https://eventos.utfpr.edu.br//sicite/sicite2020

Classificação de imagens histopatológicas de câncer de mama usando pequenas subimagens selecionadas

Breast cancer histopathological image classification using selected small patches

Henrique Frederico Trentini
fredtrentini@gmail.com
Universidade Tecnológica Federal do Paraná,Toledo, Paraná, Brasil

Gabriel Fernando Ferrazoli
gabriel@ferrazoli.com
Universidade Tecnológica Federal do Paraná,Toledo, Paraná, Brasil

Jefferson Gustavo Martins
martins@utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Toledo, Paraná, Brasil

Fabio Alexandre Spanhol
faspanhol@utfpr.edu.br
Universidade Tecnológica Federal do Paraná,Toledo, Paraná, Brasil

RESUMO

Neste trabalho foram realizados experimentos com a base de imagens BreakHis aplicando uma rede neural convolucional (CNN) para classificar imagens histopatológicas em tumores benignos ou malignos (câncer). Evitando a necessidade de executar uma segmentação explícita das imagens, este método é baseado na extração de várias pequenas subimagens (*patches*) aleatórias para treinamento e na combinação dessas subimagens para reconhecimento. Visando aumentar o desempenho do modelo classificador foi proposta uma abordagem consistindo em previamente selecionar subimagens que sejam mais representativas de cada classe, permitindo assim discriminar melhor entre padrões malignos e benignos. Os resultados alcançados pela abordagem de filtragem pré-treinamento da CNN mostraram um ganho na acurácia para os dois maiores fatores de aumento disponíveis no conjunto de imagens, 200× e 400×.

PALAVRAS-CHAVE: Câncer de mama. Imagem histopatológica. Reconhecimento de padrões. Rede neural convolucional.

ABSTRACT

In this work, it was conducted experiments on the BreaKHis dataset using a convolutional neural network (CNN) to classify histopathological images into benign tumors or malign tumors (cancer). Avoiding the necessity of performing an explicit segmentation of the images, this method is based on the extraction of several small random subimages (patches) for training, and the combination of these subimages for recognition. In order to increase the performance of the classifier model, an approach was proposed consisting of previously selecting sub-images that are more representative of each class, thus letting better discrimination between malignant and benign patterns. The results achieved by CNN's pre-training filtering approach showed a gain in accuracy for the two highest magnification factors available in the set of images, 200× and 400×.

KEYWORDS: Breast cancer. Histopathologic image. Pattern recognition. Convolutional neural network.

Recebido: 19 ago. 2020.
Aprovado: 01 out. 2020.

INTRODUÇÃO

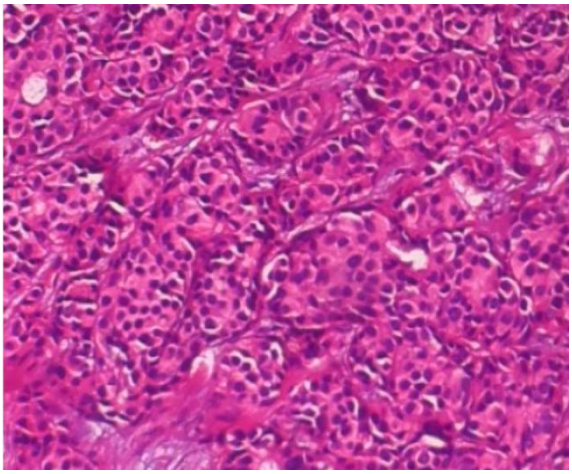
Apesar dos notáveis avanços em seu diagnóstico e tratamento, o câncer continua constituindo um massivo problema de saúde pública em todo mundo. Fatores como o envelhecimento populacional e a adoção de hábitos não saudáveis — antes restritos aos países industrializados — estão contribuindo para o avanço da incidência dessa doença. Somente na última década, houve um crescimento de 20% nos casos de câncer no mundo, segundo o *World Cancer Research Fund* (WCRF, 2020) e até 2030 projeta-se 27 milhões de novos casos de câncer (BOYLE; LEVIN, 2008). Considerados todos os tipos, o câncer é a segunda causa mais comum de mortes em países desenvolvidos e recentemente está tomando o lugar das doenças cardíacas como a principal causa de morte nos países ocidentais (KUMAR *et al.*, 2013). Dados da *International Agency for Research on Cancer* (IARC), da Organização Mundial de Saúde (OMS), confirmam um número global de 9,5 milhões de mortes por câncer em 2018 (BRAY *et al.*, 2018). Além disso, a OMS também projeta 16 milhões de mortes por câncer de 2018 até 2040, sendo os países em desenvolvimento os mais atingidos (BRAY *et al.*, 2018).

