e *SIPaKMeD* evidencia o impacto da diversidade morfológica e do tamanho da base de dados na estabilidade dos resultados. Embora o *Herlev* tenha proporcionado acurácias elevadas em cenários binários, o *SIPaKMeD* apresentou maior consistência em múltiplas configurações, refletindo a robustez do modelo em contextos de maior variabilidade intra e interclasse.

Do ponto de vista clínico, o **Recall** mostrou-se a métrica mais relevante. O fato de o *ensemble* manter elevados valores de sensibilidade e F1-score em ambos os conjuntos reforça sua aplicabilidade em triagens automatizadas, nas quais a prioridade é reduzir a probabilidade de falsos negativos, garantindo que casos suspeitos sejam devidamente encaminhados à avaliação humana.

5.3.5 Comparativo com a Abordagem Proposta e os Resultados da Literatura

A Tabela 20 apresenta uma síntese comparativa entre a abordagem proposta e os principais trabalhos da literatura. Nos experimentos realizados, observa-se que o método proposto atingiu o estado da arte nos cenários de classificação binária, tanto no conjunto *Herlev* quanto no *SIPaKMeD*, alcançando acurácias superiores a 98% e demonstrando alta capacidade discriminativa entre amostras normais e anormais. Esse desempenho evidencia a efetividade do *ensemble* na integração de representações extraídas por arquiteturas complementares, como CNNs e Vision Transformers.

Nos cenários multiclasse, a abordagem manteve resultados expressivos, especialmente no conjunto *SIPaKMeD*, em que a acurácia permaneceu acima de 98% nas configurações com três e cinco classes. Tal comportamento indica que o modelo foi capaz de generalizar de forma consistente mesmo em contextos de maior complexidade morfológica, caracterizados por uma maior sobreposição entre classes citológicas.

Por outro lado, no conjunto *Herlev*, a tarefa multiclasse apresentou uma redução mais acentuada de desempenho, alcançando cerca de 83% de acurácia. Essa queda é atribuída à limitação intrínseca do conjunto de dados, que possui menor variabilidade e número de amostras em determinadas classes, além de uma distribuição desbalanceada entre categorias normais e anormais.

De forma geral, os resultados obtidos reforçam que a abordagem proposta é eficaz e robusta, apresentando desempenho comparável ou superior aos métodos de referência da literatura em diferentes cenários experimentais. Além disso, a consistência entre as métricas de acurácia, precisão, *recall* e F1-score demonstra o equilíbrio do modelo entre sensibilidade e especificidade, aspecto essencial em aplicações clínicas.

5.3.6 Validação Estatística e Considerações Finais

Os resultados obtidos foram confirmados estatisticamente por meio do teste de Wilcoxon Signed-Rank (Tabela 21), cujos valores de p inferiores a 0,01 indicam diferenças significativas entre o desempenho do *ensemble* e dos modelos individuais. Essa validação corrobora a superioridade consistente da abordagem proposta e elimina a possibilidade de ganhos ocasionais decorrentes do acaso. Em síntese, a abordagem proposta combina desempenho superior, estabilidade de treinamento e coerência clínica. Sua eficácia em múltiplos conjuntos e cenários de complexidade variável demonstra seu potencial para aplicações reais em sistemas de apoio ao diagnóstico citopatológico, contribuindo para processos de triagem mais precisos e eficientes.

Tabela 20 – Comparação da abordagem proposta com os principais estudos da literatura em cada cenário do experimento realizado. As siglas represetam: acurácia (acc), precisão (prec), conjunto de dados (conjunto), divisão (divisão de dados), validação cruzada com *k-fold (k-fold)*.

