UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

DOUGLAS WENDER ANTUNES ISIDORO
Orientadora: Profª. Dra. Andrea Gomes Campos Bianchi
Coorientadora: Profª. Dra. Cláudia Martins Carneiro

CLASSIFICAÇÃO AUTOMÁTICA DE RECORTES DE CÉLULAS CERVICAIS COM BASE EM CARACTERÍSTICAS NÃO GEOMÉTRICAS

Ouro Preto, MG
2021

UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

DOUGLAS WENDER ANTUNES ISIDORO

CLASSIFICAÇÃO AUTOMÁTICA DE RECORTES DE CÉLULAS CERVICAIS COM BASE EM CARACTERÍSTICAS NÃO GEOMÉTRICAS

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação.

Orientadora: Profª. Dra. Andrea Gomes Campos Bianchi
Coorientadora: Profª. Dra. Cláudia Martins Carneiro

Ouro Preto, MG
2021

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
DEPARTAMENTO DE COMPUTACAO

FOLHA DE APROVAÇÃO

Douglas Wender Antunes Isidoro

Classificação Automática de Recortes de Células Cervicais com base em Características Não Geométricas

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Ciência da Computação

Aprovada em 22 de Abril de 2021.

Membros da banca

Andrea Gomes Campos Bianchi (Orientadora) - Doutora - Universidade Federal de Ouro Preto
Cláudia Martins Carneiro (Coorientadora) - Doutora - Escola de Farmácia-UFOP
Dayanne Gouveia Coelho (Examinadora) - Doutora - Universidade Federal de Ouro Preto
Guilherme Tavares de Assis (Examinador) - Doutor - Universidade Federal de Ouro Preto

Andrea Gomes Campos Bianchi, orientadora do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 22/04/2021.

Documento assinado eletronicamente por **Andrea Gomes Campos Bianchi**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 22/04/2021, às 15:21, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0160349** e o código CRC **E05E6F4C**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35400-000
Telefone: 3135591692 - www.ufop.br

Agradecimentos

Gostaria de agradecer ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (PIBITI-CNPq), à Universidade Federal de Ouro Preto (UFOP) e à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (PPSUS-FAPEMIG/APQ-03740-17).

Agradeço também à minha mãe Maria das Graças e minha irmã Rita de Cássia por todo apoio durante esse processo e por serem sempre exemplos de determinação.

Agradeço aos amigos Evair Félix, Bárbara Neves, Mateus Honorato, Thiago Roberto, Maycon Amaro, Flávia Nunes, Daniela Pralon, Fábio Augusto, Pedro Machado, Gabriel Rodrigues, Ayron Teodoli, Débora Lage, Julia Azevedo dentre outros por serem companhia e apoio durante esse caminho e por terem contribuído cada um da sua forma para o meu crescimento pessoal.

Agradecimento especial ao meu namorado Paulo Henrique por estar do meu lado sempre e por acreditar em mim mesmo quando eu mesmo não conseguia.

Por fim, gostaria de agradecer às minhas orientadoras Andrea Gomes Campos Bianchi e Cláudia Martins Carneiro pela orientação e por depositarem toda a confiança que sempre depositaram em mim. Vocês foram cruciais para que eu conseguisse chegar até o fim. Muito obrigado também por serem grandes exemplos para mim, levarei os ensinamentos com vocês aprendidos para sempre comigo.

Aos demais amigos, professores e colegas de jornada aqui não citados, obrigado pela companhia.

Resumo

A tecnologia aqui proposta visa possibilitar a análise de amostras de células cervicais de exames de Papanicolaou por metodologias baseadas em aprendizagem de máquina e inteligência computacional. Apresenta-se uma ferramenta computacional para a classificação eficiente de células cervicais baseada em características não geométricas extraídas de áreas nucleares de interesse e ignorando as características relativas ao citoplasma. Essa abordagem baseia-se na hipótese de que os núcleos celulares armazenam grande parte da informação sobre as lesões pré-neoplásicas, além de suas áreas serem mais visíveis mesmo com alto nível de sobreposição celular, fato comum nas imagens de exames de Papanicolaou. Foram utilizados algoritmos de aprendizado supervisionado para o desenvolvimento de um sistema de classificação em duas e três classes. Os experimentos utilizaram, para treinamento e testes, 400 imagens reais de citologia convencional do exame de Papanicolaou obtidas em uma base de dados construída pela equipe de pesquisa. Os resultados demonstraram alto desempenho na classificação com 92,3% de F1 para o sistema de 3 classes e 89,5% de F1 para o sistema de 2 classes. Além disso o sistema também apresenta alta eficiência para aplicabilidade em ambientes realísticos tanto computacionais quando biológicos. A tecnologia poderá auxiliar e melhorar a qualidade dos exames feitos pelos laboratórios, proporcionando maior confiabilidade, efetividade e equidade nos resultados.

Palavras-chave: Reconhecimento de padrões. Características de textura. Célula cervical. Classificação. Esfregaço de Papanicolaou.

Abstract

The technology proposed here aims to enable the analysis of cervical cell samples from Papanicolaou exams by methodologies based on machine learning and computational intelligence. We present a computational tool for the efficient classification of cervical cells based on non-geometric characteristics extracted from nuclear areas of interest and ignoring the characteristics related to the cytoplasm. This approach is based on the hypothesis that cell nuclei store much of the information on pre-neoplastic lesions, in addition to their areas being more visible even with a high level of cell overlap, a common fact in Pap smear images. Supervised learning algorithms were used to develop a classification system in two and three classes. The experiments used for training and tests, 400 real images of conventional cytology of the Pap smear obtained in a database built by the research team. The results demonstrate high performance in the classification and high efficiency for applicability in realistic environments, both computational and biological. The technology can assist and improve the quality of the tests performed by the laboratories, providing greater reliability, effectiveness and equity in the results.

Keywords: Pattern recognition. Texture features. Cervical cell. Classification. Pap smear.

Lista de Ilustrações

Figura 1.1 – Célula isolada (a) e células em sobreposição (b) presentes em esfregaço de Papanicolaou. FONTE: (REZENDE et al., 2020). 2
Figura 2.1 – O hiperplano ótimo separando os dados com a máxima margem ρ , os vetores-suporte e uma distribuição dos dados no R^2 6
Figura 2.2 – Exemplo de padrão binário local para o pixel '7' apresentado em (a). 8
Figura 2.3 – Exemplo de histograma de gradientes orientados. Fonte: (SILVA; TEIXEIRA, 2017). 9
Figura 3.1 – Arquitetura do sistema proposto. 11
Figura 3.2 – Exemplo da etapa de pré-processamento. A numeração representada nesta imagem é irrealista apenas para fins de demonstração, assim como o tamanho dos recortes. 12
Figura 3.3 – Exemplo de comparação entre recortes onde: (A) representa o recorte a partir da posição central de um núcleo; (B) representa o recorte sendo realizado pelo pré processamento naquele instante. (C) representa a interseção entre esses dois recortes que, nesse caso, possui valor maior ou igual a 60% da área individual deles. 13
Figura 3.4 – Sistema de classificação usado. 15
Figura 4.1 – Curva ROC de análise do desempenho do sistema de classificação variando a dimensão dos recortes. 18
Figura 4.2 – Matriz de Confusão para os sistema de classificação em 2 classes. 18
Figura 4.3 – Matriz de Confusão para os sistema de classificação em 3 classes. 19
Figura 4.4 – Exemplo de recortes classificados pelo sistema de 2 classes. FONTE: (REZENDE et al., 2020). 20
Figura 4.5 – Exemplo de recortes classificados pelo sistema de 3 classes. FONTE: (REZENDE et al., 2020). 21

Lista de Abreviaturas e Siglas

ASC-H	Células Escamosas Atípicas, Não Podendo Afastar Lesão de Alto Grau
ASC-US	Células Escamosas Atípicas de Significado Indeterminado
CNN	Rede Neural Convolucional
CRIC	*Center For Recognition and Inspection of Cells*
DECOM	Departamento de Computação
EM	Expectativa-Maximização
FN	Falso Negativos
FP	Falso Positivos
GLCM	Matriz de Co-ocorrência de Nível de Cinza
HOG	Histograma de Gradientes Orientados
HSIL	Lesão Intraepitelial Escamosa de Alto Grau
LBP	Padrões Binários Locais
LDA	Análise Linear Discriminante
LSIL	Lesão Intraepitelial Escamosa de Baixo Grau
NILM	Negativo para Lesão Intra-epitelial ou Malignidade
PCA	Análise de Componentes Principais
QUALICITO	Qualificação Nacional em Citopatologia
RF	Floresta Aleatória
ROC	Característica de Operação do Receptor
SCC	Carcinoma Escamoso
SVM	Máquina de Vetores de Suporte
TN	Verdadeiros Negativos
TP	Verdadeiros Positivos
UFOP	Universidade Federal de Ouro Preto

Sumário

1 Introdução 1
1.1 Objetivos 3
1.1.1 Objetivo Geral 3
1.1.2 Objetivos Específicos 3
1.2 Organização do Trabalho 3
2 Revisão Bibliográfica 4
2.1 Trabalhos Relacionados 4
2.2 Fundamentação Teórica 5
2.2.1 Métricas de Avaliação 9
3 Desenvolvimento 11
3.1 Metodologia 11
3.2 Base de Dados 12
3.3 Pré Processamento e Treinamento de Dados 12
3.4 Extração de Características 14
3.5 Classificação e Testes 15
4 Resultados 17
4.1 Definição de Parâmetros 17
4.2 Experimentos de Classificação 18
4.3 Discussão 19
5 Considerações Finais 22
5.1 Conclusão 22
5.2 Trabalhos Futuros 23
5.3 Premiações e Publicações Realizadas 23

