Classificação de Imagens Histológicas HE de Câncer Colorretal Utilizando Aprendizado Supervisionado e Profundo

Luiz Moitinho
Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo
São José dos Campos, Brasil

Rosemeri Borges
Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo
São José dos Campos, Brasil

Wellington de Souza Vieira
Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo
São José dos Campos, Brasil

***Resumo*—Câncer colorretal é o terceiro mais comum em homens e o segundo mais frequente em mulheres em todo o mundo. No Brasil, são previstos 21.970 casos em homens e 23.660 casos em mulheres entre os anos de 2023 e 2025, contabilizando um total de 45.630 novos casos. O diagnóstico pode ser realizado por meio do exame histopatológico de biópsias, o qual tem como objetivo avaliar as características celulares e dos tecidos de pacientes. Este trabalho realizou um estudo com o apoio da base de imagens histológicas de câncer colorretal proposto por [9] nomeado de *Enteroscope Biopsy Histopathological H&E Image Dataset* (EBHI) e o UniToPatho presente em [1], a fim de avaliar o desempenho dos classificadores clássicos em conjunto com métodos extratores de características e métodos de aprendizagem profunda para classificar as imagens histológicas de câncer colorretal em normal ou anormal. Os resultados apontaram que a acurácia máxima obtida pelo método *InceptionV3* no conjunto de testes do EBHI foi de 95% enquanto a maior acurácia nos métodos clássicos foi de 71% obtida pelo método BIC combinado ao classificador *RandomForest*. O trabalho concluiu que a classificação foi satisfatória e demonstrou que o EBHI pode ser utilizado para essa atividade, mesmo com um tamanho reduzido de imagens foi possível obter bons resultados.**

***Index Terms*—Câncer; colorretal; classificação; métodos clássicos;aprendizado profundo; EBHI; UNITOPATHO; H&E; e imagens histológicas.**

I. Introdução

Câncer colorretal é constituído por tumores que iniciam no intestino grosso (cólon) a partir de pólipos adenomatosos, os quais são caracterizados como lesões benignas que podem crescer na parede interna do intestino grosso [11]. Este tipo de câncer é o terceiro mais comum em homens e o segundo mais frequente em mulheres em todo o mundo. No Brasil, são previstos 21.970 casos em homens e 23.660 casos em mulheres entre os anos de 2023 e 2025, contabilizando um total de 45.630 novos casos.

Esta doença fornece condições ideais para que seja detectada de forma precoce, pois na maioria dos casos a evolução dos pólipos permeiam durante um período de 10 a 15 anos, permitindo que seja detectável durante uma longa fase pré-clinica [4].

O diagnóstico do câncer colorretal pode ser realizado por meio do exame histopatológico de biópsias, o qual tem como objetivo avaliar as características celulares e dos tecidos de pacientes. Posteriormente, a fim de destacar a estrutura das células e tecidos, os segmentos histopatológicos são tratados e corados com Hematoxilina e Eosina (H&E) [9]. Onde a hematoxilina é constituída de tons de roxo e azul (núcleos e certas estruturas intracelulares) e a eosina é composta por regiões em tons róseos (como por exemplo, o citoplasma) [18].

O processo do diagnóstico dos resultados destes exames é um problema nesta área, uma vez que é altamente subjetivo e complexo em termos de precisão em relação a detecção da região afetada com o câncer, e devido às longas jornadas de trabalhos dos especialistas somado ao volume de exames a serem realizados, permitem que hajam a perda de informações durante a análise destes resultados, e com isto, impactando no diagnóstico, cura e tratamento da doença [9].

Com os tecidos coletados, é possível, então realizar o processo de digitalização para o formato de imagens digitais, as quais possuem informações visuais complexas e que com o apoio de métodos de aprendizado profundo (do inglês, *Deep Learning* - DL) torna-se possível extrair e classificar as suas características. Estas soluções de DL têm apresentado resultados favoráveis que permitem auxiliar patologistas durante o processo de diagnósticos [1].

Para tanto, uma vez apresentado a importância de métodos computacionais para a tarefa de classificação de imagens médicas, o presente trabalho realizou um estudo com o apoio da base de imagens histológicas de câncer colorretal proposto por [9] nomeado de *Enteroscope Biopsy Histopathological H&E Image Dataset* (EBHI) e o UniToPatho presente em [1], a fim de avaliar o desempenho dos classificadores clássicos em conjunto com métodos extratores de características e métodos de aprendizagem profunda para classificar as imagens histológicas de câncer colorretal em normal ou anormal.

II. Trabalhos Relacionados

Entre os trabalhos relacionados ao tema, tem-se o estudo de [12] onde foi desenvolvido um modelo de DL utilizando a ResNet-152 (rede pré-treinada com ImageNet). Neste trabalho foram realizadas as classificações das imagens em seis classes de pólipos colorretais distribuídas em 2.074 *patches* extraídos de 236 imagens de lâminas inteiras (do ingles, *Whole Slide Images* - WSI), atingindo uma acurácia de 93,0% e *recall* igual a 88,3%.

Já em [1] foi proposto uma nova base de imagens intitulada UniToPatho contendo 9.536 *patches* extraídos de 292 WSIs, onde foram utilizados três classificadores ResNet-18 (rede pré-treinada com ImageNet) em cascata a fim de classificar seis tipos de pólipos colorretais e diferentes graus de displasia (alto e baixo), alcançando 67% de acurácia.