Conjunto	**Modelo**	**Acc**	**Prec**	***Recall***	***F1***	**Divisão**
Herlev (2 classes)	Abordagem proposta	**98,35**	98,37	98,35	**98,38**	*5-fold*
	Ensemble	97,83	96,15	**98,53**	97,25	*Holdout*
	(SHOLIK; FATICHAH; AMALIAH, 2024)	97,83	**98,75**	92,86	95,52	*Holdout*
Herlev (7 classes)	Abordagem proposta	**83,40**	83,59	**83,42**	**83,34**	*5-fold*
	Ensemble	79,35	**84,01**	82,22	82,56	*Holdout*
SIPaKMeD (2 classes)	Abordagem proposta	99,73	99,73	**99,73**	99,73	*5-fold*
	Ensemble	**99,75**	**99,75**	99,69	**99,74**	*Holdout*
	(ALOHALI et al., 2024)	99,01	99,01	99,01	99,01	*Holdout*
SIPaKMeD (3 classes)	Abordagem proposta	**98,96**	**98,97**	**98,96**	**98,96**	*5-fold*
	Ensemble	97,53	97,11	97,11	97,09	*Holdout*
	(SHOLIK; FATICHAH; AMALIAH, 2024)	98,52	98,81	97,78	98,28	*Holdout*
SIPaKMeD (5 classes)	Abordagem proposta	98,08	98,09	98,08	98,08	*5-fold*
	Ensemble	97,53	97,57	97,52	97,54	*Holdout*
	(ALOHALI et al., 2024)	98,81	**98,81**	**98,81**	**98,81**	*Holdout*
	(S. et al., 2024)	**99,70**	96,40	97,50	96,00	*K-fold*

Tabela 21 – *P*-valores obtidos pelo WSR, com nível de confiança de 99%, comparando a abordagem proposta com os melhores modelos em cada cenário de experimentos da validação cruzada.

Conjunto	**Comparação**	**Acurácia**	**Precisão**	***Recall***	***F1-score***
Herlev (2 classes)	*EdgeNeXt_base*	$9{,}92 \cdot 10^{-4}$	$3{,}29 \cdot 10^{-4}$	$9,92 \cdot 10^{-4}$	$6,53 \cdot 10^{-4}$
Herlev (7 classes)	*EfficientViT_b2*	$2{,}39 \cdot 10^{-4}$	$1,40 \cdot 10^{-4}$	$6,4 \cdot 10^{-5}$	$7,5 \cdot 10^{-5}$
SIPaKMeD (2 classes)	*EfficientViT_b2*	$2,848 \cdot 10^{-3}$	$4,003 \cdot 10^{-3}$	$2,482 \cdot 10^{-3}$	$3,495 \cdot 10^{-3}$
SIPaKMeD (3 classes)	*EfficientViT_b2*	$9,28 \cdot 10^{-4}$	$1,116 \cdot 10^{-3}$	$2,136 \cdot 10^{-3}$	$1,477 \cdot 10^{-3}$
SIPaKMeD (5 classes)	*EfficientViT_b2*	$1,31 \cdot 10^{-4}$	$2,0 \cdot 10^{-6}$	$4,0 \cdot 10^{-6}$	$2,0 \cdot 10^{-6}$

6 Conclusões e Trabalhos Futuros

Este trabalho se propôs a desenvolver e validar uma abordagem de *ensemble learning* para a classificação automatizada de células cervicais, um desafio de grande relevância clínica e computacional. A metodologia proposta, fundamentada na combinação sinérgica de arquiteturas de visão computacional de última geração, demonstrou não apenas ser viável, mas altamente eficaz, estabelecendo um novo padrão de desempenho na área.

6.1 Conclusões e Trabalhos Futuros

A presente dissertação propôs e validou uma estratégia de *ensemble learning*, inovando ao integrar as arquiteturas EfficientViT, EVA-02 e EdgeNeXt para a classificação automática de células cervicais. Os resultados, validados em protocolos robustos e conjuntos públicos amplamente utilizados, demonstram que a fusão de modelos com vieses indutivos complementares representa uma abordagem superior para a tarefa, atingindo o estado da arte em múltiplos cenários experimentais.

Apesar dos excelentes resultados obtidos, reconhecem-se limitações importantes: a validação foi realizada exclusivamente em bases públicas e a aplicabilidade em cenários clínicos reais permanece a ser explorada, sobretudo diante da heterogeneidade de equipamentos, populações e possíveis vieses institucionais. Ademais, ressalta-se a necessidade de avanços em interpretabilidade para que o uso clínico seja amplamente aceito e auditável.