Referências 24

1 Introdução

As análises estruturais e funcionais de esfregaços citológicos e cortes histológicos são comuns em um grande número de campos da biologia. A microscopia ótica tem se mostrado essencial na pesquisa biomédica moderna, seguida por todos os tipos de microscopia e sistemas de diagnóstico por imagem. A avaliação de esfregaços citológicos permite diferenciar os elementos celulares por suas características morfológicas e identificar alterações nucleares e citoplasmáticas compatíveis com diferentes graus de lesões. Além disso, em associação com marcações específicas utilizando anticorpos e sondas de oligonucleotídeos, é possível avaliar aspectos funcionais que incluem expressões gênicas e protéicas em locais específicos nas células. Entretanto, uma limitação das imagens microscópicas reside no fato de que são apenas qualitativas, o que gera informações muitas vezes subjetivas. Investigações quantitativas de amostras celulares e de tecidos têm sido realizadas desde que ferramentas morfométricas apropriadas tornaram-se acessíveis. Seguindo essa linha de raciocínio, este projeto faz parte de uma estratégia de desenvolvimento de instrumentos avaliativos semi-automatizados e automatizados, fruto de uma parceria entre os Departamentos de Análises Clínicas e de Computação da Universidade Federal de Ouro Preto.

Com aproximadamente 530 mil casos novos por ano no mundo, o câncer do colo do útero é o terceiro tumor mais frequente na população feminina, atrás do câncer de mama e do colorretal. Ele é responsável por 265 mil óbitos por ano, sendo a quarta causa mais frequente de morte por câncer em mulheres (World Health Organization, 2018). Porém, essa neoplasia tem desenvolvimento lento, o que eleva a chance de cura quando as lesões precursoras são identificadas precocemente pelo exame citopatológico ou Papanicolaou.

Introduzido pelo médico George Nicholas Papanicolaou, a popularização deste exame deu-se devido ao seu baixo custo e fácil acesso em países em desenvolvimento, além também da sua grande capacidade de diferenciação dos tipos de lesões encontradas, algo extremamente importante para o diagnóstico correto. Apesar dessa técnica ser amplamente utilizada, ela apresenta taxas de resultados falso negativos, falso positivos e insatisfatórios, com causas atribuídas a diferentes etapas do processo. Por isso, no Brasil, os laboratórios de citopatologia seguem a Qualificação Nacional em Citopatologia (QualiCito) que exige a realização do monitoramento interno e externo de qualidade com o objetivo de melhorar o desempenho dos resultados obtidos nesse exame.

Baseado neste objetivo, novas tecnologias, como a citologia em meio líquido, a biologia molecular e a semi-automação da leitura do exame citopatológico, vêm sendo implementadas na área da saúde. Os benefícios da citologia em meio líquido não justificam seu alto custo, o que impede seu uso em larga escala por grandes laboratórios, inclusive aos que servem o Sistema Único de Saúde (SUS) no Brasil. O diagnóstico eficiente a partir da biologia molecular só é

realmente efetivo quando utilizado em conjunto com o exame citopatológico, pois o resultado positivo para HPV requer a confirmação de lesão por meio da análise do esfregaço citopatológico. Dessa forma, a citologia convencional é a técnica escolhida para a realização do exame de Papanicolaou, não excluindo a necessidade de associação com outras estratégias para melhorar a qualidade do rastreio do câncer de colo do útero.

A interpretação citopatológica é realizada por especialistas devidamente habilitados e baseia-se no reconhecimento visual, em microscópio óptico, das alterações do núcleo, citoplasma, e outras informações celulares presentes no esfregaço. Essas alterações são ainda associadas a pareceres clínicos. No entanto, devido à alta de complexidade dos campos analisados microscopicamente, esse processo configura-se como lento, exaustivo e passível de erros. Observa-se na Figura 1.1 que a área do citoplasma é nitidamente visível em A enquanto em B as bordas exatas das áreas do citoplasma são ambíguas. Após a realização da interpretação citopatológica são obtidas imagens ampliadas contendo as características celulares por meio de fotodocumentação dos esfregaços do exame de Papanicolaou.

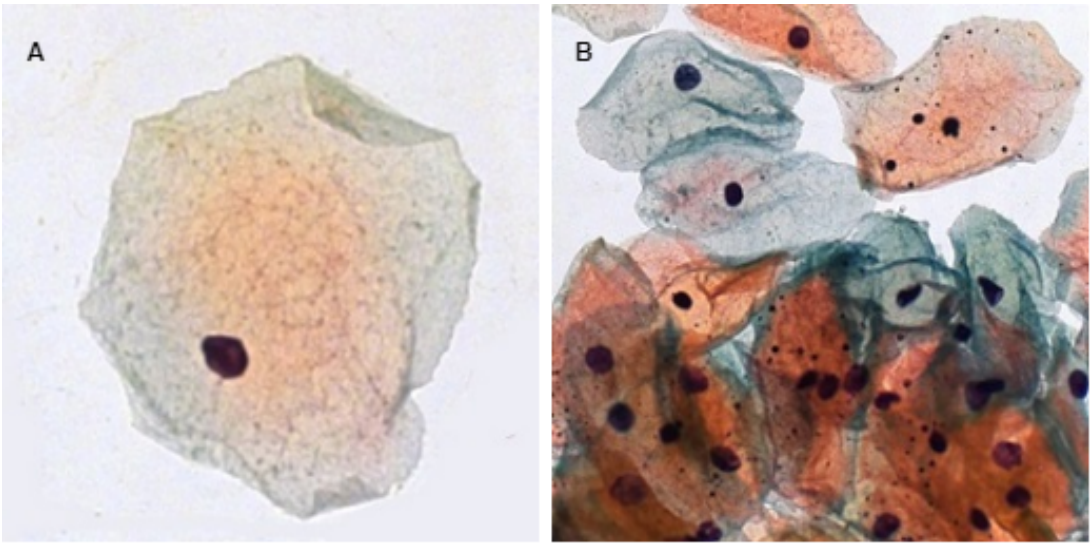

Figura 1.1 – Célula isolada (a) e células em sobreposição (b) presentes em esfregaço de Papanicolaou. FONTE: (REZENDE et al., 2020).

Em imagens contendo sobreposição e aglomerados celulares, a detecção do limite citoplasmático é um problema difícil e, até o momento, não existem métodos eficientes capazes de realizar isso de maneira automática. No entanto, a detecção e a segmentação de núcleos em imagens, contendo determinado grau de sobreposição e aglomerado de células, têm sido abordadas com sucesso por vários estudos (PLISSITI; NIKOU; CHARCHANTI, 2011; PLISSITI; NIKOU; CHARCHANTI, 2011; SOBREVILLA et al., 2010). Os resultados, que utilizam em sua grande maioria imagens oriundas do exame citopatológico do meio líquido, indicam que as características nucleares podem ser suficientes para diferenciar lesões presentes em imagens cervicais. Em relação à classificação das lesões, alguns autores também apresentam bons resultados com núcleos mas utilizando células isoladas (LORENZO-GINORI et al., 2013).

Baseado no que foi anteriormente citado, nota-se que as medidas adotadas para garantia de qualidade do exame citopatológico ainda podem ser exploradas. Portanto, a partir da viabilidade proporcionada pelo emprego de métodos computacionais para aperfeiçoar esse processo, evidencia-se a necessidade de investigar estratégias tecnológicas que busquem aprimorar os serviços de saúde no cenário do rastreio desta neoplasia.

1.1 Objetivos

1.1.1 Objetivo Geral

Desenvolver e validar uma metodologia para a classificação de imagens reais de citologia convencional do colo do útero (exames de Papanicolaou) visando o auxílio à tomada de decisão do profissional quanto ao diagnóstico.

1.1.2 Objetivos Específicos

1. Investigar a extração de características não geométricas das imagens, sem utilizar técnicas de segmentação ou redes neurais convolucionais;
2. Fornecer uma análise sistemática da classificação de imagens reais provenientes do esfregaço de Papanicolaou em sistemas de 2 e 3 classes;
3. Avaliar a importância das informações nucleares para o processo de classificação de imagens de citologia convencional do colo do útero.

1.2 Organização do Trabalho

Este trabalho está organizado da seguinte forma: o Capítulo 2 apresenta alguns estudos baseados na classificação de células cervicais e as referências nas quais se baseia essa pesquisa. O Capítulo 3 detalha a metodologia adotada e descreve todas as suas etapas. Os resultados obtidos são discutidos no Capítulo 4 e as conclusões e considerações finais são expostas no Capítulo 5.

2 Revisão Bibliográfica

Neste capítulo são apresentadas as referências que fornecem embasamento teórico para o trabalho. Na Seção 2.1 são apresentados os trabalhos relacionados ao problema aqui tratado e na Seção 2.2 são apresentadas as definições das técnicas utilizadas.