O trabalho de [17], obteve uma acurácia geral de 70%, também adotou o classificador ResNet-18 em sua abordagem com o banco de imagens proposto em [1] incrementado de novas imagens, contendo cerca de 457 WSIs.

Por fim, no trabalho de [9], responsável por propor o banco de imagens EBHI, realizou a classificação de cinco tipos de classes e avaliou diversos métodos de extração de características juntamente com métodos clássicos de aprendizado de máquina. Ao final, os dois melhores classificadores obtiveram uma acurácia de 95,3% com o VGG16 e o ANN 76,02%, utilizando como extrator o histograma orientado a gradientes.

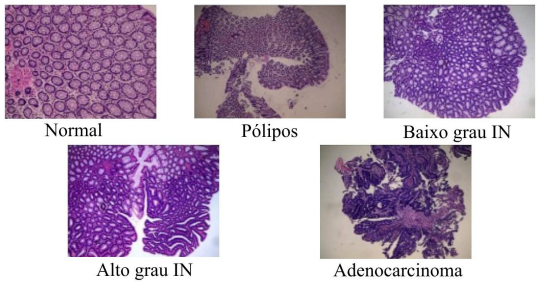
A partir dos trabalhos citados anteriormente é possível notar a contribuição que se tem ao utilizar arquiteturas de aprendizagem profunda para a resolução do problema de classificação de câncer colorretal em imagens histológicas. Contudo, estes estudos apresentam problemáticas em seus trabalhos que dificultam a reprodutibilidade dos experimentos, seja ele para obtenção das bases de imagens, como também durante a avaliação dos modelos utilizados a fim de avaliar o seu desempenho, o que impactam diretamente em seus resultados.

Dito isto, para o desenvolvimento deste trabalho foi realizado um estudo dos métodos de extratores de características, classificadores clássicos e redes de aprendizado profundo adotando a base de imagens disponível em [9], a fim de avaliar uma magnificação acima (400×) da adotada em [9], como também a avaliação de desempenho destes métodos na base proposta por [1].

III. Base de Imagens

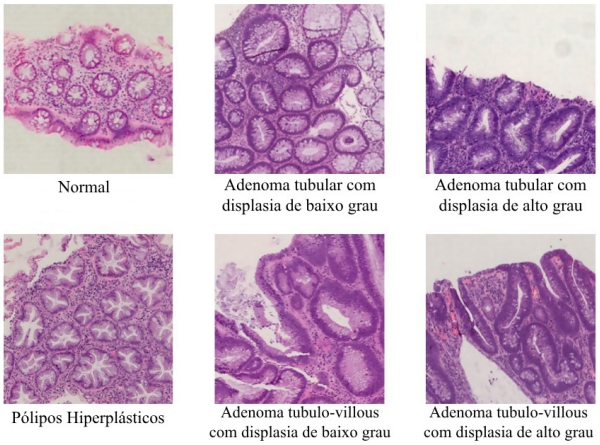
Para o presente trabalho, foram utilizadas duas bases de imagens histológicas de câncer colorretal H&E. A primeira, e principal base adotada durante o estudo foi a EBHI [9], idealizada para permitir que novos pesquisadores experimentassem seus algoritmos de classificação para o diagnóstico de câncer colorretal, que contou com o apoio de dois patologistas do *Cancer Hospital of China Medical University* para realizar a anotação das lâminas. Esta base dispõe de quatro magnificações e em cinco estágios de diferenciação tumoral representando um total de 5.532 imagens histológicas (Fig. 1) com uma resolução de 2048×1536 pixels anotadas em cinco categorias, estando dispostas em quatro magnificações: 40×, 100×, 200× e 400× (Tabela I), sendo a de 400× a magnificação selecionada para realização do estudo.

Figura 1: Exemplo de imagens da base EBHI



A segunda base de imagens foi a UNITOPATHO (Fig. 2) [1], que é uma base de imagens anotadas e em alta resolução, adquiridas através de 292 WSIs obtidas mediante um *scanner Hamamatsu Nanozoomer S210* com magnificação de 20×. Cada lâmina pertence a um paciente com câncer colorretal e foi anotada em seis classes por patologistas. Ao todo, a base é composta por 9.536 *patches* organizados em dois grupos, assim como pode ser visto na Tabela II. O primeiro e adotado neste estudo, os *patches* foram extraídos com um σ igual a 800, resultando em 8.699 imagens com uma resolução de 1.812×1.812 pixels. Já o segundo grupo, é composto por *patches* extraídos com um σ igual a 7.000 que originou cerca de 867 imagens com resolução de 15.855×15.855 pixels.

Figura 2: Exemplo de imagens da base UNITOPATHO



IV. Metodologia

Para este trabalho a base EBHI foi divida em dois grupos de imagens (anormal e normal). As imagens foram submetidas a um processo de classificação em duas abordagens distintas: a primeira utilizando métodos clássicos de extração e classificação e a segunda utilizando aprendizagem profunda por transferência. Na primeira etapa foram selecionados 10 métodos de extração clássicos e 5 classificadores da biblioteca S*cikit-learn*, os métodos e classificadores serão descritos nas

Tabela I: Distribuição das imagens no conjunto EBHI

	Magnificação				
Tipos de imagens	40×	100×	200×	**400**×	Total
Normal	17	29	61	79	186
Pólipos	119	165	254	304	842
Baixo grau IN	204	341	603	660	1.808
Alto grau IN	47	80	130	161	418
Adenocarcinoma	205	471	790	812	2.278
Total	592	1.086	1838	2.016	5.532