Como caminhos futuros de pesquisa, recomenda-se:

- Exploração de *self-supervised learning* para aproveitar grandes volumes de imagens não rotuladas, aumentando a capacidade de generalização dos modelos;
- Adoção de técnicas de *distilação de modelo* para transferir o conhecimento de grandes arquiteturas para redes leves, como EdgeNeXt ou MobileNet, visando aplicações em ambientes restritos em termos de recursos computacionais;
- Incorporação de técnicas de explicabilidade, como Grad-CAM ou SHAP, com vistas a aumentar a confiança dos especialistas;
- Utilização de lâminas inteiras, com aplicação de modelos de segmentação ou detecção de células, aproximando ainda mais o cenário experimental da realidade prática em citopatologia;
- Expansão para análise multimodal, integrando dados de imagem com informações clínicas e agentes inteligentes.

Tais diretrizes reforçam o compromisso com solidez científica, impacto translacional e abertura para inovação contínua em sistemas inteligentes de apoio ao diagnóstico médico.

6.2 Avaliação do Cumprimento dos Objetivos Propostos

Os objetivos estabelecidos nesta investigação, tanto em âmbito geral quanto específico, foram atingidos de maneira satisfatória e rigorosa. A dissertação propôs, implementou e validou uma abordagem inovadora de *ensemble learning*, baseada na combinação de arquiteturas Efficient-ViT, EVA-02 e EdgeNeXt, visando aprimorar a classificação automática de células cervicais. Os experimentos realizados, empregando os conjuntos de dados *Herlev* e *SIPaKMeD* em múltiplos cenários de classificação, resultaram em métricas superiores às reportadas na literatura recente, atestando o êxito metodológico da proposta.

A obtenção de resultados de estado da arte, evidenciada por acurácias superiores a 98% nas tarefas binárias e desempenho altamente competitivo em cenários multiclasse, demonstra o alcance do objetivo central de contribuição para o desenvolvimento de sistemas computacionais sofisticados no contexto do rastreamento automatizado do câncer cervical.

Adicionalmente, os procedimentos de validação cruzada, análise estatística via teste de Wilcoxon e discussão das limitações experimentais revelam uma condução científica adequada, sustentando a robustez e credibilidade dos achados. Os desafios inerentes à diferenciação de subtipos celulares e à limitação de bases públicas são reconhecidos como oportunidades para pesquisas futuras, e não como entraves ao cumprimento dos objetivos delineados.

Conclui-se, portanto, que todos os objetivos propostos foram integralmente atingidos, consolidando relevantes contribuições acadêmicas e práticas para o avanço da área de visão computacional aplicada à citopatologia cervical.

Referências

ABD-ALHALEM, S. M. et al. Cervical cancer classification based on a bilinear convolutional neural network approach and random projection. *Engineering Applications of Artificial Intelligence*, v. 127, p. 107261, 2024. ISSN 0952-1976. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0952197623014458>. Citado 3 vezes nas páginas 9, 20 e 24.

AHMED, R. et al. Early detection and categorization of cervical cancer cells using smoothing cross entropy-based multi-deep transfer learning. *IEEE Access*, v. 12, p. 157838–157853, 2024. Citado na página 19.

AKBAR, A. H. et al. Layer Selection on Residual Network for Feature Extraction of Pap Smear Images. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, v. 36, n. 2, p. 56–66, dez. 2023. ISSN 2462-1943. Number: 2. Disponível em: <https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/article/view/3221>. Citado na página 17.

AKHILA, N. M. K.; AHILAN, A. Classification of cervical cancer using an autoencoder and cascaded multilayer perceptron. *IETE Journal of Research*, Taylor & Francis, v. 70, n. 1, p. 26–36, 2024. Disponível em: <https://doi.org/10.1080/03772063.2022.2142859>. Citado na página 22.