No Brasil, o câncer também é um problema de saúde extremamente preocupante. Estimativas do Ministério da Saúde (MS) e do Instituto Nacional do Câncer José Alencar Gomes da Silva (INCA) apontam que 625 mil novos casos de câncer diagnosticados para o triênio 2020-2022 (INCA, 2020). Destes, a maior incidência está nos cânceres de pele (177 mil casos), de próstata (66 mil), de mama (66 mil), cólon e reto (41 mil), pulmão (30 mil) e estômago (21 mil) (INCA, 2020). Dentre todos os tipos de câncer, excluindo o câncer de pele, o câncer de mama é o segundo mais comum entre as mulheres (BRAY *et al.*, 2018). Embora o câncer de mama ainda seja mais prevalente nas regiões mais desenvolvidas, a mortalidade é relativamente maior nos países menos desenvolvidos devido a dificuldade de diagnóstico precoce e restrições enfrentadas pelas mulheres no acesso aos avanços clínicos de combate à doença.

A detecção e o diagnóstico do câncer de mama podem ser feitos através de métodos não invasivos e biópsia. Métodos não invasivos são basicamente procedimentos de imagem: mamografia (raios-x), imagem por ressonância magnética (MRI) das mamas, ultra-som (sonografia) e termografia. Apesar do uso de técnicas de imagem para diagnóstico do câncer estar difundido, a biópsia é único meio de informar, com segurança, se o câncer está realmente presente. Dentre as técnicas de biópsia, destacam-se procedimentos como aspiração por agulha fina (FNA), biópsia de agulha grossa (CNB), biópsia mamária assistida à vácuo (VABB) e biópsia cirúrgica (SOB)(KUMAR *et al.*, 2013). Os procedimentos de biópsia coletam amostras de células ou tecido. Tais amostras devem ser fixadas em uma lâmina para microscopia para a subsequente coloração e análise microscópica. Os patologistas usam os benefícios de uma ampla variedade de corantes para obter informações úteis sobre as lesões e a composição dos tecidos. Na Figura 1 é mostrado um tumor mamário maligno (câncer) com as estruturas celulares evidenciadas pela hematoxilina e eosina (HE), uma combinação de corantes usada “rotineiramente” em amostras de tecido para revelar as estruturas subjacentes e sua condição (HERRINGTON, 2014). Em resumo, o diagnóstico a partir de imagens histopatológicas permanece sendo o “padrão-ouro” para

diagnosticar a maioria dos tipos de câncer, incluindo o câncer de mama (RUBIN *et al.*, 2012).

Figura 1 – Detalhe de uma seção de carcinoma ductal corada com HE (aumento de 100×).



Fonte: SPANHOL (2018).

Apesar da relevância, a análise patológica ainda é bastante manual e subjetiva, dependente do especialista humano. Dado o crescente volume de casos a serem avaliados, principalmente de câncer, almeja-se que sistemas automatizados possam auxiliar o patologista na tarefa de classificação dessas doenças em menor tempo e com maior acurácia no diagnóstico. Neste contexto, considerando o impacto do câncer na saúde pública, especialmente o câncer de mama pela incidência e letalidade na população feminina, somado a urgência de prover ferramentas de suporte ao patologista, propomos um modelo de classificador que possa identificar o câncer de mama através da análise de imagens digitalizadas de lâminas histopatológicas apoiando a decisão do profissional médico.

