Computers & Graphics (2021)

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Nova Abordagem Totalmente Automática para Identificação de Tecidos em Exames Histopatológicos com uso de Transfer Learning⋆

Matheus Santos[a,*], Anderson Bessa[a], Adriell Gomes[a], Iágson Silva[a,1], Luís Fabrício Freitas[a,b,1], Pedro Pedrosa Rebouças[a,b,2]

[a]*Laboratório de Processamento de Imagens, Sinais e Computação Aplicada (LAPISCO) - IFCE, Fortaleza, Brazil*
[b]*Address, City, Postcode, Country*

ARTICLE INFO

Article history:
Received September 27, 2021

Keywords: Classificação de Câncer de Mama, Redes Neurais Convolucionais, Transfer Learning, Diagnóstico Assistido por Computador

ABSTRACT

A utilização de técnicas computacionais no processamento de imagens histopatológicas permite o estudo da organização estrutural dos tecidos e suas alterações mediante ação de doenças. O presente trabalho propõe a análise dessas imagens com o auxílio de CNNs associadas à extratores de Machine Learning através do processo de Transfer Learning. Foi utilizado nos experimentos o dataset BreakHis, constituído de imagens histopatológicas de Câncer de mama com diferentes escalas de ampliação de tumores que podem ser classificados como Maligno ou Benigno. Neste estudo foram realizados diferentes combinações de Extrator-Classificador, buscando assim a comparação do melhor modelo. Dentre os resultados alcançados, o melhor conjunto Extrator-Classificador formado foi a CNN DenseNet201, atuando como extrator, com o classificador SVM RBF, obtendo acurácia de 95,39% e precisão de 95,43% para o fator de magnificação de 200X. A fim de validar os experimentos, foi gerado diferentes modelos, comparados entre si, e validados em uma comparação baseados em métodos da literatura, mostrando assim a eficácia no modelo proposto.

© 2021 Elsevier B.V. All rights reserved.

1. Introduction

O câncer de mama é a causa de morte por câncer mais comum entre mulheres na faixa etária de 40 à 45 anos, além de ser o principal fator da mortalidade no sexo feminino [1]. Segundo a Organização mundial da saúde (OMS), em 2020, cerca de 2,3 milhões de mulheres foram diagnosticadas com câncer de mama, sendo que até o final do mesmo ano, cerca de 7,8 milhões de mulheres viviam após receber o diagnóstico. Por fim, em 2020 ocorreram cerca de 685.000 mortes em todo o mundo, esses dados afirmam o câncer de mama como o mais prevalente do mundo [2]. Nesse contexto, um paciente com esse tipo de patologia gasta anualmente cerca de US $ 2.300 de forma direta e cerca de US $ 3325 a $ 5.545 com custos indiretos devido às despesas médicas da doença [3].

Quando se tem a necessidade de um diagnóstico preciso para um preparatório cirúrgico, sobretudo quando não é possível determinar a causa do tumor com uma biópsia por agulha [4] ou demais técnicas, a biópsia cirúrgica aberta (OSB) é o exame mais recomendado, obtendo praticamente todas as lesões suspeitas em seu procedimento com uma alta taxa de confiabilidade. Mesmo sendo uma técnica incisiva e com relativo custo. A OSB ainda é uma técnica considerada eficaz para definir a totalidade da patologia nos casos em que a massa é pequena e difícil de localizar pelo toque ou se uma área parece suspeita em uma mamografia, mas não pode ser sentida [5].

⋆Only capitalize first word and proper nouns in the title.

*Corresponding author:
e-mail: matheus.santos@lapisco.ifce.edu.br (Matheus A. dos Santos)

[1]Prof. Fabrício

[2]Prof. Dr. Pedro Pedrosa

As imagens de microscópio providas pelo exame OSB são denominadas imagens histopatológicas [6], após a obtenção desse tipo de imagem, se faz necessário classificar o tumor como maligno ou benigno, geralmente esse processo é manual, realizado pelo médico especialista [7], o que pode gerar uma demora na classificação além de inúmeros fatores causadores de erros como a alta carga de trabalho ou fatores externos como iluminação [8]. A seguir são abordados estudos que se utilizaram principalmente de Transferência de Aprendizagem e do dataset BreakHis para a classificação do câncer de mama.

Com objetivo de tornar não somente este tipo de classificação, mas diversas outras aplicações médicas mais rápidas e precisas, muitos métodos de Diagnóstico Assistido por Computador (*Computer Aided Diagnosis* - CAD) foram propostos ao longo dos anos. Esses métodos são baseados principalmente em Processamento Digital de Imagens[9], Machine Learning [10] e deep learning [11], que se caracterizam por utilizar da Inteligência Artificial, ou auxiliar, em atividades humanas de forma automática e precisa [12]. Trabalhos como o de Sharma et al [13] que propuseram duas propostas de classificação baseada em algoritmos Machine e Deep Learning Learning para o conjunto de multiclasses presentes no dataset BreakHis [14]. A primeira proposta é baseada em recursos artesanais extraídos usando Momentos de Hu, histograma de cores e texturas Haralick para posteriormente a classificação desses atributos serem classificados por classificadores convencionais. Já a segunda proposta é baseada no uso das redes VGG16, VGG19 e ResNet50 para extração e classificação de atributos. Os resultados foram satisfatórios, sobretudo para as redes neurais convolucionais, alcançando acurácias de 93,97%.

Então, levando em consideração a importância de um sistema CAD para a classificação das histopatologias de forma rápida e eficiente, baseado em diferentes técnica na literatura com a mesma finalidade, além do dataset BreakHis, o método proposto busca a classificação para as diferentes classes de histopatologias de câncer de mama, especialmente a classificação entre benigno ou maligno através do uso de algoritmos de Machine e Deep Learning. Como contribuições, este estudo apresenta resultados significativamente melhores que os métodos encontrados na literatura atual, além de ter um rápido tempo de classificação e apresentar não somente a classificação de acordo com o fator de escala do dataset BreakHis, mas também ótimos resultados na classificação generalizada do dataset, isto é, com todas as escalas de uma só vez.

2. Related works

O uso de *deep learning* ganhou força nos últimos anos, ao ser aplicado como abordagem para problemas envolvendo *machine learning*, como reconhecimento de objetos e classificação [15]. Uma rede neural convolucional (CNN) é composta por camadas convolucionais (usadas para gerar atributos), camadas de *pooling* (usadas para unir informações de uma região) e camadas chamadas de *fully-connected*, usadas na classificação.

As imagens histopatológicas (H&E) são difundidas na detecção de diversos tipos de patologias, como câncer de mama, câncer colorretal e câncer de pulmão. Os vários usos desse tipo de imagem médica demonstram o vasto escopo para melhorias no uso de exames histológicos e a importância de contrastar essas abordagens [16]. A técnica de (*transfer learning*) usa das camadas convolucionais e de *pooling* da CNN, enquanto substitui as camadas *fully-connected* por classificadores tradicionais (como KNN, SVM, etc). Dessa forma, a rede neural convolucional funciona como um extrator de atributos, ou extrator de características.