Tabela II: Distribuição das imagens no conjunto UNITO-PATHO

Tipos de imagens	σ=7000×	σ=**800**×	Total
Pólipos Hiperplásticos	59	545	604
Normal	74	950	1.024
Adenoma tubular com displasia de alto grau	98	454	552
Adenoma tubular com displasia de baixo grau	411	3.618	4.029
Adenoma *tubulo-villous* com displasia de alto grau	93	916	1.009
Adenoma *tubulo-villous* com displasia de baixo grau	132	2.186	2.318
Total	867	8.669	9.536

subseções seguintes. Na segunda etapa forma selecionados 3 métodos de aprendizagem por transferência o VGG16, o *InceptionV3* e o *Resnet50V2*, eles foram responsáveis pela extração e classificação e também são descritos nas subseções seguintes.

A. Extratores de características

O extrator BIC [23] (do inglês, *Border/Interior pixel Classification*) foi criado por Stehling. Ele classifica os pixels da imagem em pixels de borda ou pixels de interior, esse descritor funciona em três passos. Primeiro o algoritmo de extração quantiza a imagem em 64 níveis (bins). Depois o algoritmo classifica cada pixel da imagem em interior ou borda, e para isto considera os vizinhos do pixel atual. Finalmente, são construídos dois histogramas (um para interior e outro para borda), esses dois histogramas unidos dão origem ao vetor de extração de características de cor de 128 posições.

O ACC [10] (do inglês, *Auto Color Correlogram*) foi criado por Huang et al. em 1997. O auto-correlograma de cor mapeia a informação do espaço de cores através de um auto-correlograma. O auto-correlograma é uma técnica que indica a probabilidade de encontrar dois pixels de cor e distâncias idênticas um do outro. Seus passos são: quantização do espaço de cores, depois cálculo do auto-correlograma para as distâncias escolhidas. No final temos um vetor com os valores de probabilidade para cada distância considerada.

CEDD [2] (do inglês, *Color and Edge Directivity Descriptor*) criado por Chatzichristofis e Boutalis (2008) funciona da seguinte forma: divisão da imagem de qualquer tamanho em 1600 áreas retangulares de imagem. Esses blocos de imagem (*Image-Block*) são então manipulados independentemente para extrair suas informações de cor e textura. Cada *Image-Block* é representado por um vetor quantizado que captura seus atributos de cor e textura. Quando todos os 1600 vetores de bloco de imagem tiverem sido calculados, os mesmos são combinados (fundidos) para formar um único vetor de imagem. O descritor CEED final é produzido pela normalização e quantificação em 8 níveis predefinidos do vetor de imagem acima mencionado.

FCTH [3] (do inglês, *Fuzzy Color and Texture Histogram*) foi criado também por Chatzichristofis e Boutalis (2008) e resulta da combinação de 3 sistemas *fuzzy*. O tamanho do FCTH é limitado a 72 *bytes* por imagem, tornando este descritor adequado para uso em grandes bancos de dados de imagens. É um descritor que combina em seu histograma cor e textura. Inicialmente a imagem é segmentada em um número predefinido de blocos. Cada bloco passa sucessivamente por todas as 3 unidades *fuzzy*. A primeira unidade efetua a extração de um histograma *Fuzzy-Linking* derivado do espaço de cores HSV. Vinte regras são aplicadas em um sistema *fuzzy* de três entradas para gerar eventualmente um histograma de 10 bins. Cada caixa corresponde a uma cor predefinida. Na segunda unidade, propõe um sistema *fuzzy* de duas entradas, a fim de expandir o histograma de 10 bins para um histograma de 24 bins, importando assim informações relacionadas a matiz de cada cor apresentada. Em seguida, na terceira unidade, cada bloco de imagem é transformado com a transformada *Haar Wavelet* e um conjunto de elementos de textura é exportado. Esses elementos são usados como entradas em um terceiro sistema *fuzzy* que converte o histograma de 24 bins em um histograma de 192 bins, importando informações de textura. Nesta unidade, oito regras são aplicadas em um sistema *fuzzy* de três entradas. Com o uso do classificador *fuzzy* de Gustafson Kessel, são modeladas 8 regiões, que são então utilizadas para quantizar os valores dos 192 fatores FCTH no intervalo 0–7, limitando assim o comprimento do descritor em 576 bits por imagem.

GCH [24] (do inglês, *Global Color Histogram*), é um descritor de cores proposto por Stricker e Orengo (1995), muito popular na literatura, devido sua simplicidade e eficiência em operações relacionadas a imagens. Ele retrata a distribuição global das intensidades de uma imagem, considerando seus canais de cores, quando normalizado, fornece a probabilidade de um determinado pixel da imagem ser de uma determinada cor. Seus passos são quantizar a imagem de entrada em uma profundidade parametrizável, e em seguida gerar o histograma da mesma, representando as características de cor da imagem em um vetor de características. A requantizacão é novamente realizada para evitar a alta dimensionalidade gerada.

LCH [22] (do inglês, *Local Color Histogram*) foi criado por Smith e Chang em 1996. Aplica a mesma lógica do GCH, porém em regiões isoladamente (quadrantes). O LCH é um dos descritores mais populares que segue a abordagem baseada em regiões fixas. Seu algoritmo de extração de características divide a imagem em células de tamanho fixo e calcula um histograma de cor para cada célula. Ao final, os histogramas das células são concatenados para formar um único grande histograma gerando uma dimensionalidade mais elevada que o GCH [16].