2.1 Trabalhos Relacionados

Para o cenário dessa pesquisa, a base de imagens pública mais utilizada é a de Herlev, que contém 917 imagens de células cervicais individualizadas e é fornecida pelo Hospital Universitário de Herlev, na Dinamarca (JANTZEN et al., 2005). Seu conjunto de dados consiste em 7 classes de imagens de citologia cervical com única célula, sendo elas: escamosas superficiais, escamosas intermediárias, colunares, displasia leve, displasia moderada, displasia severa e carcinoma in situ. Alguns trabalhos que se baseiam nessa coleção de imagens são mencionados a seguir.

Mariarputham e Stephen (2015) apresentam um sistema de classificação de células, baseado em informações de textura extraídas do núcleo e do citoplasma, utilizando redes neurais e métodos de aprendizado supervisionado. É ressaltado que não existe um conjunto exclusivo de características capazes de fornecer diagnósticos eficientes para todas e quaisquer classes de um sistema de classificação proposto. Além disso, o algoritmo Máquina de Vetores de Suporte (SVM) é apresentado como uma solução eficiente para a classificação.

LAKSHMI e KRISHNAVENI (2014) descrevem um método para extração automatizada de múltiplas características de núcleos e citoplasmas, utilizando função de densidade de probabilidade paramétrica associada ao algoritmo Expectativa-Maximização (EM) e clusterização. Seus resultados confirmam a eficiência do sistema proposto para diferenciar entre Lesão Intraepitelial Escamosa de Baixo Grau (LSIL) e Lesão Intraepitelial Escamosa de Alto Grau (HSIL).

Singh et al. (2015) propõem uma técnica de classificação utilizando Floresta Aleatória (RF). Para isso, consideram características de textura da região do núcleo e do citoplasma, investigam a eficiência dessas características para reconhecimento de células normais e alteradas e apresentam métricas para mensurar a eficiência do sistema.

WALKER et al. (1994) utilizam uma base de dados própria e apresentam resultados preliminares para a classificação de núcleos de células utilizando características texturais da Matriz de Co-ocorrência de Nível de Cinza (GLCM) e Análise Linear Discriminante (LDA) para reduzir a dimensionalidade do espaço de características extraídas. Seus estudos demonstram que as características de textura extraídas da GLCM fornecem um meio eficiente para discriminar células cervicais normais e alteradas.

Todos os trabalhos anteriormente citados utilizam a extração de características de textura

associadas à extração de características geométricas e a um pré processamento que contém segmentação celular. De forma distinta, Plissiti e Nikou (2012) apresentam uma estrutura para a classificação eficiente de células cervicais em categorias normais e alteradas baseado exclusivamente na extração de características de textura da área do núcleo. São utilizados classificadores não supervisionados e técnicas de clusterização não lineares, além de algoritmos genéticos para seleção de recursos. Também é examinada a eficiência de esquemas de redução de dimensionalidade não linear para produzir uma representação precisa das múltiplas características celulares. Este fornece resultados que permitem analisar a eficiência de sistemas baseados na extração exclusiva de características provenientes dos núcleos quando comparados aos sistemas que utilizam extração de características da célula completa.

No estado da arte atual, encontram-se métodos que empregam *Deep Learning*, assim como em Araújo et al. (2019). Este artigo apresenta uma ferramenta computacional para análise citológica utilizando Rede Neural Convolucional (CNN) para a segmentação celular em imagens de esfregaços convencionais de Papanicolaou que contêm alto grau de sobreposição de células. Seus resultados demonstram eficiência para a abordagem proposta, além de robustez para a presença e interferência de neutrófilos, ruídos e outros artefatos muito comuns em imagens de Papanicolau que comprometem a classificação automatizada. Apesar da efetividade, técnicas que utilizam *Deep Learning* e redes neurais, apresentam algumas disfunções: necessitam de um grande número de imagens para treino; apresentam alto custo de processamento computacional; e, principalmente, não permitem a identificação dos atributos utilizados e suas respectivas importâncias. Tais disfunções dificultam a correlação da técnica computacional com o contexto biológico e, dessa forma levam a buscar soluções alternativas, assim como o sistema aqui apresentado.

A partir desses trabalhos foram definidas as características a serem extraídas das imagens, o método de classificação e as métricas de avaliação aplicadas. Diferentemente deles, a nossa proposta utiliza uma base de dados de imagens reais de citologia multicelular e introduz o conceito de recortes nas imagens. Além disso, a segmentação e a extração de características geométricas não são utilizadas, com o objetivo de permitir a classificação de imagens onde não é possível identificar limites citoplasmáticos e nucleares.

2.2 Fundamentação Teórica

Para a classificação, o algoritmo utilizado foi o SVM, que de acordo com SEMOLINI (2002), é uma técnica de aprendizado de máquina fundamentada nos princípios indutivos da Minimização do Risco Estrutural. Ele implementa um mapeamento não-linear (executado por um produto interno kernel escolhido a priori) dos dados de entrada para um espaço característico de alta-dimensão, em que um hiperplano ótimo é construído para separar os dados linearmente em duas classes. Quando os dados de treinamento são separáveis, o hiperplano ótimo no espaço

característico é aquele que apresenta a máxima margem de separação ρ assim como exemplificado na Figura 2.1. Para dados de treinamento em que as amostras das duas classes apresentam superposição (dados não separáveis), uma generalização deste conceito é utilizada.

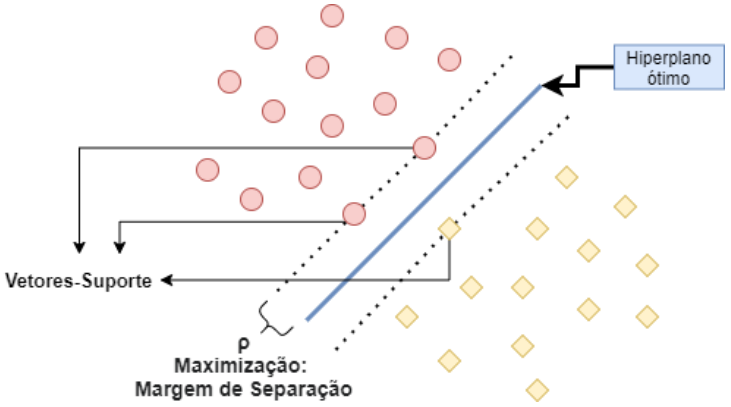

Figura 2.1 – O hiperplano ótimo separando os dados com a máxima margem ρ , os vetores-suporte e uma distribuição dos dados no R^2 .

Foram usadas características extraídas de cada parte da imagem, elas são valores que representam a imagem com um número menor de informações e em sua grande maioria apresentam informações invariantes a iluminação, rotação, translação e outros. As características extraídas foram baseadas em textura, optamos por não incluir as informações geométricas uma vez que não temos as segmentações. As características extraídas associadas ao algoritmo de classificação são:

- Características de Textura de Haralick (HARALICK, 1979):

  Sendo p(i,j) o pixel da linha i, coluna j da imagem p. O número de linhas e colunas variam de 0 até N e M, respectivamente. Sendo assim temos:

  - Momento Angular: $\sum i \sum j p(i,j)^2$;
  - Contraste: $\sum_{n=0}^{N_g-1} n^2 \{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)\}, |i-j| = n$;
  - Correlação: $Corr(p) = \dfrac{\sum i \sum j (ij) p(i,j) - \mu_x \mu_y}{\sigma\sigma_y}$ onde μ_x, μ_y, σ_x e σ_y são as médias e desvio padrão de p_x e p_y, as funções de densidade de probabilidade parcial;
  - Variância: $\sigma = \sum i \sum j \,(i-\mu)^2 \, p\,(i,j)$;
  - Momento de Diferença Inversa: $\sum i \sum j \dfrac{1}{1+(i-j)^2} p\,(i,j)$;
  - Soma de Médias: $\sum_{i=2}^{2N_g} i p_{x+y}(i)$ onde x e y são as coordenadas de uma entrada na matriz de co-ocorrência e $p_{x+y}(i)$ é a probabilidade das coordenadas da matriz de co-ocorrência somarem $x+y$;

  - Soma de Variâncias: $\sum_{i=2}^{2N_g} (i - f_8)^2 p_{x+y}(i)$;
  - Soma de Entropias: $-\sum_{i=2}^{2N_g} p_{x+y}(i) \, log\{p_{x+y}(i)\} = f_8$;
  - Entropia: $H = -\sum i \sum jp(i,j) \, log(p(i,j))$;
  - Diferença de Variâncias: $\sum_{i=0}^{N_g-1} i^2 p_{x-y}(i)$;
  - Diferença de Entropias: $-\sum_{i=0}^{N_g-1} p_{x-y}(i) \, log\{p_{x-y}(i)\}$;
  - Medida de Informação de Correlação 1: $\frac{HXY - HXY1}{max\{HX, HY\}}$;
  - Medida de Informação de Correlação 2: $(1 - exp[-2(HXY2 - HXY)])^{\frac{1}{2}}$ onde $HXY = -\sum i \sum jp(i,j) \, log(p(i,j))$, HX, HY são as entropias de p_x e p_y, $HXY1 = -\sum i \sum jP(i,j) \, log\{p_x(i) \, p_y(j)\}$, $HXY2 = -\sum i \sum jp_x(i) \, p_y(j) \, log\{p_x(i) \, p_y(j)\}$;
  - Coeficiente Máximo de Correlação: raiz quadrada do segundo maior autovalor de Q onde $Q(i,j) = \sum k \frac{p(i,k) \, p(j,k)}{p_x(i) \, p_y(k)}$;