ALMOHIMEED, A. et al. ViT-PSO-SVM: Cervical Cancer Predication Based on Integrating Vision Transformer with Particle Swarm Optimization and Support Vector Machine. *Bioengineering*, v. 11, n. 7, p. 729, jul. 2024. ISSN 2306-5354. Number: 7 Publisher: Multidisciplinary Digital Publishing Institute. Disponível em: <https://www.mdpi.com/2306-5354/11/7/729>. Citado 2 vezes nas páginas 12 e 16.

ALOHALI, M. A. et al. Swin-GA-RF: genetic algorithm-based Swin Transformer and random forest for enhancing cervical cancer classification. *Frontiers in Oncology*, v. 14, jul. 2024. ISSN 2234-943X. Publisher: Frontiers. Disponível em: <https://www.frontiersin.org/journals/oncology/articles/10.3389/fonc.2024.1392301/full>. Citado 2 vezes nas páginas 15 e 77.

ANAND, M. S. et al. Jellyfish search optimizer with deep learning driven cervical cancer recognition and classification model on biomedical images. In: *2024 Second International Conference on Advances in Information Technology (ICAIT)*. [S.l.: s.n.], 2024. v. 1, p. 1–6. Citado na página 22.

ANAND, V.; BACHHAL, P. Cervical net: An effective convolution neural network for five-class classification of cervical cells. In: *2024 2nd International Conference on Device Intelligence, Computing and Communication Technologies (DICCT)*. [S.l.: s.n.], 2024. p. 51–55. Citado na página 14.

ASHFAQUE, K.; BEIJI, Z.; KUI, X. Cervix Visionator ELM: A Novel Approach to Early Detection of Cervical Cancer. In: *2023 8th International Conference on Control, Robotics and Cybernetics (CRC)*. [s.n.], 2024. p. 254–259. Disponível em: <https://ieeexplore.ieee.org/document/10488648/?arnumber=10488648>. Citado na página 18.

BEDELL, S. L. et al. Cervical cancer screening: Past, present, and future. *Sexual Medicine Reviews*, v. 8, n. 1, p. 28–37, 2020. ISSN 2050-0521. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2050052119301015>. Citado na página 1.

BRIDGE, P.; SAWILOWSKY, S. Increasing Physicians' Awareness of the Impact of Statistics on Research Outcomes: Comparative Power of the t-test and Wilcoxon Rank-Sum Test in Small Samples Applied Research. *J CLIN EPIDEMIOL*, 1999. Citado 2 vezes nas páginas 46 e 51.

CAI, H. et al. Efficientvit: Multi-scale linear attention for high-resolution dense prediction. *arXiv preprint arXiv:2205.14756*, 2022. Citado 3 vezes nas páginas 9, 31 e 39.

CHITRA, S. S. K. B.; SUBBULEKSHMI, D. Prediction models applying convolutional neural network based deep learning to cervical cancer outcomes. *IETE Journal of Research*, Taylor & Francis, v. 70, n. 8, p. 7058–7068, 2024. Disponível em: <https://doi.org/10.1080/03772063.2024.2305831>. Citado 2 vezes nas páginas 9 e 19.

CHUNG, J. et al. *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling*. 2014. Disponível em: <https://arxiv.org/abs/1412.3555>. Citado na página 35.

DEININGER, L. et al. *A comparative study between vision transformers and CNNs in digital pathology*. arXiv, 2022. ArXiv:2206.00389 [eess]. Disponível em: <http://arxiv.org/abs/2206.00389>. Citado na página 2.

DEVARAJ, S. et al. Deep learning approaches for analysing papsmear images to detect cervical cancer. *Wireless Personal Communications*, Springer, p. 1–18, 2024. Citado 2 vezes nas páginas 9 e 14.

DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. [S.l.: s.n.], 2019. p. 4171–4186. Citado na página 37.

DOSOVITSKIY, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: *International Conference on Learning Representations*. [s.n.], 2021. Disponível em: <https://openreview.net/forum?id=YicbFdNTTy>. Citado 3 vezes nas páginas 9, 2 e 38.

FABBRI, S. C. et al. Managing literature reviews information through visualization. In: *ICEIS (2)*. [S.l.: s.n.], 2012. p. 36–45. Citado na página 5.