Nesse sentido, nos estudos de Celik *et al.* [17] houve uma investigação quanto à detecção automática de carcinoma ductal invasivo (IDC), que é o subtipo mais comum de câncer de mama, usando a técnica de aprendizagem de transferência profunda, sobretudo com o auxílio dos extratores ResNet-50 e DenseNet-161 para detecção de IDC. O método desenvolvido pelos autores foi aplicado no conjunto de dados de imagem BreaKHis para classificação entre tumores benignos e malignos da mama, obtendo acurácias médias de 88 % para a classificação do conjunto total, sem separação por escala. Entretanto vale ressaltar que o dataset foi balanceado para uma melhor performance, o que pode vir a comprometer a capacidade de generalização do método.

O trabalho de Zhi *et al.* [18] investiga o uso de aprendizagem por transferência com redes neurais convolucionais para diagnosticar automaticamente o câncer de mamas em manchas de imagens histopatológicas providas pelo BreakHis. os autores Aliaram a aprendizagem por transferência com a CNN VGGNet em uma arquitetura customizada mais superficial. Para a classificação, o dataset foi separado de acordo com as escalas, e as duas principais classes como sendo maligno e benigno. O método obteve resultados satisfatórios, com métricas superiores ao mesmo modelo de CNN treinada sem transfer Learning, como acurácias de 94% para a classificação na escala de 200%. Entretanto, a comparação entre o modelo proposto e demais métodos encontrados na literatura para o dataset BreakHis foi embasada somente na acurácia, o que pode vir a comprometer a confiabilidade de tal comparação, tendo em vista a ausência de observação de outros pontos do método, como sua detecção de falsos positivos e falsos negativos.

Também levando em consideração as vantagens dessa abordagem, e o mesmo dataset, Mehra *et al.* [19] dispuseram de aprendizagem de transferência em comparação com redes totalmente treinadas na modalidade de imagem histopatológica. Foram analisadas três redes pré-treinadas: VGG16, VGG19 e ResNet50, observou-se o comportamento das redes para ampliação da escala da imagem, a classificação levou em conta os dois principais tipos de câncer. O método obteve 92% de acurácia e 95% de ROC, comprovando a eficácia do uso de Transfer Learning sobre as redes. Entretanto os autores realizaram balanceamento no dataset para uma melhor performance, sem que fosse especificado de maneira mais detalhada tal procedimento, isso torna a performance do método menos confiável, além disso não foi realizada a classificação por escala.

Song *et al.* [20] apresentaram uma abordagem baseada em transferência de aprendizagem para a classificação de imagens histopatológicas. Além disso, foi-se utilizado o recurso de imagem de Fisher Codificação Vetorial (FV) de características lo-

cais, que são extraídas usando o modelo de Rede Neural Convolucional (CNN) pré-treinado na ImageNet. Novamente, o conjunto de dados de imagem utilizado foi o BreaKHis, para classificação entre tumores benignos e malignos da mama. Os autores utilizaram somente a acurácia como métrica de avaliação, oque impossibilita uma comparação mais ampla com trabalhos no estado da arte, e somente realizaram a classificação por escala.

Deniz *et al.* [21] nortearam o tema voltado para o câncer de mama em que a transferência de aprendizado e métodos de extração de características são usados para adaptar um modelo CNN pré-treinado ao problema em questão. Os modelos AlexNet e Vgg16 são utilizados para extração de recursos. Os atributos obtidos são classificados por máquinas de vetores de suporte (SVM). O conjunto de dados usado foi novamente o BreakHis, sendo dividido entre as escalas de imagem para as classes maligno e benigno. O método alcançou 95% de acurácia com o classificador SVM RBF na escala de 200X. Entretanto não foi realizada a classificação sem distinção por fator de escala, oque poderia vir a demonstrar a capacidade de generalização do método.

Portanto, é notório que as abordagens citadas acima norteiam a aprendizagem profunda e utilizam de CNNs para extrair atributos das imagens, bem como métodos de aprendizado de máquina para classificá-las. Além disso, todos propuseram utilizar o conjunto de dados BreaKHis. Entretanto, é possível afirmar que os resultados realizados por nosso método para extração são mais propícios para análises médicas, uma vez que os experimentos do método proposto por este estudo superou em acurácia comparadas a outros métodos encontrados na literatura assim como no estado da arte encontrado nesta Seção de Trabalhos Relacionados. Logo, de fato, o método proposto baseia-se em uma metodologia capaz de alcançar bons resultados.

3. Materiais e Métodos

Nesta seção são apresentados os métodos utilizados nos experimentos, nas diferentes combinações de métodos, por meio da técnica de transfer learning para a geração de diferentes modelos computacionais para classificação de tecidos em exames histopalógico. Bem como as métricas de avaliação utilizados nos experimentos.

3.1. Extração de atributos através de redes neurais convolucionais

O método de Transfer Learning consiste em usar uma rede neural convolucional (CNN), pré-treinada com uma grande base de dados, para extrair características de forma e textura de outro dataset, de forma a fazer a extração dos atributos das imagens para usar na etapa de classificação, utilizando de classificadores tradicionais [22]. Ou seja, para ser usada exclusivamente como extrator de características, a CNN é desprovida de suas camadas de classificação, chamadas de fully-connected layers [23, 24]. Dessa forma, a rede não precisa ser treinada, diferente do que ocorre com outros métodos como fine tuning ou learning from scratch, que buscam treinar a rede com o dataset em questão [25].

Nesse trabalho foi utilizado as topologias de CNN's com pesos já treinados originalmente com o grande banco de imagens ImageNet [26], que consiste em milhões de imagens de objetos do dia a dia, com cerca de 1000 classes diferentes. Para todas, as fully-connected layers foram removidas, sendo a saída de cada uma das redes, um vetor resultante da última camada convolucional ou de pooling.

As arquiteturas de CNN's usadas nesse paper foram duas: a arquitetura VGG [27] e a arquitetura DenseNet [28]. A primeira foi implementada em duas configurações diferentes (VGG16 e VGG19), enquanto a segunda foi implementada em três (DenseNet121, DenseNet169 e DenseNet201).

As configurações VGG16 e VGG19 se diferenciam pelo seu número de camadas de pesos, a primeira com 16 camadas e a segunda com 19 camadas. Essa arquitetura se diferencia pelo uso de filtros convolucionais pequenos, o que aumenta seu poder de profundidade [27].

Em paralelo, as três diferentes configurações de DenseNet se diferenciam pelos diferentes números de camadas que as compõem, sendo compostas por camadas convolucionais, de transição e pelos chamados "DenseBlocks". Sua principal característica é suas conexões densas entre as camadas, alimentando um sistema de forte propagação de atributos para as próximas camadas e o seu reuso, sendo necessários poucos parâmetros [28].