O *Moments* [6] representa os primeiros 4 momentos es-

tatísticos: Média, Desvio Padrão, Assimetria e Curtose, utilizado para extração de textura.

LBP [7] padrões binários locais: a extração de características invariantes de rotação local ou global tem sido amplamente utilizada na classificação de texturas. Características invariantes locais, por exemplo, padrão binário local (LBP), têm a desvantagem de perder informação espacial global, enquanto características globais preservam pouca informação de textura local. Este método analisa padrões de ocorrência na vizinhança de um dado pixel central em análise.

Gabor [26] utiliza a *wavelet* de Gabor como uma análise de textura muito útil. É um método de recuperação de imagens baseado no filtro Gabor. Os recursos de textura são encontrados calculando a média e a variação da imagem filtrada de Gabor.

Haralick [8] foi criado em 1973 e é um extrator de textura. Haralick propõem 14 medidas estatísticas a serem computadas a partir de matrizes de coocorrência.

B. Classificadores

Para o presente trabalho, os cinco tipos de imagens presentes no EBHI [9] (Tabela I) com a magnificação de 400× e os seis tipos de classes presentes no UNITOPATHO [1] com σ = 800 (Tabela II) foram classificados utilizando métodos clássicos e aprendizagem profunda.

1) Métodos Clássicos: Com o apoio dos descritores de características abordados anteriormente, foram experimentados os métodos clássicos de classificação: *Decision Tree*, *k-Nearest Neghbor* (*k*-NN), *Support Vector Machine* (SVM), *Random Forest* (RF) e *Artificial Neural Network* (ANN).

O *Decision Tree*, é um método de aprendizado supervisionado capaz de prever o valor de uma variável de destino enquanto aprende regras de decisão inferidas a partir de um conjunto de dados, onde cada árvore descreve uma decisão tomada com base em uma determinada característica e os seus ramos refletem os possíveis resultados desta decisão [20]. O *k*-NN é um método de aprendizagem supervisionada comum na área de aprendizagem de máquina, o qual se baseia na busca (utilizando cálculos de distância, como por exemplo euclidiana) por k vizinhos mais próximos do padrão detectado. Para o presente trabalho foi adotado um $k = 3$ [5]. SVM, originalmente é um classificador que tem como objetivo encontrar um hiperplano de segmentação entre duas classes diferentes, podendo resolver problemas de classificação linear e não-linear. Neste estudo, foi adotado o SVM não-linear que utiliza o *kernel* RBF, onde é realizado uma transformação do conjunto de dados não-linearmente separáveis em um espaço de alta dimensionalidade, permitindo a solução do problema de modo linear [14]. O *Random Forest*, baseia-se na combinação de classificadores de árvore onde cada classificador é construído utilizando um vetor aleatório amostrado, de forma independente do vetor de entrada, ao final cada árvore fornece o seu voto para uma classe e a mais votada é atribuída como o rótulo final para o vetor [15]. Por fim, as ANNs são modelos que visam simular a estrutura do cérebro, uma vez que tem como objetivo simular o comportamento humano em diversos processos, entre eles a aprendizagem, associação, generalização e abstração, e são estes processos que permitem a aplicação deste algoritmo na função de classificar imagens [13].

2) Métodos de Aprendizagem Profunda: Métodos de aprendizado profunda também foram utilizados neste trabalho para desempenhar a função de classificadores das imagens histológicas H&E de câncer colorretal, sendo eles: *VGG16*, *InceptionV3* e *ResNet50*.

O VGG16 é considerado uma rede neural convolucional (do inglês, *Convolutional Neural Network* - CNN) foi proposto em [21], onde foi analisado o impacto do aumento do número de camadas para obtenção de melhores acurácias, que por conta dos pequenos filtros convolucionais utilizados (3×3) possibilitou otimização da função de decisão e redução em até 80% da quantidade de parâmetros da rede.

A arquitetura *InceptionV3* é, também, considerada uma CNN e foi proposta no ano de 2015 pelo *Google Research* tendo como objetivo otimizar o problema de eficiência em desempenho de redes de aprendizado profundo. Esta arquitetura é capaz de realizar extrações de características de forma eficiente em diversas escalas e resoluções devido ao *Inpection Modules*, que utiliza diferentes filtros de convoluções (1×1, 3×3 e 5×5) em paralelo para obter estas informações em diferentes escalas [19]. Por fim, a arquitetura *ResNet-50* (do inglês, *Residual Network* 50) é uma CNN de 50 camadas desenvolvida no ano de 2015 pela *Microsoft Research* e que conta com o uso de blocos residuais para resolver o problema de desvanecimento de gradientes em redes profundas ocasionado devido a grande quantidade de camadas e a diminuição do tamanho dos gradientes conforme são utilizadas as camadas mais profundas [19].

V. Análise Experimental

A. Conjunto de dados

O conjunto de dados EBHI I foi dividido de maneiras distintas no aprendizado supervisionado clássico e no aprendizado por transferência. No aprendizado supervisionado clássico o percentual de divisão foi de 80% para treino e 20% para teste. Para as técnicas de aprendizado por transferência foram utilizados um conjunto de validação para ajustar os hiperparâmetros durante o treino, sendo divido em 70% para treino, 15% teste e 15% validação.