MATERIAIS E MÉTODOS

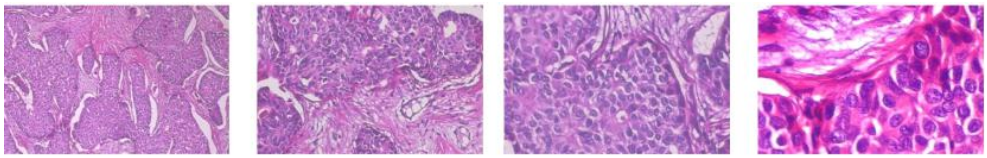
Infelizmente, há uma falta de bancos de images histopatológicas públicos e abrangentes destinados a pesquisa em sistemas de diagnóstico assistido por computador CAD/CADx. Em uma revisão recente, (VETA *et al.*, 2014) destacaram que o maior obstáculo no desenvolvimento de novos métodos de análise de imagens histopatológicas é a falta de grandes bases de dados públicas rotuladas por especialistas. Bases rotuladas também são essenciais para desenvolver e validar sistemas de aprendizado de máquina.

BASE DE IMAGENS BREAKHIS

Para mitigiar a escassez de bases de dados públicas de imagens histopatológicas destinadas à pesquisa, em (SPANHOL *et al.*, 2016) foi disponibilizada uma nova base de imagens chamada BreaKHis, a qual é composta de imagens microscópicas de lâminas de tumores mamários. As imagens do conjunto estão divididas em tumores malignos e benignos, coletadas usando quatro diferentes fatores de aumento: 40×, 100×, 200× e 400×. Uma imagem representativa de cada fator de aumento pode ser vista na Figura 2. São

mostradas áreas distintas, pertencentes a uma mesma lâmina de um tumor mamário maligno (corado com HE), capturadas em diferentes fatores de aumento: (a) 40×, (b) 100×, (c) 200× e (d) 400×.

Figura 2 – Exemplos de imagens da base BreaKHis.



Fonte: SPANHOL (2018).

Tal base de dados foi construída em colaboração com o laboratório P&D[1] – Anatomia Patalógica e Citopatologia, Cascavel, Paraná, Brasil. A base BreaKHis é licenciada sob licença *Creative Commons* 4.0 e está disponível no repositório do Laboratório de Visão Robótica e Imagem (VRI)[2] da Universidade Federal do Paraná (UFPR), através de requisição, para propósito de pesquisa. Atualmente conta com mais de 1200 usuários registrados pelo mundo e mais de 350 citações em publicações acadêmicas indexadas. A base BreaKHis foi utilizada nos experimentos.

ABORDAGEM APRENDIZADO PROFUNDO

A primeira abordagem baseada em aprendizado profundo (*deep learning*) usando a base BreaKHis foi originalmente publicada em (SPANHOL *et al.*, 2016b) . Em tal trabalho, os autores apresentaram os resultados alcançados usando uma CNN (*Convolutional Neural Network*) treinada diretamente com as imagens histopatotógicas da BreaKHis, considerando os 4 aumentos e 5 distribuições diferentes de pacientes (partições) nos conjuntos de treinamento e teste.

CLASSIFICAÇÃO USANDO SUBIMAGENS

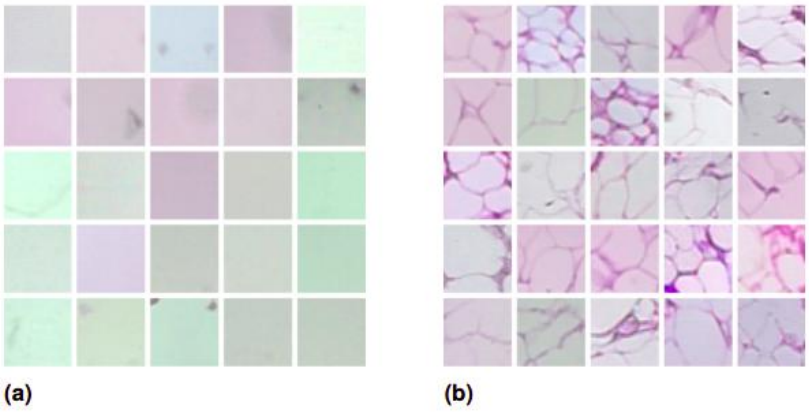
Dado que CNNs exigem grandes conjunto de dados para treinamento, os autores utilizaram a técnica de extrair pequenas subimagens (*patches*) das imagens originais, tanto nas fases de treinamento quanto de teste (SPANHOL *et al.*, 2016b). A ideia é aumentar o conjunto de instâncias disponíveis para treinamento extraindo de cada imagem original do conjunto de treinamento um grande número de *patches* selecionados de posições randômicas. Na fase de testes os *patches* extraídos das imagens originais do conjunto de teste são classificados individualmente e então o resultado é combinado para classificar a imagem original como sendo tumor maligno (câncer) ou benigno.