3.2. Classificação utilizando técnicas de machine learning

Após a etapa de extração de características, os vetores de atributos, extraídos pelas topologias apresentadas na seção anterior, têm diferentes tamanhos, a depender da rede, e são fornecidos aos sete classificadores utilizados nesse trabalho. São eles: Naive Bayes, MLP, Nearest Neighbors, Random Forest e três diferentes versões de SVM (Linear, Polinomial e RBF).

O primeiro classificador utilizado, o Naive Bayes, é um algoritmo que faz uma análise estatística do vetor de atributos, se baseando na Regra de Decisão de Bayes, em análise condicional e na função de densidade de probabilidade. O método calcula um valor de probabilidade para uma amostra pertencer a cada uma das classes em questão, e no fim a rotula com a classe de maior probabilidade [29].

Já o classificador MLP (Multi-layer Perceptron) é um algoritmo composto por várias camadas do neurônio artificial, chamado perceptron [30]. Entre a entrada, que recebe os atributos, e a saída do MLP, há diversas camadas de perceptrons com diferentes pesos que propagam a informação inicial por toda sua extensão, aprendendo com os valores fornecidos a ele, para predizer a classe da amostra em sua saída [31].

O Nearest Neighbors, ou KNN, é um método de machine learning supervisionado que classifica uma amostra através de sua distribuição espacial com as outras já rotuladas [32]. Ele se baseia no seu parâmetro k, que deve ser ímpar e é o número de vizinhos mais próximos da atual amostra. Logo, a nova amostra recebe o rótulo mais frequente entre os k vizinhos.

Baseando-se nas árvores de decisão, e na forma de classificação do cérebro humano, tem-se o algoritmo Random Forest [33]. Esse método é considerado não supervisionado e tem inicialização aleatória, utilizando-se de estimadores para o tratamento das informações de entrada.

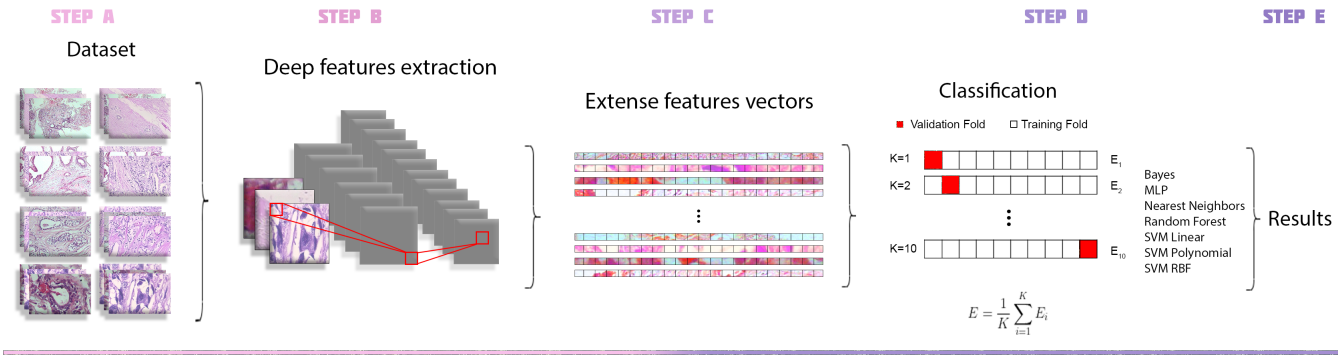


Fig. 1. Flowchart of transfer learning process, the methodology used in this work. The Step A represents the split of dataset, depending on magnification factor. The Step B presents the feature extraction of images by each CNN model. The Step C represents the extense feature vectors resulting from extraction stage. The Step D shows the classification stage of each feature vector into benign or malignat classes. And the Step E represents the results reached by each model in terms of the evaluation metrics.

As Support Vector Machine (SVM), por sua vez, são métodos de classificação que utilizam de análise estatística e hiperplanos de separação ótima que dependem diretamente do kernel utilizado para análise da disposição espacial das amostras [34]. Os kernels usados para as SVM's foram o Linear, o Polinomial e o RBF (Radial Basis Function).

3.3. Métricas de Avaliação

As métricas utilizadas para avaliar a classificação de cada combinação foram: acurácia (Acc), precisão, recall ou sensitividade, F1-Score (F1) e Coeficiente de Correlação de Matthews (MCC).

Todas as equações usam verdadeiros positivos (VP), falsos negativos (FN), verdadeiros negativos (VN) e falsos positivos (FP), todos presentes na matriz de confusão, para cálculo das métricas de avaliação que podem ser vistas nas Equações de 1 à 5. A matriz de confusão é uma ferramenta que faz a comparação entre a classe real de cada amostra classificada e a classe predita pelo método.

$$Acc = \frac{VP + VN}{VP + VN + FP + FN} \tag{1}$$

$$Precision = \frac{VP}{VP + FP} \tag{2}$$

$$Recall = \frac{VP}{VP + FN} \tag{3}$$

$$F1 = \frac{2VP}{2VP+FP+FN} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

$$MCC = \frac{VP \times VN - FP \times FN}{\sqrt{(VP+FP)(VP+FN)(VN+FP)(VN+FN)}} \tag{5}$$

A acurácia (Acc) é uma métrica que mede a quantidade de amostras classificadas corretamente pelo número total de amostras, e como trata diretamente do índice de acerto do método, é muitas vezes colocada como principal. O F1-Score, por outro lado, usa das métricas intermediárias, Precision e Recall, para medir um balanço entre elas, através de sua média harmônica. Dessa forma, um alto índice de F1-Score é considerado um fator de uniformidade na classificação de tecidos, pois mostra um número baixo de falsos positivos e falsos negativos [16].

Por fim, o Coeficiente de Matthews (MCC) é uma medida de correlação amplamente utilizado em problemas de classificação binários. Baseado nos índices de correlação de Pearson, esse coeficiente pode ser tratado como uma medida balanceada mesmo quando se tem um dataset desbalanceado.

Em muitas dessas métricas, existe um resultado para cada classe em questão. Para facilitar a análise nos resultados, foram feitas as médias aritméticas simples dos resultados de classe para representação em tabelas. Os valores serão representados em porcentagem.

3.4. Teste Estatístico

O teste estatístico utilizado foi o teste de Kolmogorov-Smirnov, esse teste é não paramétrico, comumente utilizado para avaliar a similaridade estatística entre dois vetores de atributos. Através do teste estatístico é possível afirmar com maior propriedade se uma média com seu respectivo desvio padrão é diferente da outra, pois não somente seus valores pontuais são comparados, e sim o vetor de valores que a gerou [35].

Para a realização de tal comparação, é necessário se estabelecer um coeficiente de confiabilidade, isto é um α, para que então o algoritmo de comparação entre os vetores de atributos forneça um valor de P ao final de sua apuração. Esse valor P é então comparado ao alfa, aceitando a hipótese de semelhança estatística caso seja maior, e recusando a hipótese caso seja menor que o valor α. Normalmente o valor de α escolhido é de 5% (0.05) [36].