- Momentos de Hu: De acordo com Huang e Leng (2010), Momentos de Hu é um conjunto de 7 números calculados usando momentos centrais que são invariantes às transformações da imagem. Os primeiros 6 momentos provaram ser invariantes à translação, escala, rotação e reflexão. Enquanto o sinal do 7º momento muda para reflexão da imagem. Sendo η_{ij} o momento central normalizado na posição i,j da imagem, os 7 valores são calculados segundo as seguintes fórmulas:

  - $h_0 = \eta_{20} + \eta_{02}$;
  - $h_1 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2$
  - $h_2 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2$
  - $h_3 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2$
  - $h_4 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})\left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right] + (3\eta_{21} - \eta_{03})\left[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right]$
  - $h_5 = (\eta_{20} - \eta_{02})\left[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2 + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})\right]$
  - $h_6 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})\left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right] + (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})\left[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right]$

- Padrões Binários Locais (LBP): Os padrões binários locais segundo Peterson (2009) são um método para extração de descritores e realização de análises de imagens em tons de

cinza. Apesar do algoritmo ser simples, ele fornece excelentes resultados em problemas de classificação de imagens. Em sua versão mais básica o método consiste em:

1. Comparar cada pixel da imagem com seus oito vizinhos, seguindo um movimento circular que sempre parte de uma mesma posição na vizinhança e segue sempre o mesmo sentido. Na Figura 2.2 (a) o pixel central é o 7.
2. Se o pixel central for maior que seu vizinho escreva '1', senão, escreva '0'
3. Com isso, teremos um código com 8 números '0' ou '1', ou seja um número binário de 8 bits representando a variação entre os vizinhos, como pode ser visto na Figura 2.2 (b).
4. Transformando este número binário para decimal obtemos um valor entre 0 e 255.
5. O descritor finalmente é o histograma destes valores decimais atribuídos a cada pixel.

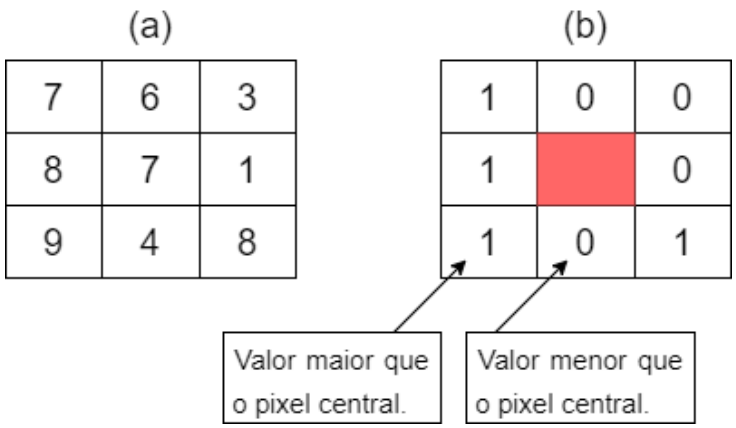

Figura 2.2 – Exemplo de padrão binário local para o pixel '7' apresentado em (a).

- Histograma de Gradientes Orientados (HOG): De acordo com Silva e Teixeira (2017), consiste na ideia de que a aparência de um objeto dentro de uma imagem pode ser descrita pela distribuição de gradientes de intensidade ou direções de bordo. Assim, como exibido na Figura 2.3, a imagem é dividida em pequenas regiões conectadas chamadas células, e para os pixels dentro de cada célula, um histograma de instruções da inclinação é compilado. O descritor é a concatenação desses histogramas. Para melhor precisão, os histogramas locais podem ser normalizados pelo cálculo de uma medida da intensidade através de uma região maior da imagem, denominada um bloco, e, em seguida, utilizando este valor para normalizar todas as células dentro do bloco. Esta normalização resulta em melhor invariância a mudanças na iluminação e sombreamento. O descritor HOG tem algumas vantagens importantes sobre outros descritores. Uma vez que opera em células locais, é invariante a transformações geométricas e fotométricos, com exceção de orientação a objetos. Tais mudanças só aparecem em regiões espaciais maiores.

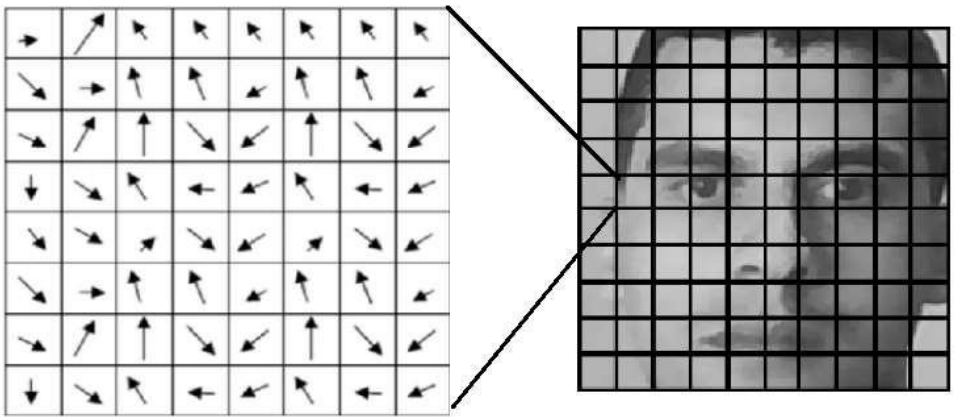

Figura 2.3 – Exemplo de histograma de gradientes orientados. Fonte: (SILVA; TEIXEIRA, 2017).

- GLCM: Segundo (SCHWARTZ; PEDRINI, 2003) uma Matriz de Co-Ocorrência é uma tabulação de quantas combinações diferentes de valores de intensidade dos pixels ocorrem em uma imagem. Ela considera a relação entre dois pixels por vez, um chamado de pixel referência e o outro de pixel vizinho. Cada pixel dentro da imagem torna-se o pixel referência, iniciando no canto superior esquerdo e procedendo até o inferior direito. A co-ocorrência, na sua forma geral, pode ser especificada por uma matriz de frequências relativas P (i, j; d, θ), na qual dois elementos de textura vizinhos, separados por uma distância d em uma orientação θ ocorrem na imagem, um com propriedade i e o outro com propriedade j. Instanciando essa definição para co-ocorrência de níveis de cinza, os elementos de textura são pixels e as propriedades são os níveis de cinza. Em outras palavras a matriz de co-ocorrência pode ser definida como P(i, j, d, θ) onde: i e j são pixels vizinhos de acordo com a distância d. Esta distância entre os pixels i e j é analisada de acordo com uma direção θ .

2.2.1 Métricas de Avaliação

Para a avaliação dos resultados e do desempenho da classificação, algumas métricas foram extraídas. Primeiramente, para cálculo das medidas estatísticas, foi mensurado o número de Verdadeiros Positivos (TP), que se referem aos núcleos alterados corretamente classificados como alterados, Verdadeiros Negativos (TN), que se referem aos núcleos normais corretamente classificados como normais, Falso Positivos (FP), que se referem a núcleos normais erroneamente classificados como alterados e Falso Negativos (FN), que se referem aos núcleos alterados erroneamente classificados como normais. As medidas estatísticas são descritas a seguir:

- Precisão: é a porcentagem de dados que são classificados corretamente para a classe a qual verdadeiramente pertencem. A precisão é definida como:

$$Precisão = \frac{TP}{TP + FP}$$

- Revocação: é a porcentagem dos dados alterados que são corretamente classificados como alterado (verdadeiro positivo). A Revocação é definida como:

$$Revocacao = \frac{TP}{TP + FN}$$

- Especificidade: é a porcentagem dos dados normais que são corretamente classificados como normais (verdadeiro negativo). A especificidade é definida como:

$$Especificidade = \frac{TN}{TN + FP}$$

- F1: é uma relação entre a precisão e a sensibilidade para fornecer uma medição única e balanceada para o sistema. O F1 é definido como:

$$F1 = 2 * \frac{Precisao * Revocacao}{Precisao + Revocacao}$$

3 Desenvolvimento

Neste capítulo, são descritas as fases do desenvolvimento deste projeto. Na Seção 3.1, é apresentada a metodologia adotada na construção dessa estrutura. Na Seção 3.2, é especificado o banco de imagens utilizado. Na Seção 3.3, é descrita a etapa de pré-processamento ao qual são submetidas as imagens. Na Seção 3.4, são apresentadas as características texturais extraídas do conjunto de imagens previamente selecionado. Por fim, na Seção 3.5, é introduzido o método de aprendizado supervisionado utilizado para classificação, além de descrever o sistema de divisão de classes escolhido para esse sistema.

3.1 Metodologia

A arquitetura do sistema de análise proposto está representada na Figura 3.1. As imagens do banco de dados são utilizadas como entrada do sistema e são cortadas em sub-imagens 50x50 pixels. Essas regiões têm suas características extraídas e são usadas para o treinamento de um modelo de classificação supervisionada. Uma parte dos dados é usado para o treinamento e outra para os testes. Finalmente, são realizadas medidas de avaliação sobre os resultados.

Figura 3.1 – Arquitetura do sistema proposto.