Entretanto, os padrões de alguns *patches* é muito similar, independente se a imagem original provém de um tumor maligno ou benigno. Exemplos de tais padrões podem ser vistos na Figura 3: em (a) *patches* de áreas de fundo e (b) *patches* de tecido adiposo. Essa intersecção pode ser parcialmente explicada pela presença natural de certos tipos de tecido (como tecido adiposo, tecido rico em

[1] https://www.prevencaoediagnose.com.br/
[2] https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/

colágeno, tecido conectivo, etc.) em muitas das amostras histopatológicas de mama.

Figura 3 – Exemplos de padrões recorrentes



Fonte: SPANHOL (2018).

Além disso, um significante número de *patches* extraídos apresenta texturas quase planas, normalmente correspondendo a áreas de formação de líquido, à cavidade central de estruturas tubulares ou mesmo a espaços no tecido (fundo). Definitivamente os *patches* contendo apenas tal tipo de padrão plano não são representativos para distinguir entre tumores malignos e benignos. Assim, a proposta é descartar os *patches* que apresentem padrões comuns a ambas as classes. Para tanto, foi aplicada uma técnica de agrupamento (*clustering*) em tais *patches* visando separá-los em três grupos: puramente maligno, puramente benigno e misto. Logo, apenas os grupos puros são considerados como entrada de treinamento para a CNN. Espera-se que usando *patches* mais discriminativos melhore-se a taxa de reconhecimento do modelo de classificação.

EXTRAÇÃO DE CARACTERÍSTICAS *DECAF*

Na extração de características, usando os *patches* já obtidos, é aplicada uma CNN pré-treinada para obter características DeCAF (*Deep Convolutional Activation Feature*) a partir da camada ip1 dessa CNN. Essa camada é uma camada *InnerProduct* (camada totalmente conectada) que trata a imagem de entrada como um vetor e gera como saída também um vetor. Neste modelo tem-se um vetor v_i de 64 dimensões que é salvo em V.

AGRUPAMENTO DE *PATCHES*

Uma vez extraída a representação a partir da camada ip1, aplicou-se um algoritmo de agrupamento *k-means* para criar k = 50 agrupamentos no conjunto de dados de entrada. Foram avaliados variações do número de agrupamentos k tal que $k \in \{10, 20, 50\}$. Analisando a distribuição dos *patches* nos respectivos *clusters* foi escolhido o valor de k = 50 para manter os agrupamentos menos densos. Foi executado um agrupamento particionado, isto é, a divisão do conjunto de instâncias de entrada (vetores de características) em subconjuntos não sobrepostos (agrupamentos). Espera-se dividir os *patches* em grupos distintos,

baseando-se nas propriedades de coerência e similaridade aprendidos automaticamente pela CNN DeCAF.

FILTRAGEM

Este passo objetiva avaliar cada *cluster* (agrupamento) e descartar aqueles que eventualmente possuam padrões mistos. Assim, usando os vetores em V, cria-se os *clusters* C e avalia-se a taxa de pureza p^C para cada *cluster*, descartando os *clusters* que apresentem pureza inferior a um limiar fixado. A taxa de pureza p^C é computada pela Eq. (1). Aqui o termo pureza representa a qualidade de um *cluster* possuir majoritariamente *patches* provindos apenas de uma determinada classe, isto é, minimize a ocorrência de padrões que sejam recorrentes em ambas as classes.

$$p^{C_q} = max\left(\frac{|L_1|}{N_P}, \ldots, \frac{|L_m|}{N_P}\right) \quad \text{tal que, } \forall b, L_b \in C_q \tag{1}$$

Tem-se na equação (1) que $1 \leq q \leq k$, e $L_b \in C_q$ representa os *patches* de L_b no cluster C_q. Dado um limiar λ, é descartado o *cluster* C_q se $p^{Cq} < \lambda$. Finalmente, para cada *cluster* remanescente, é atribuído o rótulo do *cluster* aos repectivos *patches*. Tais *patches*, presumidamente contendo padrões mais representativos, são usados como entrada para treinar a CNN. Foram avaliados diferentes valores da taxa de pureza considerando valores distintos para o limiar λ. Notou-se que assumindo taxas muito altas ($\lambda \geq 0{,}99$) causava-se a eliminação completa dos *patches* de certos conjuntos. Assim, foi fixado um limiar $\lambda = 0{,}9$ que produziu resultados de classificação satisfatórios.