4. Metodologia

Esta seção apresenta a metodologia proposta nesse estudo para a abordagem de transfer learning. A seção é dividida em subseções abordando o dataset utilizado nos experimentos e a metodologia bem c como a metodologia aplicada ao estudo, assim como os parâmetros de cada modelo utilizado.

Table 1. BreaKHis dataset images distribution according to class, subclass and magnification factor [37].

Class	Subclass	Magnification Factors				Total
		40X	100X	200X	400X	
Benign	Adenosis Fibroadenoma Phyllodes Tumor Tubular Adenoma	625	644	623	588	2,480
Malignant	Ductal Carcinoma Lobular Carcinoma Mucinous Carcinoma Papillary Carcinoma	1,370	1,437	1,390	1,232	5,429
	Total					7,909

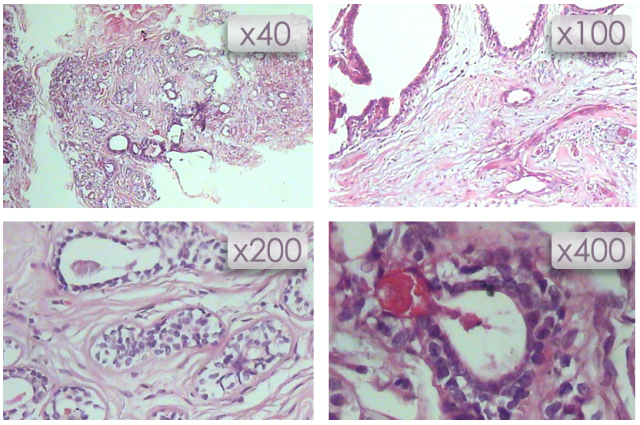


Fig. 2. benign.png

4.1. Base de dados BreaKHis

A BreaKHis é uma base de dados composta por milhares de imagens de biópsia, adquiridas por meio de microscópios, de tecidos presentes em tumores benignos e malignos em mamas [37]. O dataset foi formado entre Janeiro e Dezembro de 2014, com pacientes convidados pelo Laboratório de P&D, no Brasil.

As amostras foram coletadas por meio de surgical open byopsis (OSB) e preparadas para estudo através de um microscópio acoplado a uma câmera digital. O dataset resultante é composto por 7,909 imagens com 3 canais RGB, de 8 bits cada, no formato PNG e de dimensões de 700×460 pixels.

As imagens de tecidos, divididas nas classes principais benigno-maligno, se subdividem ainda em outras 8 subclasses, de acordo com o tipo de lesão, conforme mostra a Tabela 1. Por fim, também há uma divisão conforme o fator de magnificação, que pode ser de 40X, 100X, 200X e 400X.

Seguindo outros trabalhos que usaram o mesmo dataset, e em um contexto de auxílio ao diagnóstico médico, optou-se pela classificação binária entre as classes benigno e maligno para cada um dos fatores de magnificação. Dessa forma, as subclasses não foram consideradas para esse trabalho.

Além disso, as amostras da BreaKHis também podem ser agrupadas entre os 82 pacientes que se voluntariaram para a formação do dataset. Dessa forma, várias imagens seriam designadas para um mesmo paciente, o que promoveria um trabalho de classificação diferente. Vale ressaltar que para esse trabalho a classificação se deu somente a nível de imagem, para os diferentes fatores de magnificação, e a classificação a nível de paciente foi desconsiderada.

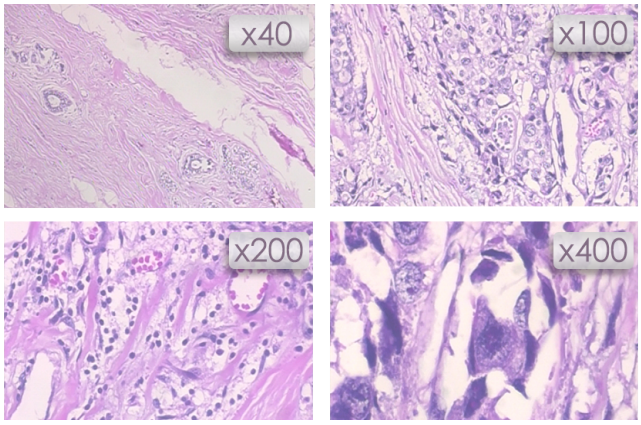


Fig. 3. malignant.png

Também ressaltamos que não houve qualquer trabalho de augmentation ou balanceamento do dataset. Logo, somente as imagens originais foram fornecidas às CNN's, somente com o rótulo benigno/maligno.

4.2. Metodologia do Estudo Proposto

A fim de diferenciar as imagens de exames histopatológicos obtidas através de biópsia, em suas classes benigno/maligno (para cada subdivisão dos fatores de magnificação), é extraído os atributos de cada uma das imagens, por meio de cinco diferentes CNN's (DenseNet121, DenseNet169, DenseNet201, VGG16 e VGG19), usadas como extratores, para uma classificação totalmente automática do final do processo de extração, posteriori por meio de sete diferentes classificadores (Naive Bayes, MLP, Nearest Neighbors, Random Forest, SVM Linear, SVM Polynomial e SVM RBF), ambos os modelos processam paralelamente. EM seguida, é avaliado o desempenho de cada combinação extrator-classificador. As arquiteturas de CNN mencionadas neste estudo, bem como os classificadores, estão apresentados na Seção 3.

A Figura 1 detalha a abordagem de da metodologia proposta com uso de técnicas de transfer learning. A metodologia foi subdividida em Etapas de A a E, abordando os diferentes pontos de cada Etapa.

No Step A, as imagens do dataset BreaKHis, apresentado acima, foram divididas segundo suas escalas de magnificação (40X, 100X, 200X e 400X), formando quatro diferentes datasets, com diferentes quantidades de imagens, rotuladas em benigno ou maligno.

É Realizado a técnica de resize para cada imagem, como pré-processamento, a fim de adequá-las à entrada para a rede, a depender da topologia que estaria sendo utilizada. Todos os outros parâmetros e configurações das CNN's utilizadas nesse trabalho foram mantidos como padrão da própria rede, como usados também nos trabalhos comparados na Seção 3.

No Step B, as topologias de CNN, citadas acima, recebem cada um desses datasets formados, de forma a atuarem como extratores de características, formando um sub-dataset composto pelos vetores de atributos extraídos de cada uma das imagens e pela sua respectiva classe (benigno - 0, maligno - 1), resultando no Step C. Os vetores de atributos têm tamanhos difer-

entes, dependendo da CNN usada na extração: 512 atributos nos casos de VGG16 ou VGG19 e 1024, 1664 e 1920 atributos nos casos de DenseNet121, DenseNet169 e DenseNet201, respectivamente.

No Step D, o dataset composto pelos vetores de atributos gerados no Step B e C, seguem o processo para os classificadores, a fim de realizar uma classificação supervisionada de cada uma dessas imagens extraídas. O treinamento se deu por cross-validation, usando o método K-fold com 10 folders. Para cada folder, foi feita a separação das imagens na proporção de 90%/10% para treino/validação.