3.2 Base de Dados

Todas as imagens utilizadas foram retiradas do banco de dados de imagens pesquisáveis *Center For Recognition and Inspection of Cells* (CRIC) (ARAúJO et al., 2019; REZENDE et al., 2020). Esse conjunto de dados consiste em 11.534 núcleos celulares marcados e classificados a partir de 400 imagens de dimensão 1392 x 1040 pixels. Para a obtenção das imagens, esfregaços de Papanicolaou são digitalizados a partir do uso de microscopia convencional em campo claro com objetiva de 40x e ocular de 10x, por meio de câmera digital Zeiss AxionCam MRc acoplada ao microscópio. Esse banco de dados privado é um dos produtos de uma equipe multidisciplinar de pesquisadores de ciência da computação e biologia que fornece uma coleção de imagens reais de esfregaços de citologia do colo do útero humano representando uma gama de 6 (seis) diferentes lesões em células cervicais, interpretadas por citopatologistas independentes, que são listadas a seguir:

- Negativo para Lesão Intra-epitelial ou Malignidade (NILM);
- Células Escamosas Atípicas de Significado Indeterminado (ASC-US);
- Células Escamosas Atípicas, Não Podendo Afastar Lesão de Alto Grau (ASC-H);
- Lesão Intraepitelial Escamosa de Baixo Grau (LSIL);
- Lesão Intraepitelial Escamosa de Alto Grau (HSIL);
- Carcinoma Escamoso (SCC).

3.3 Pré Processamento e Treinamento de Dados

Conforme mencionado, como não foi utilizada a segmentação dos núcleos e citoplasmas, todas as partes das imagem são utilizadas para o treinamento e a construção do modelo de classificação. Nessa etapa de pré-processamento as imagens do banco de dados são recortadas em 50x50 pixels em toda a sua dimensão, assim como exemplificado na Figura 3.2.

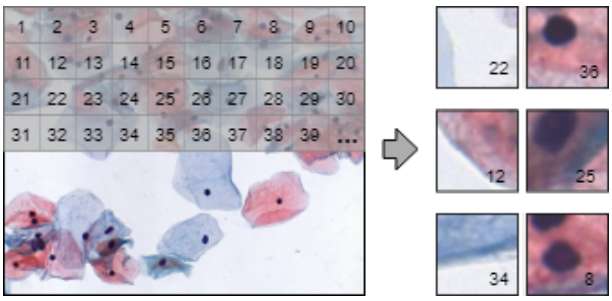

Figura 3.2 – Exemplo da etapa de pré-processamento. A numeração representada nesta imagem é irrealista apenas para fins de demonstração, assim como o tamanho dos recortes.

A realização desses recortes acontece de forma ordenada em relação à imagem, no sentido da esquerda para a direita e de cima para baixo. Porém, é arbitrária em relação à posição exata das células e núcleos, assim como ilustrado na Figura 3.2. Dessa forma as regiões recortadas podem ou não conter núcleo e citoplasmas. A definição das informações para o treinamento são muito importantes pois elas compõem de forma precisa o conjunto de informações referentes a cada lesão e levam a uma boa precisão no resultado. Dessa forma, a etapa de pré processamento resulta em imagens de dimensões reduzida que:

- Compreendem informações nucleares suficientes para serem englobadas por alguma das classes de lesões posteriormente descritas;
- Compreendem fragmentos nucleares insuficientes para serem considerados dentro de alguma classe de lesões e dessa forma são categorizadas como ruídos;
- Compreendem sobreposições celulares e apenas são enquadradas dentro de classes de lesões, se incluírem pelo menos 60% das informações nucleares de alguma das células envolvidas;
- Não compreendem nenhuma informação nuclear e por isso também são incluídas na categoria de ruídos.

Cada imagem gera 588 recortes e para constatar se em um recorte existem informações nucleares suficientes para que ele corresponda a uma das classes nucleares do sistema de classificação, na etapa de treino são utilizados os dados fornecidos pelo banco de dados a respeito da localização dos núcleos dentro das imagens. Com base nas posições dos núcleos, uma comparação é realizada entre recortes feitos a partir das posições centrais de todos os núcleos presentes na imagem e o recorte arbitrário realizado pelo pré-processamento naquele momento, assim como ilustrado pela Figura 3.3. Além disso, esse recorte ao redor da posição (x,y) central de um núcleo da imagem é utilizado como entrada para a etapa de treino e testes, juntamente com os recortes arbitrários obtidos, a fim de permitir que o classificador aprenda também a categorizar núcleos em sua globalidade.

Figura 3.3 – Exemplo de comparação entre recortes onde: (A) representa o recorte a partir da posição central de um núcleo; (B) representa o recorte sendo realizado pelo pré processamento naquele instante. (C) representa a interseção entre esses dois recortes que, nesse caso, possui valor maior ou igual a 60% da área individual deles.

Como consequência disso, é obtido o valor de sobreposição dessas imagens que, caso possua valor igual a zero, indica-se que são recortes de posições que não correspondem a núcleos

de imagens. Caso obtenha valor maior que zero, detecta-se a justaposição dos recortes e no caso específico de possuir valor maior ou igual a 60% da área do recorte, determina-se que o mesmo possui informações suficientes para ser enquadrado como região de núcleo celular e é identificado como um das classes de lesões, diferentes dos ruídos.

A proposta de recortes cria uma nova unidade delimitada na região próxima ao núcleo celular, com menor número de informações, em alternativa para a segmentação que não é utilizada. A dimensão desse recorte foi definida empiricamente a partir de uma série de testes e ajustes de parâmetros onde os melhores resultados foram apresentados pelas imagens na dimensão aqui escolhida. Todos os recortes obtidos a partir do pré processamento são utilizados como entrada para a etapa de classificação. Cabe ressaltar que a etapa de classificação não utiliza essas informações relativas à posição do núcleo, mas sim uma comparação entre os atributos de uma região da imagem com as utilizadas no treinamento.

3.4 Extração de Características

Foi definido um subconjunto de características com capacidade discriminativa a serem extraídas do conjunto de recortes definidos como treinamento. Todas as características selecionadas referem-se a informações relacionadas à textura das imagens e não passam por pré processamento de segmentação para serem obtidas, devido ao objetivo do projeto. São elas:

- 7 Momentos Hu;
- Intensidade Média;
- Valor Máximo de Intensidade;
- Valor Mínimo de Intensidade;
- LBP;
- Características de Textura de Haralick (HARALICK, 1979):

  Momento angular (energia);

  Contraste;

  Correlação;

  Variância;

  Entropia;

  Coeficiente Máximo de Correlação;

  Momento de Diferença Inversa (Homogeneidade);

  Soma das Médias, Soma de Variâncias e Soma de Entropias;

Diferença de Variância e Diferença de Entropia;

Informação Medida de Correlação I e II.

- HOG;
- GLCM;

Essas informações são usadas para a criação de um vetor de características de tamanho 108.

3.5 Classificação e Testes

A classificação manual de todo o conjunto de dados da base utilizada foi feita por especialistas da área e foi utilizada como verdade terrestre. Para a classificação aqui proposta, foram utilizados sistemas de duas e três classes, demonstrados na Figura 3.4 e descritos a seguir:

- **2 Classes:** Nesse sistema, as lesões são agrupadas em duas classes. Na primeira classe apenas as células normais e na segunda as lesões ASC-US, ASC-H, LSIL, HSIL e SCC, formando a classe de células alteradas. A classificação utilizando esse sistema possui como objetivo o ranqueamento de imagens apenas a partir da presença ou ausência de lesão.
- **3 Classes:** Nesse sistema, as lesões são agrupadas em três classes. Na primeira classe apenas as células normais. Na segunda as lesões de gravidade leve ASC-US e LSIL. E por fim, na terceira classe, as lesões de gravidade alta ASC-H, HSIL e SCC. Em todo o restante desse documento, a segunda classe é referida como de lesões leves e a terceira classe como de lesões graves. Essa proposta possui como objetivo o ranqueamento de imagens a partir do diagnóstico e tratamento dado às pacientes detectadas com esses grupos de lesões, onde: pacientes sem lesão (classe normal) após receberem dois resultados negativos do exame, só precisam repeti-lo após 3 anos; pacientes com lesões da classe leve precisam de acompanhamento e devem repetir o exame em 6 meses ou 1 ano; pacientes com lesões da classe grave necessitam de cuidados especiais e devem realizar colposcopia e/ou biopsia.

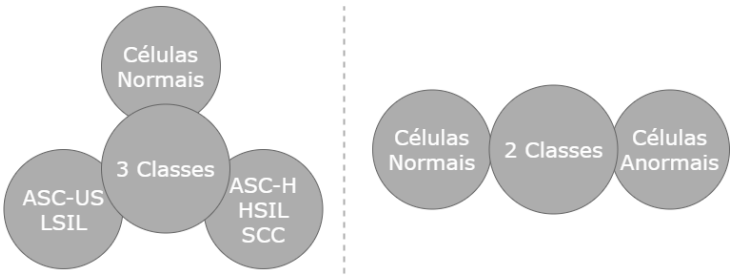

Figura 3.4 – Sistema de classificação usado.

Em ambos os sistemas, ainda é incluído outra classe que engloba os recortes que não possuem informações relevantes para a classificação, como fundo de lâmina ou outros elementos não nucleares. O objetivo dessa classe é permitir que o classificador aprenda a ignorar informações que não são importantes para a classificação celular. Essa classe será referida como ruídos no restante desse documento.