TREINAMENTO DA CNN

Finalizada a filtragem, os *patches* menos discriminativos foram descartados e um conjunto de instâncias menor é obtido para treinar a CNN. Esse modelo foi treinado usando os *patches* filtrados como entrada. O protocolo de treinamento é o supervisionado e aplicou-se o método SGD (*Stochastic Gradient Descent*) para computar os gradientes. Mini-lotes de tamanho 1 foram usados para atualizar os parâmetros da CNN, treinada por 80.000 iterações, iniciando com uma taxa de aprendizado de 10^{-6} em conjunção com um termo *momentum* de 0,9 e um decaimento de peso 4^{-5}.

MÉTRICAS DE AVALIAÇÃO

Para permitir uma comparação direta dos resultados deste trabalho principalmente foi avaliada a acurácia em cada fator de aumento (*zoom*) de forma independente, tanto na métrica nível de imagem quanto em nível de paciente. A acurácia em nível de imagem corresponde ao escore do total de imagens corretamente classificadas. Isto é mostrado na Eq. (2), onde N é o número total de imagens no conjunto de dados e N_c é o total de imagens corretamente classificadas:

$$Ac_{im} = \frac{N_c}{N} \tag{2}$$

Já a acurácia em nível de paciente corresponde a média da acurácia em nível de imagem por paciente. Mais formalmente, fazendo S o número total de pacientes, N_c o total de imagens classificadas corretamente do paciente ϕ e N_ϕ o total de imagens para o mesmo paciente tem-se a Eq. (3):

$$Ac_{pac} = \frac{\sum_{\varphi=1}^{S} \frac{N_c^{\varphi}}{N_\varphi}}{S} \tag{3}$$

RESULTADOS E DISCUSSÃO

Desde a disponibilização pública da base BreaKHis alguns métodos de classificação que utilizam tal conjunto de imagens foram propostos na literatura. Logo, considerando a base BreaKHis um *benchmark* e os resultados alcançados com descritores convencionais publicados em (SPANHOL *et al.*, 2016) uma linha base (A), podemos comparar o método de CNN sem filtragem (B e B*) (SPANHOL *et al.*, 2016b) com o método de filtragem de *patches* significativos aqui apresentado (C). A arquitetura da CNN é a mesma. Assim, a Tabela 1 compara a acurácia do método de descarte de *patches* proposto neste trabalho com as outras abordagens usando as mesmas 5 partições de treinamento-teste da linha base. A acurácia percentual é a média das 5 partições. Os melhores resultados estão em negrito com fundo cinza. A abordagem B* indica uma combinação de classificadores.

Tabela 1 – Acurácia (%) comparada com trabalhos disponíveis na literatura

Abordagem		Fator de aumento			
		40X	100X	200X	400X
paciente	A	83,8 ± 4,1	82,1 ± 4,9	85,1 ± 3,1	82,3 ± 3,8
	B	88,6 ± 5,6	84,5 ± 2,4	85,3 ± 3,8	81,7 ± 4,9
	B*	**90,0 ± 6,7**	**88,4 ± 4,8**	84,6 ± 4,2	**86,1 ± 6,2**
	C	86,4 ± 5,7	83,6 ± 5,8	**92,1 ± 7,3**	85,0 ± 4,7
imagem	A	82,8 ± 3,6	80,7 ± 4,9	84,2 ± 1,6	81,2 ± 3,6
	B	**89,6 ± 6,5**	**85,0 ± 4,8**	84,0 ± 3,2	80,8 ± 3,1
	B*	85,6 ± 4,8	83,5 ± 3,9	83,1 ± 1,9	80,8 ± 3,0
	C	85,3 ± 3,3	82,5 ± 3,0	**87,8 ± 4,9**	**82,1 ± 3,4**

Fonte: Autoria própria (2020).