FACELI, K. et al. Inteligência artificial: uma abordagem de aprendizado de máquina. 2021. Citado na página 45.

FANG, M. et al. Deep integrated fusion of local and global features for cervical cell classification. *Computers in Biology and Medicine*, v. 171, p. 108153, 2024. ISSN 0010-4825. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0010482524002373>. Citado na página 15.

FANG, M. et al. A systematic review on deep learning based methods for cervical cell image analysis. *Neurocomputing*, v. 610, p. 128630, 2024. ISSN 0925-2312. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0925231224014012>. Citado 5 vezes nas páginas 2, 7, 10, 44 e 48.

FANG, Y. et al. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, Elsevier, v. 149, p. 105171, 2024. Citado 4 vezes nas páginas 9, 31, 40 e 41.

FAWCETT, T. An introduction to roc analysis. *Pattern Recognition Letters*, v. 27, n. 8, p. 861–874, 2006. ISSN 0167-8655. ROC Analysis in Pattern Recognition. Disponível em: <https://www.sciencedirect.com/science/article/pii/S016786550500303X>. Citado na página 44.

FERLAY, J. et al. *Global Cancer Observatory: Cancer Today*. 2024. Acessado em 17 de abril de 2024. Disponível em: <https://gco.iarc.who.int/media/globocan/factsheets/cancers/23-cervix-uteri-fact-sheet.pdf>. Citado na página 1.

FERLAY, J. et al. *Global Cancer Observatory: Cancer Tomorrow (version 1.1)*. Lyon, France: International Agency for Research on Cancer, 2024. Acessado em 27 de dezembro de 2024. Disponível em: <https://gco.iarc.fr/tomorrow/en/dataviz/isotype?cancers=23&single_unit=500&years=2025&populations=76>. Citado na página 1.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. <http://www.deeplearningbook.org>. Citado 2 vezes nas páginas 33 e 34.

GOSWAMI, A.; GOSWAMI, N. G.; SAMPATHILA, N. Deep learning-based classification of cervical cancer using pap smear images. In: *2024 4th International Conference on Intelligent Technologies (CONIT)*. [S.l.: s.n.], 2024. p. 1–6. Citado na página 13.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural computation*, MIT press, v. 9, n. 8, p. 1735–1780, 1997. Citado na página 35.

HUSSAIN, E. Liquid based cytology pap smear images for multi-class diagnosis of cervical cancer. *Data Brief*, v. 4, 2019. Citado na página 23.

IKENBERG, H. et al. Comparison of the hologic genius digital diagnostics system with the thinprep imaging system—a retrospective assessment. *Cancer Cytopathology*, v. 131, n. 7, p. 424–432, 2023. Disponível em: <https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.1002/cncy.22695>. Citado na página 2.

JANTZEN, J. et al. Pap-smear benchmark data for pattern classification. *Nature inspired smart information systems (NiSIS 2005)*, p. 1–9, 2005. Citado 5 vezes nas páginas 9, 23, 29, 30 e 48.

JIANG, P. et al. A systematic review of deep learning-based cervical cytology screening: from cell identification to whole slide image analysis. *Artificial Intelligence Review*, Springer, v. 56, n. Suppl 2, p. 2687–2758, 2023. Citado 4 vezes nas páginas 1, 7, 10 e 48.

JIANG, P. et al. Interpretable detector for cervical cytology using self-attention and cell origin group guidance. *Engineering Applications of Artificial Intelligence*, v. 134, p. 108661, 2024. ISSN 0952-1976. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0952197624008194>. Citado 2 vezes nas páginas 9 e 19.

JOYNAB, N. S. et al. A federated learning aided system for classifying cervical cancer using pap-smear images. *Informatics in Medicine Unlocked*, v. 47, p. 101496, 2024. ISSN 2352-9148. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2352914824000522>. Citado na página 13.

KAUR, A. et al. Predicting cervical cancer with deep learning: Comparative analysis of vggnet, googlenet, and densenet121 models. In: *2024 2nd World Conference on Communication & Computing (WCONF)*. [S.l.: s.n.], 2024. p. 1–4. Citado na página 14.