Para a classificação, foi utilizado o algoritmo SVM que é o mais utilizado na literatura apresentada no Capítulo 2 e além disso demonstrou a maior precisão na classificação dos dados e construiu o modelo com um hiperplano que permitiu a melhor separação das classes. Esse método de aprendizado supervisionado é um classificador discriminativo formalmente que recebe como entrada um conjunto de dados e prediz, para cada entrada dada, qual das possíveis classes a entrada faz parte, fazendo dele um classificador linear não probabilístico. Na etapa de treinamento, foram utilizadas todos os recortes de imagens definidos na Seção 3.3. Além disso, é aplicado a validação cruzada K-Fold para garantir a generalidade e imparcialidade dos resultados obtidos.

4 Resultados

Neste capítulo, são apresentados os resultados obtidos nesse trabalho. Na Seção 4.1, é apresentado a definição de parâmetros relativos ao tamanho dos recortes que serão utilizados. Na Seção 4.2, são apresentados os experimentos de classificação realizados e as métricas estatísticas obtidas. Por fim, na Seção 4.3, são trazidas discussões em diferentes vieses a respeito dos resultados obtidos e da aplicabilidade do sistema aqui proposto.

4.1 Definição de Parâmetros

A classificação foi realizada a partir das características de textura extraídas das imagens obtidas a partir do exame de Papanicolaou. Diferentes combinações entre as características foram experimentadas a partir do uso da Análise de Componentes Principais (PCA) como técnica de redução de dimensionalidade. Além disso, diferentes tamanhos para os recortes foram investigados a fim de identificarmos a dimensão capaz de fornecer os melhores resultados. Essa análise é apresentada a seguir na Tabela 4.1, onde são exibidos os valores de precisão, F1, revocação e especificidade para os diferentes tamanhos de recortes.

Tabela 4.1 – Resultados da classificação em 3 classes com diferentes tamanhos de recortes.

Métricas	**30x30**	**50x50**	**70x70**	**100x100**	**150x150**
F1	0,861	**0,923**	0,887	0,882	0,871
Precisão	0,858	**0,921**	0,881	0,877	0,873
Revocação	0,863	**0,926**	0,893	0,887	0,869
Especificidade	0,715	**0,803**	0,760	0,752	0,748

O desempenho do sistema de classificação também foi analisado a partir da variação do seu limiar de discriminação. Sua curva Característica de Operação do Receptor (ROC) é exibida na Figura 4.1 e apresenta a razão de verdadeiros positivos *versus* falsos positivos obtidos. O gráfico indica que o modelo ótimo do sistema é obtido com recortes de tamanho 50x50, pois apresenta maiores valores de sensibilidade e menores valores de *(1-especificidade)*.

Ao comparar espacialmente os recortes, os de dimensão 50x50 também apresentam os melhores resultados. Recortes de dimensão 30x30 não conseguem englobar toda as informações nucleares de uma célula enquanto recortes de dimensão 150x150 acabam englobando muita informação citoplasmática e outros elementos presentes nas lâminas de esfregaços de Papanicolaou. Dessa forma, o recorte de dimensão 50x50, que compreende de forma precisa as informações nucleares das células, é definido como tamanho ideal a ser utilizado pelo sistema.

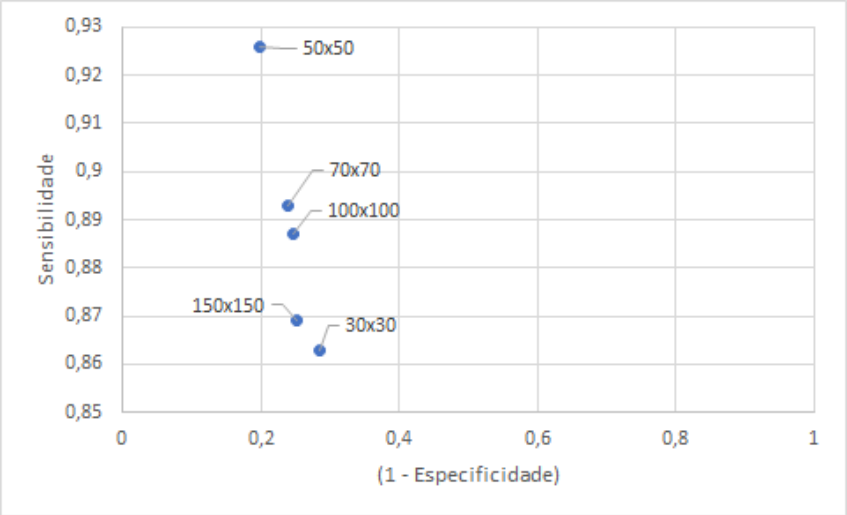

Figura 4.1 – Curva ROC de análise do desempenho do sistema de classificação variando a dimensão dos recortes.

4.2 Experimentos de Classificação

Os melhores resultados foram apresentados utilizando todas as características descritas na Seção 3.2. O procedimento passo a passo da metodologia aqui proposta é exibido na Figura 3.1. Conforme discutido na Seção 3.5, são propostos dois experimentos distintos de classificação, um com duas classes, investigando a presença ou não de lesão na imagem, e outro com três classes, que permite a diferenciação não só entre presença ou ausência de lesões, mas também entre lesões leves ou graves. Para ambos os sistemas, as Figuras 4.2 e 4.3 exibem os resultados de classificação para cada classe do modelo.

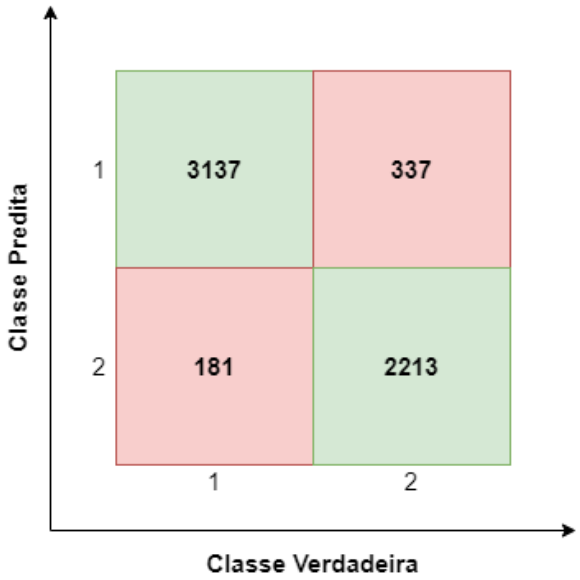

Figura 4.2 – Matriz de Confusão para os sistema de classificação em 2 classes.

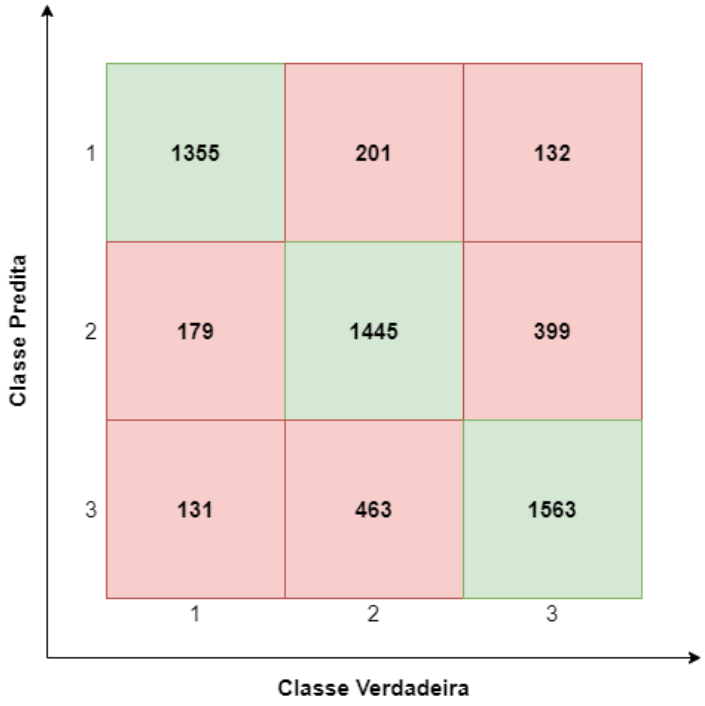

Figura 4.3 – Matriz de Confusão para os sistema de classificação em 3 classes.

O método que se demonstrou mais eficiente para a separação do conjunto de dados foi o SVM com F1 de 89,5% para a classificação em duas classes e 92,3% para a classificação em três classes. Outros resultados obtidos são apresentados na Tabela 4.2. É importante ressaltar que, para cálculo dessas métricas, as informações a respeito da classe de ruídos foram desconsideradas, ou seja, quaisquer valores obtidos a partir da matriz de confusão que envolvam a classe de ruídos não foram contabilizados.

Tabela 4.2 – Resultados da classificação em duas e três classes.

Métricas	**2 Classes**	**3 Classes**
F1	0,895	0,923
Precisão	0,868	0,921
Revocação	0,924	0,926
Especificidade	0,903	0,803

4.3 Discussão

A fim de fornecer auxílio eficiente, com aplicação em ambiente real, na detecção de lesões em imagens obtidas a partir do exame de Papanicolaou, a eficácia dos resultados obtidos foi investigada tanto em viés computacional quanto biológico. Nosso objetivo é fornecer uma ferramenta de classificação ou ranking preliminar das imagens obtidas a partir das lâminas de Papanicolaou. Nessa estrutura, a função do citopatologista continua sendo essencial para oficializar o diagnóstico no cenário de provável presença de lesão.