Um algoritmo de busca aleatória foi utilizado para buscar as melhores configurações para cada um dos classificadores para o problema atual. O método de Random Search foi utilizado para buscar os melhores hiperparâmetros para cada um deles, de acordo com os intervalos de busca, que são apresentados na Tabela 2. Foram feitas 20 iterações com cross-validation de 5 folds para buscar os parâmetros de cada classificador.

Table 2. Parameters intervals provided to Random Search algorithm for each classifier used in the transfer learning approach.

Classifier	Parameters of Search	Intervals of Search
Naive Bayes	-	-
MLP	Hidden layers	[2, 1001]
Nearest Neighbors	Number of Neighbors	1, 3, 5, 7, 9, 11, 13, 15
Random Forest	Maximum depth	6, Unlimited
	Bootstrap	True, False
	Criterion	Gini, Entropy
SVM Linear	Regularization parameter (C)	$2^x \vert x \in [-5, 15]$
SVM Polynomial	Degree	3, 5, 7, 9
	Regularization parameter (C)	$2^x \vert x \in [-5, 15]$
SVM RBF	Gamma	$2^x \vert x \in [-15, 3]$
	Regularization parameter (C)	$2^x \vert x \in [-5, 15]$

Além dos parâmetros encontrados através da busca aleatória, destacam-se outros parâmetros fixos para os classificadores: Para o MLP, foi utilizado um limite de 1000 iterações e uma taxa de aprendizagem inicial no valor de 5×10^{-4}; para o Random Forest, além da inicialização aleatória, foram usados 3000 estimadores; e, por fim, para os três tipos de SVM foram usadas tolerâncias fixas de 10^{-3}. Os demais parâmetros desses classificadores, que não foram mencionados aqui, foram mantidos fixos e iguais aos valores padrões da biblioteca scikit-learn 0.20.2, usada na implementação desses métodos.

Por fim, no Step E, foram usadas as métricas de avaliação vistas na Seção 3 para a avaliação de cada modelo formado pela combinação CNN-classificador. Os resultados dessas métricas são apresentados na Seção 5. Além da comparação entre os modelos, a fim de encontrar a melhor combinação para o problema em questão, são comparados melhor modelo com outros trabalhos da literatura que usaram o mesmo banco de dados em problemas de classificação binária, tanto para os fatores de magnificação individualmente, quanto para o dataset inteiro, sem distinção entre as escalas.

5. Experimental Results

Esta seção trata dos resultados obtidos por cada uma das 35 combinações CNN × classificador (5 × 7), para cada um dos fatores de magnificação, à luz das métricas de avaliação apresentadas na seção anterior. Além disso, nas subseções que se seguem, foi realizado a comparação da melhor combinação feita pela o estudo (Melhor modelo) com outros trabalhos da literatura, e por fim, é realizado um experimento classificatório, usando o dataset inteiro, simulando um augmentation por escala, Afim de validar o experimento também comparado com trabalho da literatura.

Todo o processamento foi realizado utilizando o sistema operacional Linux (distribuição Ubuntu 16.04) com 16GB de RAM e processador AMD Ryzen 5 3400G. Os processos de extração com CNN's foram acelerados por meio de uma GPU Nvidia GeForce GTX 1660 Super com memória dedicada de 6GB. Os modelos de *deep learning* (VGG's e DenseNet's) foram implementados na linguagem de programação Python (versão 3.7), por meio das bibliotecas tensorflow-gpu 1.14, keras 2.2.4 e com uso da OpenCV 4.1.0, enquanto os classificadores foram implementados com a biblioteca scikit-learn 0.20.2.

A Tabela 3 mostra os resultados obtidos com o fator de magnificação de 40X, para uma classificação binária benigno-maligno das imagens histopatológicas. Para todas as arquiteturas de CNN usadas, vê-se que o classificador SVM RBF atingiu acurácias acima de 91.9%. Os resultados de precisão e recall, para esse mesmo classificador, também ficaram acima de 90%. Para as métricas restantes, F1-Score e Coeficiente de Matthews, os resultados ficaram na faixa de 80% ou 90%, o que sinaliza uma uniformidade e balanço na classificação de ambas as classes.

A Tabela 3 também mostra que a arquitetura DenseNet se mostra superior, como extrator de características, se comparada a arquitetura VGG. Entretanto, essa diferença não é muito significativa para a maioria dos classificadores. Diferentemente do SVM RBF, o SVM Polynomial foi o pior de todos os classificadores, não alcançando nenhum valor métrico acima de 50%, o que mostra que esse classificador tende a errar mais do que acertar referente a regiões pertencentes ao tecido mamário. Por fim, os classificadores MLP e Nearest Neighbors também conseguiram resultados próximos ao do SVM RBF para a maioria das métricas, mas perdem para ele quando se trata de tempo de treino (maior processamento) e tempo de teste, respectivamente, visto que o tempo de treino do MLP é em torno de 120 a 190 ms, enquanto o tempo de predição do Nearest Neighbors se dá na faixa de 9 a 30 ms, com alto desvio-padrão (enquanto o SVM RBF treina e prediz numa média mais rápida que os outros dois e com desvio-padrão menor).

A Tabela 4 mostra os resultados da classificação para o dataset BreaKHis, restrito ao fator de magnificação de 100X. De forma bem parecida com os resultados da tabela anterior, a combinação DenseNet201 com SVM RBF também conseguiu os maiores valores para todas as métricas. Novamente os classificadores Naive Bayes e SVM Polynomial se mostraram os inferiores para esse problema de classificação do tecido mamário, apesar de uma melhora do SVM Polynomial para essas imagens