KHANARSA, P.; KITSIRANUWAT, S. Deep Learning-based Ensemble Approach for Conventional Pap Smear Image Classification. *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, v. 18, n. 1, p. 101–111, fev. 2024. ISSN 2286-9131. Number: 1. Disponível em: <https://ph01.tci-thaijo.org/index.php/ecticit/article/view/254621>. Citado na página 21.

KHARE, S. K. et al. A systematic review and research recommendations on artificial intelligence for automated cervical cancer detection. *WIREs Data Mining and Knowledge Discovery*, v. 14, n. 6, p. e1550, 2024. Disponível em: <https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1550>. Citado 3 vezes nas páginas 7, 10 e 48.

KHOWAJA, A.; ZOU, B.; KUI, X. Enhancing cervical cancer diagnosis: Integrated attention-transformer system with weakly supervised learning. *Image and Vision Computing*, v. 149, p. 105193, set. 2024. ISSN 02628856. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0262885624002981>. Citado na página 16.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. Red Hook, NY, USA: Curran Associates Inc., 2012. (NIPS'12), p. 1097–1105. Citado na página 2.

LECUN, Y. et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, v. 86, n. 11, p. 2278–2324, 1998. Citado 3 vezes nas páginas 9, 33 e 35.

LI, M. et al. VTCNet: A Feature Fusion DL Model Based on CNN and ViT for the Classification of Cervical Cells. *International Journal of Imaging Systems and Technology*, v. 34, n. 5, p. e23161, 2024. ISSN 1098-1098. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ima.23161. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/ima.23161>. Citado na página 16.

LIMA, M.; BRITTO, M.; NETO, L. B. Classificação de células cervicais cancerígenas com cnns e vits em ensemble para auxílio ao diagnóstico médico. In: *Anais do XXV Simpósio Brasileiro de Computação Aplicada à Saúde*. Porto Alegre, RS, Brasil: SBC, 2025. p. 497–508. ISSN 2763-8952. Disponível em: <https://sol.sbc.org.br/index.php/sbcas/article/view/35521>. Citado na página 4.

LIU, Y. et al. A novel ensemble framework based on CNN models and Swin transformer for cervical cytology image classification. In: *Seventh International Conference on Computer Graphics and Virtuality (ICCGV 2024)*. SPIE, 2024. v. 13158, p. 150–161. Disponível em: <https://www-spiedigitallibrary-org.ez17.periodicos.capes.gov.br/conference-proceedings-of-spie/13158/131580F/A-novel-ensemble-framework-based-on-CNN-models-and-Swin/10.1117/12.3029445.full>. Citado na página 21.

MAAZ, M. et al. Edgenext: Efficiently amalgamated cnn-transformer architecture for mobile vision applications. In: SPRINGER. *CADL*. [S.l.], 2022. Citado 5 vezes nas páginas 9, 31, 42, 43 e 44.

MAESAROH, I. S. et al. Cervical cancer classification from pap smear using xception model. In: *2024 10th International Conference on Wireless and Telematics (ICWT)*. [S.l.: s.n.], 2024. p. 1–5. Citado na página 13.

MAHAJAN, P.; KAUR, P. Improving cervical cancer classification in PAP smear images with enhanced segmentation and deep progressive learning-based techniques. *Diagnostic Cytopathology*, v. 52, n. 6, p. 313–324, 2024. ISSN 1097-0339. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/dc.25295. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/dc.25295>. Citado na página 17.

MATHEW, B.; CHERIYAN, R. A detailed review on classification and risk factor analysis of cervical cancer using artificial intelligence. In: *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*. [S.l.: s.n.], 2024. v. 2, p. 1–6. Citado 2 vezes nas páginas 11 e 48.

MOHER, D. et al. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *International journal of surgery*, Elsevier, v. 8, n. 5, p. 336–341, 2010. Citado na página 5.

MUNIRATHINAM, R. et al. Manifold cervical abnormality categorization using deep learned features and support vector machine. In: *2024 Tenth International Conference on Bio Signals, Images, and Instrumentation (ICBSII)*. [S.l.: s.n.], 2024. p. 1–6. Citado 2 vezes nas páginas 20 e 24.