O EBHI foi dividido em duas classes e suas imagens distribuídas entre essas categorias conforme abaixo.

1) C01 = Anormal (Alto Grau IN, Adenocarcinoma)
2) C02 = Normal (Normal, Pólipos e Baixo Grau IN)

Para o treinamento utilizando aprendizado por transferência foi feita uma etapa extra de teste utilizando o conjunto UNITOPATHO já apresentado na tabela II. Foram selecionadas 600 imagens dividas em 2 conjuntos conforme abaixo.

1) C01 = Anormal (Adenoma Tubular Displasia de Alto Grau, Adenoma Túbulo-Viloso, Displasia de Alto Grau)
2) C02 = Normal (Normal, Pólipo Hiperplástico e Adenoma Tubular Displasia de Baixo Grau)

O conjunto Adenoma Túbulo-Viloso, Displasia de Baixo Grau não foi utilizado nos experimentos, pela dificuldade em atribuí-lo a uma das classes.

B. Configurações do ambiente

Os experimentos foram executados no ambiente *Google Colaboratory*, com uma GPU T4 com RAM de 15GB e memória RAM do sistema de 12,7 GB. A linguagem de implementação dos códigos executados foi *Java 11* com a biblioteca *JFeatureLib* [6] e *Python* versão 3.10.0. As análises foram feitas utilizando gráficos e funções das bibliotecas, *ScikitLearn*, *Matplotlib*, *Keras* e *Plotly* utilizando *Python*.

C. Critérios de análise

A avaliação dos resultados foi realizada utilizando as métricas de avaliação: acurácia, precisão, *recall* e *f1-score*. A matriz de confusão foi utilizada para visualizar os resultados mais facilmente, ela identifica os casos como verdadeiros positivos (VP), verdadeiros negativos (VN), falsos positivos (FP) e falsos negativos (FN).

Segundo [25] a acurácia busca responder a seguinte questão: qual o percentual de casos verdadeiros em relação a todos os resultados? Sendo assim ela não é sinônimo de precisão e pode ser obtida pela seguinte equação 1:

$$\frac{(VP + VN)}{(VP + VN + FP + FN)} \tag{1}$$

A precisão por sua vez busca responder à questão: Que porcentagem de VPs prevista está correta? [25] A equação é definida como 2:

$$\frac{VP}{(VP + FP)} \tag{2}$$

O *Recall* também chamado de sensibilidade (Retorno) nos responde à questão: Qual a porcentagem de todos os VPs que foram previstos corretamente? [25] Sua equação é 3:

$$\frac{VP}{(VP + FN)} \tag{3}$$

A última métrica utilizada é o *F1-Score* que associa duas métricas quaisquer, em classificadores binários do *machine learning*, a precisão e sensibilidade. É obtido pela média harmônica entre ambas. [25] Sua equação 4:

$$\frac{2 * Precision * Recall}{(Precision + Recall)} \tag{4}$$

D. Resultados e discussão

Os dados do EBHI foram submetidos a duas análises: a primeira utilizou métodos clássicos de extração e classificação sendo aplicados os extratores e classificadores apresentados na seção metodologia. A segunda aplicou métodos de aprendizagem por transferência também apresentados na metodologia.

Na primeira etapa as imagens do conjunto EBHI foram divididas em treino e teste na proporção de 80% para treino e 20% para testes. Essas imagens sofreram processo de extração de características pelos métodos clássicos gerando arquivos do tipo *arff* (*weka*). Esses arquivos foram submetidos a análise utilizando métodos de classificação da biblioteca *Scikit-learn* em *Python*.

Os resultados das acurácias obtidas são apresentados na tabela III. Podemos visualizar que apesar das boas acurácias obtidas pelos métodos nos conjuntos de treino elas não se repetiam no conjunto de teste. Podemos destacar a boa performance geral dos classificadores *Decision Tree*, *RandomForest* e *k*-NN enquanto o SVM não apresentou bons resultados.

Os extratores que apresentaram os melhores resultados foram o GCH, LCH combinados ao *k*-NN e a maior acurácia foi observada na combinação *RandomForest* e BIC com 71%. O extrator GCH teve excelentes resultados com todos os classificadores.

A tabela V apresenta os resultados para a precisão, *recall* e *F1-score*. Nela podemos observar que o *RandomForest* foi muito bem para a maioria dos extratores, com destaque novamente para o BIC com precisão de 81% e *F1-score* de 63%. O SVM novamente não apresenta bons resultados em combinação com nenhum dos extratores.

O extrator Gabor apresenta resultados das métricas bem abaixo dos outros com todos os classificadores.