Nas Figuras 4.4 e 4.5, temos exemplos da aplicação em imagens reais do modelo de classificação para 2 classes e 3 classes, respectivamente, definidos na Seção 3.5. Na Figura 4.4, são indicados em vermelho os núcleos detectados como alterados e em verde os núcleos normais. Na Figura 4.5, são indicados em vermelho os núcleos detectados com lesões graves, em amarelo os núcleos detectados com lesões intermediárias e em verde os núcleos normais. Destaca-se que em ambas as imagens os núcleos demarcados são exemplos parciais retirados da classificação completa das imagens com o objetivo único de representar de forma sucinta as diferentes classificações feitas pela estrutura aqui proposta.

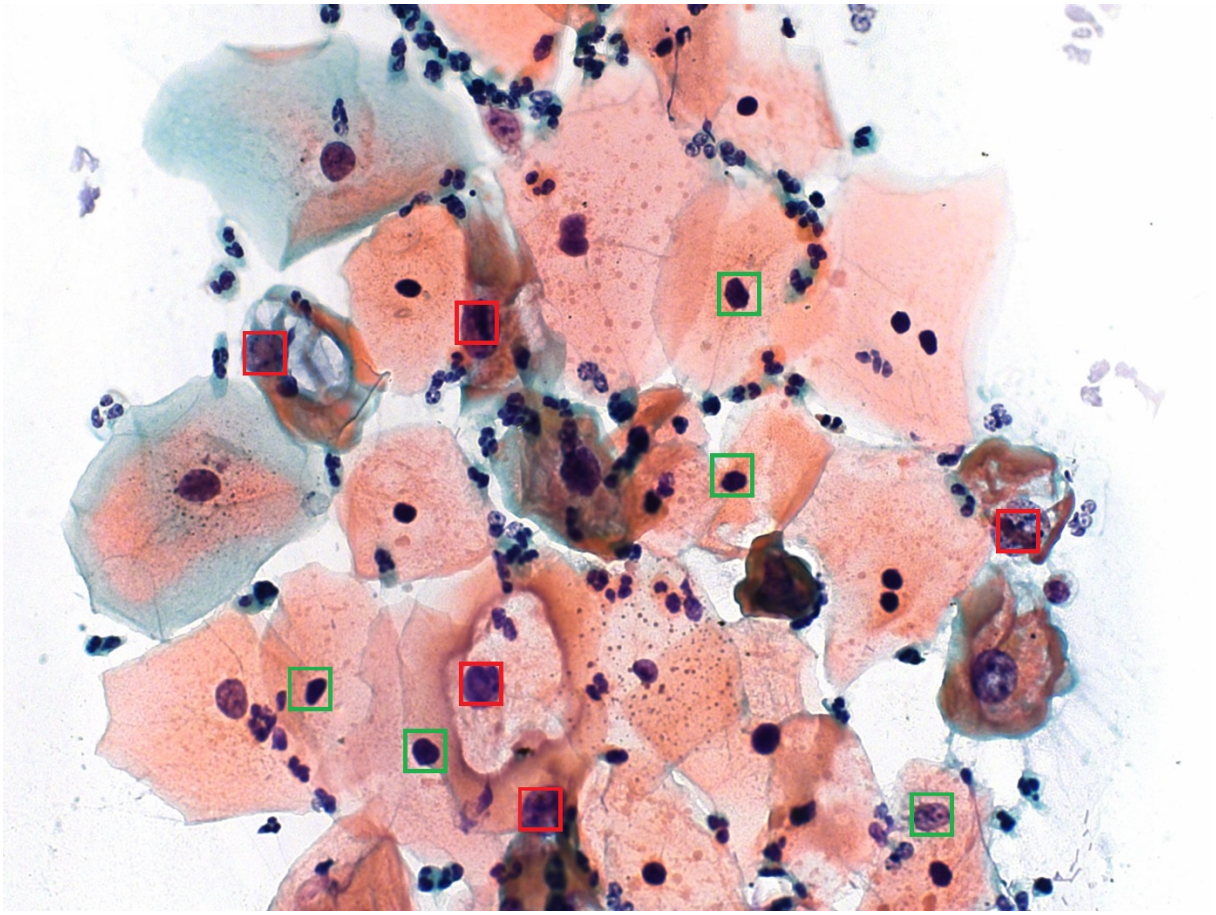

Figura 4.4 – Exemplo de recortes classificados pelo sistema de 2 classes. FONTE: (REZENDE et al., 2020).

Dessa forma, de acordo com os especialistas que realizaram a classificação manual do banco de dados, o principal problema que deve ser combatido dentro da estrutura proposta são as classificações falso negativas, ou seja, não encontrar lesões que realmente existem. Nessas situações, o diagnóstico de uma lesão não é realizado, fazendo com que o paciente crie uma falsa tranquilidade e continue em risco de desenvolver o câncer. Os especialistas ainda explicam que diagnósticos falso positivos não são problemas essenciais no contexto da aplicação real desse sistema, pois essas imagens serão analisadas pelo citopatologista para o diagnóstico final, e caso a lâmina indicada com presença de lesões não as tenham de fato, o diagnóstico será dado como normal e o paciente receberá o diagnóstico correto. Porém, em diagnósticos falso negativos, a

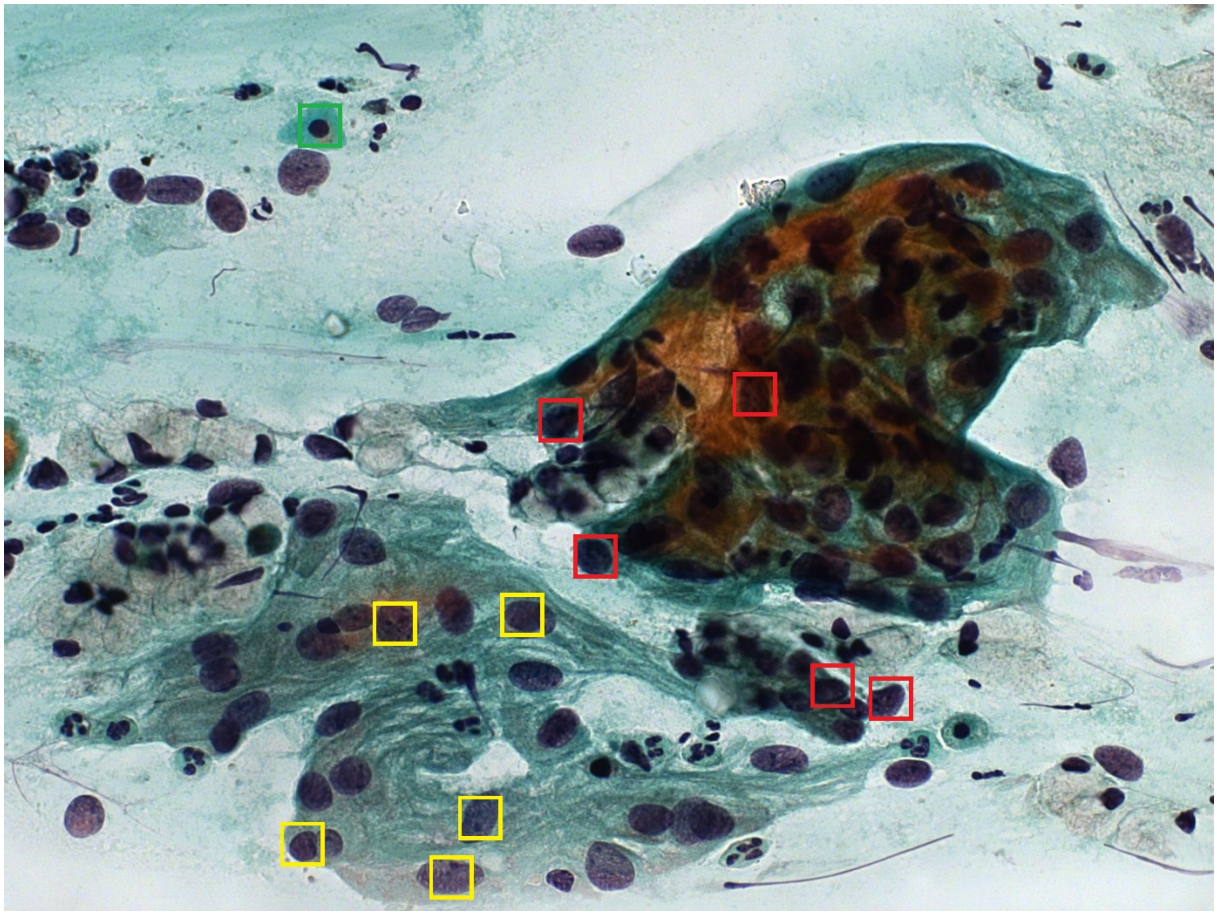

Figura 4.5 – Exemplo de recortes classificados pelo sistema de 3 classes. FONTE: (REZENDE et al., 2020).

lâmina que contém lesões e é classificada como normal não é designada para a avaliação em segunda instância pelo profissional citopatologista. E, dessa forma, o sistema poderia levar a diagnósticos incorretos.