Feature Extractor	Classifier	Accuracy	Precision	Recall	F1-Score	Matthews	Train Time	Predict Time
DenseNet121	Bayes	73.83±0.68	79.63±4.12	59.15±0.56	58.00±0.70	32.90±3.14	0.049±0.022	0.106±0.010
	MLP	90.83±1.34	89.66±2.14	89.10±1.31	89.30±1.46	78.73±3.03	187.838±8.430	0.677±0.318
	Nearest Neighbors	91.73±0.69	91.21±0.73	89.37±1.26	90.17±0.91	80.54±1.69	0.104±0.015	16.512±2.471
	Random Forest	87.67±2.76	85.90±3.37	85.37±3.11	85.59±3.17	71.25±6.38	15.718±1.708	5.709±0.454
	SVM Linear	90.38±1.54	88.37±1.86	90.08±1.39	89.09±1.66	78.43±3.19	6.940±0.638	3.758±0.195
	SVM Polynomial	31.33±0.00	15.66±0.00	50.00±0.00	23.85±0.00	0.00±0.00	17.794±1.158	11.071±0.536
	SVM RBF	93.13±1.28	92.17±1.83	91.91±1.23	92.01±1.45	84.07±2.92	8.139±0.764	4.701±0.499
DenseNet169	Bayes	78.20±0.88	83.31±0.75	66.24±1.56	67.79±1.98	46.47±2.46	0.083±0.030	0.087±0.029
	MLP	92.98±1.95	91.87±2.42	91.85±2.06	91.85±2.23	83.72±4.47	117.196±8.137	0.590±0.323
	Nearest Neighbors	93.78±1.11	93.29±0.95	92.17±1.82	92.67±1.40	85.44±2.66	0.123±0.018	24.993±3.575
	Random Forest	89.32±1.14	90.04±1.47	84.83±1.51	86.79±1.45	74.68±2.79	14.133±1.672	10.158±3.413
	SVM Linear	91.63±1.71	89.89±1.95	91.25±1.99	90.45±1.90	81.12±3.74	9.602±0.873	5.606±0.351
	SVM Polynomial	31.33±0.00	15.66±0.00	50.00±0.00	23.85±0.00	0.00±0.00	26.836±0.972	17.894±0.309
	SVM RBF	94.09±1.32	93.63±1.83	92.56±1.32	93.05±1.52	86.18±3.07	11.621±0.960	7.104±0.440
DenseNet201	Bayes	78.35±1.05	84.11±1.62	66.31±1.74	67.87±2.23	47.11±3.04	0.071±0.030	0.100±0.034
	MLP	93.18±0.81	92.64±1.02	91.38±1.02	91.96±0.96	84.01±1.92	121.425±2.298	0.836±0.076
	Nearest Neighbors	93.18±1.36	93.60±1.58	90.43±1.70	91.79±1.65	83.96±3.27	0.122±0.031	29.353±4.152
	Random Forest	88.12±1.45	86.55±1.85	85.65±1.54	86.04±1.61	72.19±3.21	23.778±2.721	6.171±1.180
	SVM Linear	91.63±0.80	90.07±0.88	90.69±1.23	90.35±0.96	80.75±1.98	10.820±0.638	6.393±0.282
	SVM Polynomial	31.33±0.00	15.66±0.00	50.00±0.00	23.85±0.00	0.00±0.00	30.694±1.194	20.596±0.292
	SVM RBF	**94.94±0.87**	**94.63±1.05**	**93.53±1.11**	**94.04±1.04**	**88.15±2.06**	13.418±0.768	8.655±0.804
VGG16	Bayes	73.83±0.88	72.83±1.87	61.24±1.40	61.46±1.85	32.00±2.95	0.040±0.030	0.074±0.018
	MLP	87.87±1.31	86.60±1.38	84.99±2.51	85.58±1.79	71.53±3.32	129.823±9.492	0.394±0.174
	Nearest Neighbors	89.42±0.99	89.70±1.03	85.30±1.42	87.03±1.29	74.86±2.43	0.045±0.006	9.743±1.467
	Random Forest	84.96±1.65	86.91±2.31	77.74±2.22	80.34±2.32	63.98±4.37	34.466±3.962	8.064±1.655
	SVM Linear	87.02±1.03	84.69±1.33	85.94±1.19	85.20±1.12	70.61±2.27	5.278±0.586	2.730±0.110
	SVM Polynomial	31.33±0.00	15.66±0.00	50.00±0.00	23.85±0.00	0.00±0.00	10.200±0.302	5.609±0.037
	SVM RBF	91.98±0.79	91.62±0.55	89.55±1.62	90.44±1.07	81.13±1.99	8.142±0.604	3.119±0.177
VGG19	Bayes	74.59±1.53	73.68±3.49	62.62±1.74	63.31±2.17	34.57±4.90	0.047±0.024	0.074±0.015
	MLP	88.77±0.58	87.45±1.20	86.39±1.15	86.79±0.63	73.81±1.35	178.718±8.357	0.556±0.248
	Nearest Neighbors	89.47±0.69	90.43±0.92	84.90±1.28	86.94±1.01	75.10±1.61	0.046±0.006	9.576±1.640
	Random Forest	86.17±2.93	85.12±3.33	81.84±3.90	83.13±3.70	66.86±7.21	31.504±4.190	6.997±2.559
	SVM Linear	86.42±1.70	83.98±1.94	85.28±2.31	84.52±2.00	69.25±4.11	4.965±0.543	2.400±0.087
	SVM Polynomial	31.33±0.00	15.66±0.00	50.00±0.00	23.85±0.00	0.00±0.00	10.386±1.002	5.501±0.056
	SVM RBF	91.98±1.30	91.92±1.61	89.20±1.58	90.38±1.56	81.07±3.12	9.056±1.115	3.729±0.295

Table 3. Metrics (%), train time (ms) and predict time (ms) obtained by each combination of feature extraction and classifier for the 40X magnification factor. In bold, the best results per metric and the best combination method.

de 100X.

Os demais valores de métricas atingidos pelas outras combinações tiveram resultados muito próximos da 3, o que mostra que o fator de magnificação de 100X não tem tanta diferença se comparado ao de 40X, provavelmente por serem muito próximos um do outro. Apesar dos valores de acurácia serem próximos, para o fator de 100X, houve uma diminuição no desvio-padrão em praticamente todas as combinações, mostrando que os modelos obtiveram resultados minimamente melhores para classificar o tecido em imagens Histopatol ogicas. Os valores de F1-Score na tabela, por usa vez, ligados aos classificadores MLP, Nearest Neighbors, Random Forest, SVM Linear e SVM RBF ficam na mesma faixa entre 80% e 92%. O Coeficiente de Matthews, por sua vez, tem seus valores destaque para o MLP, Nearest Neighbors e SVM RBF, principalmente para as redes DenseNet169 e DenseNet201, com valores de 85% a 88%. Em contrapartida, mais uma vez, o SVM Polynomial não conseguiu pontuar no Coeficiente de Matthews, ou seja, ele classifica as amostras de forma completamente aleatória, sem nenhuma correlação.

Por fim, nas Tabelas 5 e 6 podemos ver os resultados de cada classificação feita para os fatores de magnificação de 200X e 400X, respectivamente. A combinação DenseNet201 e SVM RBF atingiu seus melhores valores entre todas as escalas de imagem para a variação de 200X, quando teve 95.38% de acurácia, com desvio-padrão de apenas 0.40%. A precisão e o recall também conseguiram aproximadamente 95.5% e 94.5%, respectivamente, ambos com desvio-padrão também baixo. No que se refere ao F1-Score e Coeficiente de Matthews (MCC), também obtiveram os melhores resultados até agora, com 93.7% e 89%.

Por outro lado, a magnificação de 400X apresentou uma baixa nos valores atingidos pelas métricas para todas as redes em estudo, mostrando que essa aproximação já torna mais difícil uma diferenciação pelas topologias de CNN, não mostrando muita diferença entre as classes. Apesar disso, a diferença é baixa nas acurácias, tendo a melhor combinação atingido 92.64% em média. Já o desvio-padrão se elevou para acima de 1.2% em praticamente todos os casos. Para o MCC que há uma maior diferença em comparação com as outras magnificações vistas, tendo seu melhor resultado novamente para a combinação SVM RBF com DenseNet201, atingindo quase 83% em média, com desvio-padrão de 2.7%.