Tabela III: Resultados dos experimentos clássicos - Acurácia

Descritor	**Decision Tree**		**KNN**		**SVM**		**RandomForest**		**Rede Neural**	
	Treino	***Teste***	***Treino***	***Teste***	***Treino***	***Teste***	***Treino***	***Teste***	***Treino***	***Teste***
BIC	1.00	0.58	0.71	0.51	0.43	0.41	1.00	0.71	0.65	0.51
ACC	1.00	0.52	0.75	0.49	0.46	0.41	1.00	0.70	0.80	0.62
CEDD	1.00	0.53	0.77	0.54	0.62	0.51	1.00	0.62	0.73	0.58
FCTH	0.99	0.53	0.79	0.59	0.66	0.58	0.99	0.64	0.79	0.63
GCH	1.00	0.53	0.86	0.65	0.68	0.60	1.00	0.69	0.91	0.68
LCH	1.00	0.47	0.82	0.62	0.69	0.59	1.00	0.68	0.99	0.68
LBP	1.00	0.39	0.75	0.51	0.65	0.53	1.00	0.54	0.99	0.59
Moments	1.00	0.40	0.71	0.48	0.56	0.52	1.00	0.53	0.54	0.52
Gabor	1.00	0.33	0.65	0.40	0.46	0.44	1.00	0.37	0.46	0.42
Haralick	1.00	0.45	0.70	0.49	0.56	0.53	1.00	0.50	0.55	0.53

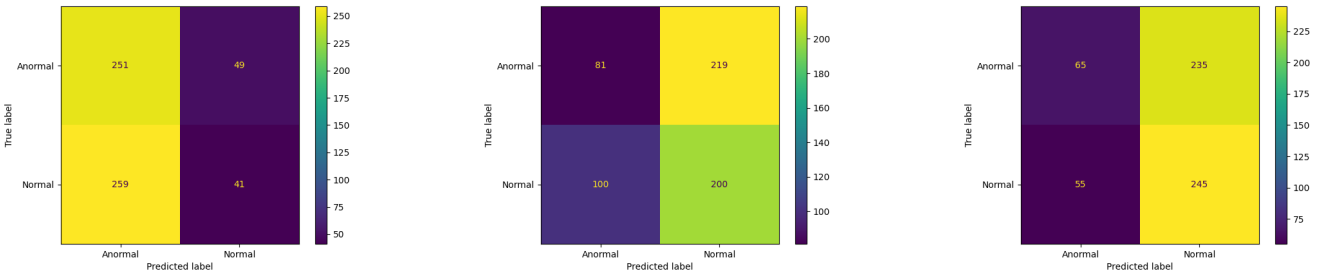
A segunda etapa dos experimentos foi conduzida utilizando aprendizado por transferência, os modelos utilizados foram o VGG16, *Resnet50V2* e o *InceptionV3*. Os dados do EBHI foram divididos nesta etapa em treino (70%), validação (15%) e teste (15%). Foi feito também uma etapa de teste utilizando o conjunto UNITOPATHO aplicando os modelos treinados utilizando o EBHI. Os modelos foram treinados utilizando 10 épocas com função de ativação *sigmoid*, otimizador *Adam* e função de perda entropia cruzada binária. Os resultados são apresentados na tabela IV.

Tabela IV: Resultados dos experimentos aprendizado por transferência no EBHI

Modelo	**Treino**	**Teste**	***Precision***	***Recall***	***F1-score***
VGG16	0.98	0.91	0.91	0.91	0.91
Resnet50V2	0.98	0.92	0.93	0.92	0.92
InceptionV3	0.98	0.95	0.95	0.95	0.95

Tabela V: Resultados dos experimentos clássicos - Métricas

Descritor	***Decision Tree***			**KNN**			**SVM**			***RandomForest***			**Rede Neural**		
	Precision	***Recall***	***F1-score***	***Precision***	***Recall***	***F1-score***	***Precision***	***Recall***	***F1-score***	***Precision***	***Recall***	***F1-score***	***Precision***	***Recall***	***F1-score***
BIC	0.53	0.52	0.52	0.43	0.35	0.36	0.31	0.21	0.16	0.81	0.57	0.63	0.45	0.38	0.40
ACC	0.43	0.42	0.42	0.44	0.37	0.39	0.28	0.22	0.19	0.82	0.52	0.58	0.56	0.51	0.53
CEDD	0.45	0.46	0.45	0.46	0.42	0.43	0.51	0.32	0.32	0.68	0.45	0.48	0.60	0.43	0.46
FCTH	0.47	0.46	0.47	0.55	0.47	0.49	0.56	0.38	0.38	0.66	0.47	0.51	0.62	0.49	0.53
GCH	0.44	0.44	0.44	0.62	0.54	0.57	0.56	0.40	0.39	0.69	0.56	0.59	0.60	0.56	0.58
LCH	0.40	0.40	0.40	0.56	0.52	0.53	0.36	0.39	0.37	0.76	0.51	0.55	0.58	0.57	0.57
LBP	0.30	0.30	0.30	0.44	0.38	0.39	0.33	0.34	0.32	0.67	0.36	0.38	0.63	0.46	0.49
Moments	0.30	0.29	0.29	0.38	0.34	0.35	0.42	0.35	0.34	0.57	0.37	0.38	0.52	0.36	0.35
Gabor	0.23	0.24	0.23	0.23	0.22	0.21	0.17	0.22	0.18	0.23	0.24	0.23	0.16	0.22	0.18
Haralick	0.36	0.37	0.36	0.43	0.34	0.36	0.29	0.30	0.29	0.46	0.33	0.35	0.37	0.31	0.31



(a) VGG16 (b) Resnet50V2 (c) InceptionV3

Figura 3: Matrizes de confusão para teste UNITOPATHO

Os modelos apresentaram excelente acurácia de treino e de teste no conjunto EBHI. O modelo *InceptionV3* apresentou a maior acurácia no conjunto de teste do EBHI, suas métricas de precisão, *recall* e *F1-score* atingiram 95%. Este resultado divergiu do trabalho [9] onde a maior acurácia também foi de 95%, porém o melhor modelo foi o VGG16. No trabalho de [9] a acurácia observada para o *InceptionV3* foi de apenas 72.9%. O modelo *Resnet50V2* deste trabalho obteve acurácia de 92% superior ao modelo *Resnet50* do trabalho [9] onde foi de 83.8%.