NOUR, M. K. et al. Computer aided cervical cancer diagnosis using gazelle optimization algorithm with deep learning model. *IEEE Access*, v. 12, p. 13046–13054, 2024. Citado 2 vezes nas páginas 17 e 24.

OLIVEIRA, A. C. A. de et al. Comparison between Conventional Cytology and Liquid-Based Cytology in the Tertiary Brazilian Navy Hospital in Rio de Janeiro. *Acta Cytologica*, v. 64, n. 6, p. 539–546, 06 2020. ISSN 0001-5547. Disponível em: <https://doi.org/10.1159/000508018>. Citado na página 1.

PACAL, I. Maxcervixt: A novel lightweight vision transformer-based approach for precise cervical cancer detection. *Knowledge-Based Systems*, v. 289, p. 111482, 2024. ISSN 0950-7051. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0950705124001175>. Citado 2 vezes nas páginas 9 e 20.

PLISSITI, M. E. et al. Sipakmed: A new dataset for feature and image based classification of normal and pathological cervical cells in pap smear images. In: IEEE. *2018 25th IEEE international conference on image processing (ICIP)*. [S.l.], 2018. p. 3144–3148. Citado 5 vezes nas páginas 9, 23, 30, 31 e 48.

PROME, T. A.; TAHITI, T. J. Mobilenetv2 based cervical cancer classification using pap smear images. In: *2024 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA)*. [S.l.: s.n.], 2024. p. 1–7. Citado na página 13.

QIN, J. et al. Cell comparative learning: A cervical cytopathology whole slide image classification method using normal and abnormal cells. *Computerized Medical Imaging and Graphics*, v. 117, p. 102427, 2024. ISSN 0895-6111. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0895611124001046>. Citado 2 vezes nas páginas 16 e 24.

RAMESH, H. et al. Cerviscan: Advancing cervical cancer detection through deep learning innovations. In: *2024 IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS)*. [S.l.: s.n.], 2024. p. 344–349. Citado na página 13.

REZENDE, M. T. et al. Cric searchable image database as a public platform for conventional pap smear cytology data. v. 8, n. 1, p. 151, 2021. ISSN 2052-4463. Publisher: Nature Publishing Group. Disponível em: <https://www.nature.com/articles/s41597-021-00933-8>. Citado na página 23.

RIANA, D. et al. Repomedunm: A new dataset for feature extraction and training of deep learning network for classification of pap smear images. In: MANTORO, T. et al. (Ed.). *Neural Information Processing*. Cham: Springer International Publishing, 2021. p. 317–325. ISBN 978-3-030-92307-5. Citado na página 23.

S., V. et al. Predicting cervical cancer using advanced machine learning algorithms. In: *2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS)*. [S.l.: s.n.], 2024. p. 1600–1604. Citado 3 vezes nas páginas 21, 24 e 77.

SHANDILYA, G. et al. Enhanced cervical cancer diagnosis using advanced transfer learning techniques. In: *2024 2nd World Conference on Communication & Computing (WCONF)*. [S.l.: s.n.], 2024. p. 1–6. Citado na página 14.

SHARMA, V.; KUMAR, A.; SHARMA, K. Digital twin application in women's health: Cervical cancer diagnosis with cervixnet. *Cognitive Systems Research*, v. 87, p. 101264, 2024. ISSN 1389-0417. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1389041724000585>. Citado 2 vezes nas páginas 9 e 21.

SHOLIK, M.; FATICHAH, C.; AMALIAH, B. Deep feature extraction of pap smear images based on convolutional neural network and vision transformer for cervical cancer classification. In: *2024 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*. [S.l.: s.n.], 2024. p. 290–296. Citado 4 vezes nas páginas 18, 24, 74 e 77.

SIMON, H. *Neural networks and learning machines*. [S.l.]: Pearson Education, Inc, 2009. Citado 3 vezes nas páginas 9, 32 e 33.

SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, v. 45, n. 4, p. 427–437, 2009. ISSN 0306-4573. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0306457309000259>. Citado 2 vezes nas páginas 45 e 46.

SONG, Y. et al. Cell classification with worse-case boosting for intelligent cervical cancer screening. *Medical Image Analysis*, v. 91, p. 103014, 2024. ISSN 1361-8415. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1361841523002748>. Citado na página 22.

SZELISKI, R. *Computer vision: algorithms and applications*. [S.l.]: Springer Nature, 2022. Citado na página 33.

TOLLES, W. E.; BOSTROM, R. C. Automatic screening of cytological smears for cancer: The instrumentation. *Annals of the New York Academy of Sciences*, v. 63, n. 6, p. 1211–1218, 1956. Disponível em: <https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-6632.1956.tb32131.x>. Citado na página 1.

UPADHYAY, D. et al. Deep learning-based vggnet, googlenet, and densenet121 models for cervical cancer prediction. In: *2024 5th International Conference for Emerging Technology (INCET)*. [S.l.: s.n.], 2024. p. 1–6. Citado na página 13.

VARGAS-CARDONA, H. D. et al. Artificial intelligence for cervical cancer screening: Scoping review, 2009–2022. *International Journal of Gynecology & Obstetrics*, v. 165, n. 2, p. 566–578, 2024. Disponível em: <https://obgyn.onlinelibrary.wiley.com/doi/abs/10.1002/ijgo.15179>. Citado 3 vezes nas páginas 7, 11 e 48.

VASWANI, A. et al. Attention is all you need. In: GUYON, I. et al. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. v. 30. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>. Citado 4 vezes nas páginas 9, 2, 35 e 36.

WANG, J. et al. Artificial intelligence enables precision diagnosis of cervical cytology grades and cervical cancer. *Nature Communications*, v. 15, n. 1, p. 4369, maio 2024. ISSN 2041-1723. Publisher: Nature Publishing Group. Disponível em: <https://www.nature.com/articles/s41467-024-48705-3>. Citado 2 vezes nas páginas 18 e 24.

WU, T. et al. Artificial intelligence strengthenes cervical cancer screening–present and future. *Cancer Biology & Medicine*, v. 21, n. 10, p. 864, 2024. Citado 3 vezes nas páginas 9, 11 e 48.

WUBINEH, B. Z.; RUSIECKI, A.; HALAWA, K. Classification of cervical cells from the Pap smear image using the RES_dcgan data augmentation and ResNet50V2 with self-attention architecture. *Neural Computing and Applications*, v. 36, n. 34, p. 21801–21815, dez. 2024. ISSN 1433-3058. Disponível em: <https://doi.org/10.1007/s00521-024-10404-x>. Citado na página 19.

WUBINEH, B. Z.; RUSIECKI, A.; HALAWA, K. Segmentation and classification techniques for pap smear images in detecting cervical cancer: A systematic review. *IEEE Access*, v. 12, p. 118195–118213, 2024. Citado 2 vezes nas páginas 11 e 48.

YADAV, S. S.; JADHAV, S. M. Deep convolutional neural network based medical image classification for disease diagnosis. *Journal of Big Data*, v. 6, n. 1, p. 113, Dec 2019. ISSN 2196-1115. Disponível em: <https://doi.org/10.1186/s40537-019-0276-2>. Citado na página 3.

YANG, T. et al. An efficient fusion-purification network for cervical pap-smear image classification. *Computer Methods and Programs in Biomedicine*, v. 251, p. 108199, 2024. ISSN 0169-2607. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0169260724001950>. Citado 3 vezes nas páginas 9, 20 e 24.

YANG, T. et al. A two-stream decision fusion network for cervical pap-smear image classification tasks. *Tissue and Cell*, v. 90, p. 102505, 2024. ISSN 0040-8166. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0040816624002064>. Citado na página 16.

YI, J. et al. Multi-scale window transformer for cervical cytopathology image recognition. *Computational and Structural Biotechnology Journal*, Elsevier, v. 24, p. 314–321, 2024. Citado 2 vezes nas páginas 9 e 14.