Por fim, analisando os tempos de treino e predição das tabelas, percebe-se que eles dependem apenas dos classificadores e do tamanho do vetor de atributos oriundo da CNN. Logo, as faixas de tempo não variam muito a depender da tabela. É visível que o tempo de treinamento do MLP varia bastante em média, com desvio-padrão bem alto, e é em sua maioria bem alto e exige um alto nível de processamento devido à complexidade do seu algoritmo. Por outro lado, o classificador Nearest Neighbors, apesar de ser treinado rapidamente,

Feature Extractor	Classifier	Accuracy	Precision	Recall	F1-Score	Matthews	Train Time	Predict Time
DenseNet121	Bayes	71.94±0.83	81.81±4.83	54.92±1.15	50.71±2.04	24.85±4.12	0.043±0.025	0.106±0.029
	MLP	91.30±1.19	90.23±1.64	89.29±1.12	89.72±1.33	79.50±2.67	35.253±1.438	0.185±0.071
	Nearest Neighbors	88.80±0.83	89.40±1.59	84.01±0.69	86.03±0.92	73.20±2.10	0.105±0.012	18.264±2.092
	Random Forest	87.89±1.46	87.46±2.02	83.61±1.90	85.12±1.80	70.95±3.60	11.120±1.281	6.901±2.060
	SVM Linear	89.96±1.38	87.91±1.72	89.17±1.19	88.47±1.49	77.06±2.89	7.637±0.108	4.216±0.229
	SVM Polynomial	61.39±15.27	30.70±7.64	50.00±0.00	37.38±6.91	0.00±0.00	19.537±1.345	11.979±0.301
	SVM RBF	93.22±1.61	92.41±2.13	91.66±1.58	92.02±1.84	84.07±3.70	8.904±0.728	5.265±0.702
DenseNet169	Bayes	74.53±0.44	84.08±1.57	59.15±0.83	57.83±1.39	35.25±1.26	0.081±0.022	0.098±0.016
	MLP	92.02±1.31	90.82±1.43	90.54±1.92	90.63±1.59	81.35±3.12	161.009±12.553	0.693±0.269
	Nearest Neighbors	91.74±1.34	91.55±1.63	88.83±1.67	90.02±1.64	80.34±3.25	0.120±0.018	27.918±4.142
	Random Forest	87.60±2.88	86.67±3.88	83.78±3.14	84.97±3.39	70.38±6.89	10.790±1.243	7.001±1.888
	SVM Linear	91.78±0.77	90.16±0.98	90.88±1.00	90.48±0.87	81.03±1.74	11.390±0.739	6.684±0.187
	SVM Polynomial	61.39±15.27	30.70±7.64	50.00±0.00	37.38±6.91	0.00±0.00	29.500±1.170	19.383±0.132
	SVM RBF	93.37±0.97	92.20±1.15	**92.33±1.19**	92.26±1.12	84.53±2.25	11.090±0.714	6.682±0.528
DenseNet201	Bayes	79.24±0.51	85.59±0.75	67.10±0.90	69.01±1.12	49.32±1.44	0.102±0.046	0.149±0.021
	MLP	93.22±0.86	92.52±0.80	91.54±1.52	91.97±1.09	84.04±2.11	48.209±3.841	0.220±0.113
	Nearest Neighbors	91.93±1.13	92.09±1.22	88.80±1.74	90.18±1.45	80.81±2.77	0.141±0.020	31.707±3.707
	Random Forest	88.56±0.91	87.60±1.20	85.12±1.17	86.18±1.11	72.67±2.22	20.628±2.545	6.859±2.040
	SVM Linear	92.89±1.18	91.53±1.04	91.94±1.94	91.70±1.46	83.46±2.96	14.042±0.807	8.433±0.321
	SVM Polynomial	61.39±15.27	30.70±7.64	50.00±0.00	37.38±6.91	0.00±0.00	33.403±0.371	22.413±0.353
	SVM RBF	**94.18±1.34**	**94.23±1.16**	92.06±2.11	**93.01±1.67**	**86.25±3.19**	22.408±2.560	11.321±1.032
VGG16	Bayes	74.10±1.13	85.69±0.45	58.23±1.97	56.19±3.24	34.01±3.98	0.056±0.032	0.079±0.021
	MLP	88.13±1.38	87.11±1.97	84.59±1.47	85.66±1.61	71.65±3.30	12.570±0.536	0.064±0.042
	Nearest Neighbors	87.22±0.98	87.41±0.90	82.00±1.74	83.96±1.47	69.18±2.49	0.051±0.004	10.680±1.631
	Random Forest	86.64±0.99	84.96±1.31	83.30±1.44	84.00±1.23	68.23±2.46	15.960±2.162	5.360±1.444
	SVM Linear	86.78±1.72	84.35±2.07	85.50±1.90	84.82±1.92	69.83±3.84	4.979±0.444	2.451±0.122
	SVM Polynomial	61.39±15.27	30.70±7.64	50.00±0.00	37.38±6.91	0.00±0.00	11.150±0.800	6.053±0.088
	SVM RBF	90.05±1.05	90.71±1.73	85.77±1.33	87.67±1.31	76.30±2.67	10.122±0.926	4.319±0.213
VGG19	Bayes	77.27±0.73	84.17±2.36	63.92±0.90	64.87±1.20	43.60±2.62	0.056±0.034	0.080±0.023
	MLP	87.60±1.21	86.08±1.38	84.43±1.62	85.16±1.50	70.48±2.95	31.048±2.769	0.117±0.060
	Nearest Neighbors	86.30±1.57	86.69±2.09	80.49±2.14	82.62±2.11	66.88±4.04	0.046±0.004	10.614±0.866
	Random Forest	86.64±1.14	87.07±1.26	80.95±1.74	83.08±1.60	67.73±2.88	17.845±2.145	6.817±0.895
	SVM Linear	87.22±0.38	84.79±0.55	86.16±0.77	85.36±0.38	70.93±0.85	5.027±0.420	2.454±0.167
	SVM Polynomial	61.39±15.27	30.70±7.64	50.00±0.00	37.38±6.91	0.00±0.00	11.243±0.988	5.991±0.196
	SVM RBF	90.82±1.07	90.72±1.34	87.62±2.06	88.85±1.45	78.25±2.70	8.794±0.905	3.290±0.234

Table 4. Metrics (%), train time (ms) and predict time (ms) obtained by each combination of feature extraction and classifier for the 100X magnification factor. In bold, the best results per metric and the best combination method.

prediz de forma mais lenta dentre as outras opções de classificadores. No mais, os tempos da arquitetura VGG (seja 16 ou 19) são menores que os tempos da arquitetura DenseNet, pois a quantidade de atributos no vetor de cada amostra é, pelo menos, metade da quantidade de atributos resultantes da arquitetura concorrente. Mesmo assim, salienta-se que os valores estão em milissegundos, e portanto, não há uma diferença gritante entre os valores.