As matrizes de confusão para os resultados dos modelos VGG16, *InceptionV3* e *Resnet50V2* aplicados ao conjunto de teste no EBHI são apresentadas na figura 4.

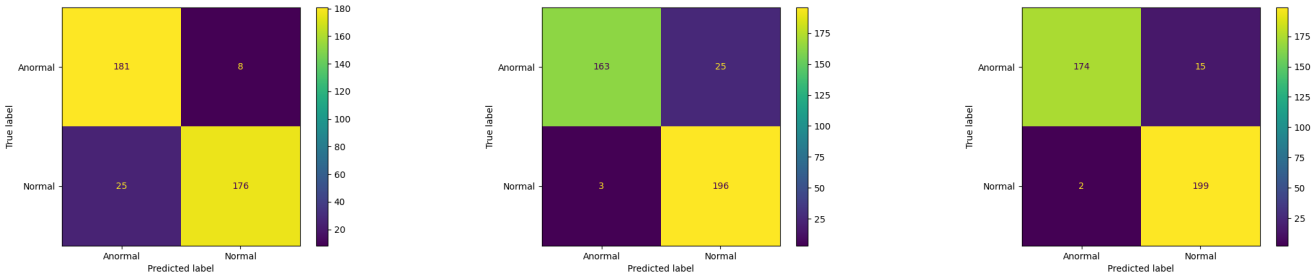
Com os resultados indicando que os métodos de aprendizado por transferência foram satisfatórios. Estes modelos treinados no EBHI foram então extrapolados para teste em um outro conjunto o UNITOPATHO. Para isso, foi selecionada uma amostra de 600 imagens do conjunto. Estas imagens foram submetidas a predição utilizando o modelo obtido na fase anterior.

Os resultados deste teste apresentados na tabela VI, demonstram uma baixa acurácia geral para todos os modelos, sendo o *InceptionV3* o melhor com 51%, enquanto o *Resnet50V2* obteve 46% e o VGG16 com 48%. A tabela ainda mostra resultados interessantes quando a previsão é analisada por classes. Os modelos tiveram resultados bem diferentes: o VGG16 teve uma acurácia de 83% para identificar as imagens anormais e apenas 13% para normais. O *Resnet50V2* obteve 66% de acurácia nas normais e apenas 27% nas anormais. O *InceptionV3* teve um resultado oposto ao VGG sendo eficiente apenas para identificar as imagens normais com acurácia de 81% enquanto para anormais de apenas 21%.

Tabela VI: Resultados dos testes no conjunto UNITOPATHO

Modelo	**Acc Geral**	***Precision***	***Recall***	***F1-score***	**Acc Anormal**	**Acc Normal**
VGG16	0.48	0.47	0.48	0.41	0.83	0.13
Resnet50V2	0.46	0.46	0.46	0.44	0.27	0.66
InceptionV3	0.51	0.52	0.51	0.46	0.21	0.81

As matrizes de confusão do teste com o UNITOPATHO são apresentadas na figura 3. Não foi encontrado um motivo específico para essa diferença no conjunto de teste do UNITOPATHO, possíveis hipóteses seriam a normalização das imagens entre os conjuntos, classes podem não terem sido perfeitamente compatíveis visto que foram criadas e anotadas por trabalhos distintos, resolução das imagens era diferente: no EBHI de 400x enquanto no UNITOPATHO de 800x. Essas hipóteses podem ser testadas em trabalhos futuros.



(a) VGG16 (b) Resnet50V2 (c) InceptionV3

Figura 4: Matrizes de confusão para teste EBHI

VI. Conclusão

Neste artigo trabalhamos o processo de classificação no conjunto de dados EBHI [9]. Foram testados métodos clássicos de extração de características e métodos de aprendizagem profunda baseado em transferência. Os resultados obtidos mostraram que os métodos de aprendizagem por transferência possuem ótima acurácia e são excelentes para realizar esta atividade. A acurácia máxima obtida pelo método *InceptionV3* no conjunto de testes do EBHI foi de 95% enquanto a maior acurácia nos métodos clássicos foi de 71% obtida pelo método BIC combinado ao classificador *RandomForest*. Apesar da acurácia máxima ser similar ao artigo de [9] os métodos foram diferentes, neste artigo o destaque ficou para o *InceptionV3* enquanto no artigo de [9] o melhor foi o VGG16.

O trabalho ainda buscou extrapolar os resultados para outro conjunto de dados o UNITOPATHO [1]. Os modelos treinados no EBHI demonstraram ser capazes de predições neste conjunto. A acurácia observada pelos métodos em cada uma das classes foi muito divergente. O *InceptionV3* conseguiu obter acurácia de 81% no conjunto normal enquanto o VGG16 obteve acurácia no conjunto anormal de 83%. Futuros trabalhos podem explorar estes resultados na busca por uma explicação da discrepância.

Os resultados da classificação foram satisfatórios e demonstram que o EBHI pode ser utilizado para essa atividade, mesmo com um tamanho reduzido de imagens é possível obter bons resultados.