Nota-se que o método de filtragem não conseguiu um desempenho superior em todos os fatores de aumento, mas sim nos dois maiores (200× e 400×) quando comparado com CNN que utiliza os patches sem filtragem. Por outro lado, o método proposto tem um desempenho melhor em todos os fatores de aumento quando comparado a linha base.

CONCLUSÕES

A técnica de descartar *patches* não discriminativos mostrou-se promissora, com resultados que melhoram o desempenho da CNN na tarefa de classificação de imagens da base BreaKHis. Contudo, novos experimentos precisam ser realizados para melhorar o desempenho nos aumentos menores (40× e 100×).

Finalmente, ressalta-se que o projeto foi parcialmente impactado em seu cronograma pela suspensão das atividades presenciais no início de março de 2020 devido às medidas de isolamento social instituídas para tentar reduzir o avanço da pandemia da COVID-19 causada pelo vírus SARS-COV-2.

REFERÊNCIAS

BOYLE, P.; LEVIN, B. (eds.). **World Cancer Report 2008**. Lyon: IARC, 2008. URL : Disponível em: http://www.iarc.fr/en/publications/pdfs-online/wcr/2008/wcr_2008.pdf. Acesso em: 02 jun/2020.

BRAY, F.; FERLAY, J.; SOERJOMATARAM, I.; SIEGEL, RL; TORRE, LA; JEMAL, A. **Global Cancer Statistics 2018: GLOBOCAN**. Disponível em: http://gco.iarc.fr. Acesso em: 02 jun/2020.

HERRINGTON, C. S. (ed.). **Muir's Textbook of Pathology**. 5 ed. Boca Raton: CRC Press, 2014.

Instituto Nacional de Câncer José de Alencar Gomes da Silva (INCA), Ministério da Saúde (MS), Secretaria de Atenção à Saúde (SAS) (2020). **Estimativa 2020– Incidência de Câncer no Brasil**. Rel. téc. Brasília, 2020. Disponível em: https://www.inca.gov.br/estimativa. Acesso em: 21 jul/2020.

KUMAR, V.; ABBAS, A. K.; ASTER, J. C. (eds). (2013). **Robbins Basic Pathology**. 9 ed. Philadelphia: Elsevier, 2013.

RUBIN, R.; STRAYER, D. S.; RUBIN, E. (eds.). **Rubin's Pathology Clinicopathologic Foundations of Medicine**. 6 ed. Philadelphia: Lippincott Williams & Wilkins, 2012.

SPANHOL, F.; OLIVEIRA, L. S.; PETITJEAN, C. HEUTTE, L. **A Dataset for Breast Cancer Histopathological Image Classification**. IEEE Transactions on Biomedical

Engineering (TBME), 63 (7), pp. 1455–1462, 2016. DOI: 10.1109/TBME.2015.2496264.

SPANHOL, F.; OLIVEIRA, L. S.; PETITJEAN, C. HEUTTE, L. **Breast cancer histopathological image classification using Convolutional Neural Networks**. *In*: 2016 International Joint Conference on Neural Networks (IJCNN). Vancouver: IEEE, 2016b. DOI : 10.1109/IJCNN.2016.7727519.

SPANHOL, Fabio Alexadre. **Automatic breast cancer classification from histopathological images: a hybrid approach**. Tese (Doutorado em Ciência da Computação) – Programa de Pós-Graduação em Informática da Universidade Federal do Paraná. Curitiba, 2018. Disponível em: https://hdl.handle.net/1884/57312. Acesso em: 22 jul. 2020.

VETA, M.; PLUIM, J. P. W.; DIEST, P. J.; VIERGEVER, Max A. **Breast cancer histopathology image analysis: a review**. IEEE Transactions On Biomedical Engineering (TBME), 61 (5), pp. 1400–1411, 2014. DOI: 10.1109/TBME.2014.2303852.

World Cancer Research Fund (WCRF). **Worldwide Cancer Data**. Disponível em: https://www.wcrf.org/dietandcancer/cancer-trends/worldwide-cancer-data. Acesso em: 02 jun/2020.