A partir disso, vários experimentos foram realizados e o desempenho da técnica de classificação foi medido variando os parâmetros definidos, a fim de obter o maior decrescimento possível da quantidade de falso negativos atingidos pelo sistema. Os resultados obtidos são satisfatórios para o contexto da aplicação em ambiente real, sendo eles: 5,28% de falso negativos no sistema de 3 classes e 3,08% no sistema de 2 classes.

Vale ressaltar que, ao longo do desenvolvimento desse projeto, as técnicas utilizadas foram escolhidas visando reduzir ao máximo o custo computacional da estrutura proposta. Isso devido ao fato de que, a partir de uma única lâmina de Papanicolaou, são obtidas aproximadamente 10 mil imagens e, como em ambiente realístico existe um demasiado número dessas lâminas para análise, para que a estrutura aqui proposta possa ter aplicabilidade funcional nesse ambiente, a classificação de uma imagem deve acontecer de forma rápida, fazendo com que técnicas elementares destaquem-se em relação à técnicas robustas.

5 Considerações Finais

Neste capítulo, é apresentado uma síntese dos elementos constantes do trabalho e a conclusão a respeito das ideias apresentadas. Na Seção 5.1, é apresentado um resumo geral do assunto pesquisado e uma discussão sobre os resultados obtidos. Na Seção 5.2, são descritos os próximos passos na continuidade do trabalho e, na Seção 5.3, são apresentados os prêmios e publicações realizadas.

5.1 Conclusão

Classificar precisamente núcleos celulares, em imagens reais de citologia convencional do exame de Papanicolaou, é uma condição necessária para fornecer diagnósticos mais precisos e confiáveis. Como nesse tipo de imagens a obtenção de informações a respeito de limites nucleares e citoplasmáticos é complexa e computacionalmente inviável, neste trabalho é apresentada uma metodologia para classificação de imagens reais de citologia cervical a partir da extração de características texturais não geométricas, sem a utilização de aprendizado a partir de redes neurais ou de algoritmos de segmentação.

A metodologia proposta possui aplicação em contexto realístico para auxílio no diagnóstico e rastreio do câncer de colo do útero. Um classificador de aprendizado supervisionado, SVM, foi utilizado para a classificação e a performance do sistema foi medida com imagens reais de citologia cervical, criadas a partir do exame de Papanicolaou, fornecidas por um banco de imagens pesquisáveis privado.

Os resultados obtidos pelos nossos experimentos indicam ótimo desempenho do sistema proposto no processo de categorização de lesões presentes nas imagens do exame de Papanicolaou, apresentando alta precisão e com índice reduzido de falso negativos. Isso implica que a caracterização eficiente de uma imagem real de citologia cervical é viável apenas com o uso de extração características não geométricas.

No contexto médico e biológico, a estrutura proposta serve como um classificador preliminar de esfregaços de Papanicolaou que determina a probabilidade de existirem lesões em uma imagem e as assinala para que um profissional citopatologista as analise, quando necessário, e forneça um diagnóstico definitivo de forma mais prática, eficiente e com garantia de qualidade. Além disso, correlacionando biologia e computação, conclusões pertinentes podem ser inferidas em relação à rigidez ou flexibilidade exigida para que um sistema possa ter aplicabilidade efetiva em ambientes reais.

5.2 Trabalhos Futuros

Como trabalhos futuros, pretende-se investigar a eficiência do sistema proposto quando aplicado em um número maior de imagens reais não catalogadas. Outro fator importante a ser examinado são os conjuntos de descritores, realizando uma investigação mais sistemática para a inclusão de informações que permitam o aumento da precisão do método.

5.3 Premiações e Publicações Realizadas

O presente trabalho foi aceito para apresentação em conferência internacional e recebeu a seguinte premiação:

- Isidoro, D.; Carneiro, C.; Resende, M.; Medeiros, F.; Ushizima, D. and Bianchi, A. (2020). Automatic Classification of Cervical Cell Patches based on Non-geometric Characteristics. In Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP, ISBN 978-989-758-402-2 ISSN 2184-4321, pages 845-852. DOI: 10.5220/0009172208450852.
- Menção Honrosa CNPq na área de Ciências Exatas E Da Terra no XXVII Seminário de Iniciação Científica do Encontro de Saberes da Universidade Federal de Ouro Preto.

Referências

ARAúJO, F. H.; SILVA, R. R.; USHIZIMA, D. M.; REZENDE, M. T.; CARNEIRO, C. M.; BIANCHI, A. G. C.; MEDEIROS, F. N. Deep learning for cell image segmentation and ranking. *Computerized Medical Imaging and Graphics*, v. 72, p. 13 – 21, 2019. ISSN 0895-6111. Disponível em: <http://www.sciencedirect.com/science/article/pii/S089561111830048X>.

HARALICK, R. M. Statistical and structural approaches to texture. *Proceedings of the IEEE*, v. 67, n. 5, p. 786–804, May 1979. ISSN 1558-2256.

HUANG, Z.; LENG, J. Analysis of hu's moment invariants on image scaling and rotation. In: . [S.l.: s.n.], 2010. v. 7, p. V7–476.

JANTZEN, J.; NORUP, J.; DOUNIAS, G.; BJERREGAARD, B. Pap-smear benchmark data for pattern classification. *Nature Inspired Smart Information Systems (NiSIS)*, 01 2005.

LAKSHMI, G. K.; KRISHNAVENI, K. Multiple feature extraction from cervical cytology images by gaussian mixture model. In: *2014 World Congress on Computing and Communication Technologies*. [S.l.: s.n.], 2014. p. 309–311.

LORENZO-GINORI, J. V.; CURBELO-JARDINES, W.; LÓPEZ-CABRERA, J. D.; HUERGO-SUÁREZ, S. B. Cervical cell classification using features related to morphometry and texture of nuclei. In: RUIZ-SHULCLOPER, J.; BAJA, G. Sanniti di (Ed.). *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 222–229. ISBN 978-3-642-41827-3.

MARIARPUTHAM, E. J.; STEPHEN, A. Nominated texture based cervical cancer classification. In: *Computational and Mathematical Methods in Medicine*. [S.l.: s.n.], 2015.

PETERSON, L. E. K-nearest neighbor. *Scholarpedia*, v. 4, n. 2, p. 1883, 2009. Revision #137311.

PLISSITI, M. E.; NIKOU, C. Cervical cell classification based exclusively on nucleus features. In: CAMPILHO, A.; KAMEL, M. (Ed.). *Image Analysis and Recognition*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. p. 483–490. ISBN 978-3-642-31298-4.

PLISSITI, M. E.; NIKOU, C.; CHARCHANTI, A. Automated detection of cell nuclei in pap smear images using morphological reconstruction and clustering. *IEEE Transactions on Information Technology in Biomedicine*, v. 15, n. 2, p. 233–241, March 2011.

PLISSITI, M. E.; NIKOU, C.; CHARCHANTI, A. Combining shape, texture and intensity features for cell nuclei extraction in pap smear images. *Pattern Recognition Letters*, v. 32, n. 6, p. 838 – 853, 2011. ISSN 0167-8655. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167865511000171>.

REZENDE, M. T.; TOBIAS, A. H. G.; SILVA, R.; OLIVEIRA, P.; MEDEIROS, F. Sombra de; USHIZIMA, D.; CARNEIRO, C. M.; BIANCHI, A. G. C. *CRIC Cervix Cell Classification*. figshare, 2020. Disponível em: <https://figshare.com/collections/CRIC_Cervix_Cell_Classification/4960286/2>.

SCHWARTZ, W. R.; PEDRINI, H. Método para classificação de imagens baseada em matrizes de coocorrência utilizando características de textura. *III Colóquio Brasileiro de Ciências Geodésicas, CuritibaPR, Brasil*, p. 1, 2003.

SEMOLINI, R. Support vector machines, inferencia transdutiva e o problema de classificação. *Dissertação (mestrado) - Universidade Estadual de Campinas, Faculdade de Engenharia Eletrica e de Computação, Campinas, SP*, 2002. Disponível em: <<http://www.repositorio.unicamp.br/handle/REPOSIP/262026>.

SILVA, L. dos S.; TEIXEIRA, Y. P. Validação facial com histograma de gradiente orientado. In: . [S.l.: s.n.], 2017.

SINGH, D.; VERMA, A.; ANEJA, M.; SINGH, B. Cervical cell classification using random forests. *International Conference on Medical Physics and Biomedical Engineering*, 07 2015.

SOBREVILLA, P.; MONTSENY, E.; VASCHETTO, F.; LERMA, E. Fuzzy-based analysis of microscopic color cervical pap smear images: Nuclei detection. *International Journal of Computational Intelligence and Applications*, v. 9, p. 187–206, 09 2010.

WALKER, R. F.; JACKWAY, P.; LOVELL, B.; LONGSTAFF, I. D. Classification of cervical cell nuclei using morphological segmentation and textural feature extraction. In: *Proceedings of ANZIIS '94 - Australian New Zealnd Intelligent Information Systems Conference*. [S.l.: s.n.], 1994. p. 297–301.

World Health Organization. *Global Cancer Statistics*. 2018. http://gco.iarc.fr/. Last checked on Oct 23, 2018. Disponível em: <http://gco.iarc.fr/>.