Referências

[1] Carlo Alberto Barbano, Daniele Perlo, Enzo Tartaglione, Attilio Fiandrotti, Luca Bertero, Paola Cassoni, and Marco Grangetto. Unitopatho, a labeled histopathological dataset for colorectal polyps classification and adenoma dysplasia grading. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 76–80. IEEE, 2021.

[2] Savvas Chatzichristofis and Yiannis Boutalis. Cedd: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. pages 312–322, 01 2008.

[3] Savvas A. Chatzichristofis and Yiannis S. Boutalis. Fcth: Fuzzy color and texture histogram - a low level feature for accurate image retrieval. In *2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services*, pages 191–196, 2008.

[4] Geraldo Magela Gomes da Cruz, Renata Magali Ribeiro Silluzio Ferreira, and Peterson Martins Neves. Câncer retal: estudo demográfico, diagnóstico e estadiamento de 380 pacientes acompanhados ao longo de quatro décadas. *Rev. bras. colo-proctol*, pages 208–224, 2004.

[5] Fabio Abrantes Diniz, Thiago Reis da Silva, and Francisco Eduardo Silva Alencar. Um estudo empírico de um sistema de reconhecimento facial utilizando o classificador knn. *Revista Brasileira de Computaçao Aplicada*, 8(1):50–63, 2016.

[6] Franz Graf. Jfeaturelib v1.6.3, October 2015.

[7] Zhenhua Guo, Lei Zhang, and David Zhang. Rotation invariant texture classification using lbp variance (lbpv) with global matching. *Pattern Recognition*, 43(3):706–719, 2010.

[8] Robert M. Haralick, K. Shanmugam, and Its'Hak Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6):610–621, 1973.

[9] Weiming Hu, Chen Li, Md Mamunur Rahaman, Haoyuan Chen, Wanli Liu, Yudong Yao, Hongzan Sun, Marcin Grzegorzek, and Xiaoyan Li. Ebhi: A new enteroscope biopsy histopathological h&e image dataset for image classification evaluation. *Physica Medica*, 107:102534, 2023.

[10] Jing Huang, S.R. Kumar, M. Mitra, Wei-Jing Zhu, and R. Zabih. Image indexing using color correlograms. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 762–768, 1997.

[11] Instituto Nacional de Câncer, INCA. Câncer de intestino, 2022. [Online; accessado em 15 de novembro de 2023].

[12] Bruno Korbar, Andrea M Olofson, Allen P Miraflor, Catherine M Nicka, Matthew A Suriawinata, Lorenzo Torresani, Arief A Suriawinata, and Saeed Hassanpour. Deep learning for classification of colorectal polyps on whole-slide images. *Journal of pathology informatics*, 8(1):30, 2017.

[13] Andrea Martiniano, Ricardo Pinto Ferreira, Arthur Ferreira, Aleister Ferreira, and Renato José Sassi. Utilizando uma rede neural artificial para aproximação da função de evolução do sistema de lorentz. *Revista Produção e Desenvolvimento*, 2(1):26–38, 2016.

[14] Petronio Diego Silva de Oliveira. Uso de aprendizagem de máquina e redes neurais convolucionais profundas para a classificação de áreas queimadas em imagens de alta resolução espacial. 2019.

[15] Mahesh Pal. Random forest classifier for remote sensing classification. *International journal of remote sensing*, 26(1):217–222, 2005.

[16] Otavio Augusto Bizetto Penatti. *Estudo comparativo de descritores para recuperação de imagens por conteudo na web*. PhD thesis.

[17] Daniele Perlo, Enzo Tartaglione, Luca Bertero, Paola Cassoni, and Marco Grangetto. Dysplasia grading of colorectal polyps through convolutional neural network analysis of whole slide images. In *International Conference on Medical Imaging and Computer-Aided Diagnosis*, pages 325–334. Springer, 2021.

[18] Michael H Ross and Wojciech Pawlina. *Histology*. Lippincott Williams & Wilkins, 2006.

[19] Jullyo Emmanuel Vieira Silva. Apoio ao diagnóstico da hanseníase por meio de redes neurais profundas. B.S. thesis, 2023.

[20] Risomario Silva and Darcy Ramos da Silva Neto. Inteligência artificial e previsão de óbito por covid-19 no brasil: uma análise comparativa entre os algoritmos logistic regression, decision tree e random forest. *Saúde em Debate*, 46:118–129, 2023.

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[22] J.R. Smith and Shih-Fu Chang. Local color and texture extraction and spatial query. In *Proceedings of 3rd IEEE International Conference on Image Processing*, volume 3, pages 1011–1014 vol.3, 1996.

[23] Renato Stehling, Mario Nascimento, and Alexandre Falcão. A compact and efficient image retrieval approach based on border/interior pixel classification. pages 102–109, 11 2002.

[24] Markus Stricker and Markus Orengo. Similarity of color images. *Proceedings of SPIE - The International Society for Optical Engineering*, 2420, 03 1995.

[25] Guanis Vilela Junior, Bráulio Lima, Adriano Pereira, Marcelo Francisco Rodrigues, José Ricardo Oliveira, Luis Silio, Anderson Carvalho, Heros Ferreira, and Ricardo Pablo Passos. MÉtricas utilizadas para avaliar a eficiÊncia de classificadores em algoritmos inteligentes. *Centro de Pesquisas Avançadas em Qualidade de Vida*, 14:1, 01 2022.

[26] Dengsheng Zhang, Aylwin Wong, Maria Indrawan, and Guojun Lu. Content-based image retrieval using gabor texture features. 2000.