5.1. Comparação da melhor combinação com os métodos da literatura para cada fator de magnificação

Visto que para todos os fatores de magnificação são analisados a melhor combinação de CNN-classificador foi usando a arquitetura DenseNet201 e o classificador SVM RBF, Desta forma, foram usados os resultados desse modelo para comparação com outros trabalhos da literatura como apresentados na tabela 7. Da mesma forma, esses autores também classificaram de forma binária as imagens, distinguindo os fatores de magnificação 40X, 100X, 200X e 400X.

A Tabela 7 mostra as acurácias comparadas o melhor modelo deste estudo com outros métodos da literatura que também fizeram a distinção benigno-maligno, separando o dataset por fator de magnificação. Assim como foi especificado na Seção 2, Song *et al.* [20] usaram CNN-based FV descriptor with adaptation layer para classificar o dataset entre as duas classes. Por outro lado, Zhi *et al.* [18] também usaram a técnica de transfer learning com VGGNet-based architeture custom model with a patch-based augmentation. Por fim, Deniz *et al.* [21] fizeram um fine-tuning no deep model AlexNet para a demanda, conseguindo resultados melhores do que nas suas outras tentativas com uso da VGG16 e dos vetores concatenados da VGG16 e da AlexNet.

Ainda na Tabela 7, é possível ver que o nosso método, tratando-se de acurácia, bate os outros métodos da literatura para os fatores de magnificação de 40X, 200X e 400X. Além disso, também chegamos bem próximo da acurácia média de Zhi *et al.* [18], melhor resultado para o fator de magnificação de 100X, mesmo sem usar a técnica de augmentation usada pelo autor, e apresentamos um desvio-padrão bem menor em comparação com o dele (o dobro do obtido por este estudo).

Os mesmos resultados podem ser vistos no gráfico da Figura 4, para efeito de melhor comparação entre as acurácias obtidas pelos métodos. Pela figura fica claro que a acurácia atingida por nosso método apresenta um desvio-padrão menor para praticamente todos os fatores de magnificação, em comparação com outros métodos.

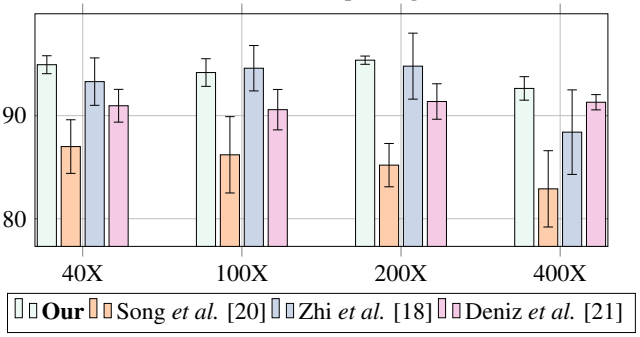
Como pode ser observado na tabela 8, o teste estatístico de Kolmogorov-Smirnov recusou a hipótese, com um α de 1%, de que os vetores de atributos possuem similaridade estatística, essa recusa mesmo para um valor de α tão pequeno, pode ter se dado devido á densa quantidade de amostras em cada um dos vetores de atributos utilizados para gerar as médias e desvio padrões observados na tabela 7.

As comparações feitas na tabela e na respectiva figura são

Feature Extractor	Classifier	Accuracy	Precision	Recall	F1-Score	Matthews	Train Time	Predict Time
DenseNet121	Bayes	79.58±1.47	83.35±1.20	68.43±2.56	70.46±3.00	49.49±4.02	0.053±0.036	0.094±0.041
	MLP	90.11±1.45	88.49±1.86	88.41±1.58	88.43±1.66	76.90±3.33	246.854±12.726	0.685±0.199
	Nearest Neighbors	90.76±1.98	90.34±2.05	87.68±2.87	88.80±2.53	77.96±4.86	0.114±0.011	17.307±2.990
	Random Forest	89.32±2.40	88.89±2.51	85.66±3.55	86.94±3.12	74.44±5.89	16.204±1.637	6.762±1.985
	SVM Linear	89.62±1.36	87.37±1.56	89.38±1.47	88.20±1.51	76.72±2.92	6.618±0.556	3.454±0.197
	SVM Polynomial	53.73±18.68	26.86±9.34	50.00±0.00	33.92±8.45	0.00±0.00	18.467±0.494	11.247±0.233
	SVM RBF	93.49±0.53	93.33±0.90	91.30±0.60	92.21±0.62	84.60±1.27	14.919±0.821	6.405±0.231
DenseNet169	Bayes	77.55±0.85	82.71±2.06	64.78±1.33	66.03±1.71	43.94±2.69	0.097±0.032	0.064±0.010
	MLP	92.10±0.95	91.40±1.51	90.12±1.71	90.62±1.16	81.49±2.39	374.174±24.015	1.402±0.838
	Nearest Neighbors	92.94±1.10	92.61±1.33	90.68±1.37	91.56±1.34	83.27±2.67	0.109±0.024	26.162±2.925
	Random Forest	87.88±0.87	86.21±1.12	85.11±1.07	85.61±1.03	71.31±2.06	10.352±1.196	6.788±1.869
	SVM Linear	91.36±0.79	89.51±0.72	90.82±1.73	90.05±1.03	80.31±2.25	11.441±0.548	6.718±0.269
	SVM Polynomial	53.73±18.68	26.86±9.34	50.00±0.00	33.92±8.45	0.00±0.00	27.726±1.383	18.152±0.530
	SVM RBF	93.34±0.90	92.74±1.30	91.68±1.51	92.12±1.07	84.39±2.19	10.987±0.577	6.693±0.475
DenseNet201	Bayes	82.96±1.29	87.35±2.17	73.40±2.05	76.24±2.19	59.08±3.54	0.089±0.035	0.108±0.027
	MLP	92.40±0.29	91.67±0.38	90.38±0.60	90.97±0.38	82.03±0.73	396.417±36.733	2.052±0.802
	Nearest Neighbors	93.54±0.77	93.64±0.84	91.17±1.40	92.23±0.99	84.75±1.81	0.136±0.019	30.310±4.724
	Random Forest	89.87±0.67	89.77±1.06	86.15±1.19	87.61±0.90	75.82±1.66	13.735±1.806	6.778±1.913
	SVM Linear	90.61±0.68	88.93±1.06	89.49±1.11	89.11±0.72	78.40±1.36	9.727±0.710	5.600±0.129
	SVM Polynomial	53.73±18.68	26.86±9.34	50.00±0.00	33.92±8.45	0.00±0.00	31.196±0.650	21.026±0.403
	SVM RBF	**95.38±0.40**	**95.43±0.59**	**93.69±0.51**	**94.49±0.48**	**89.10±0.97**	15.662±0.859	10.064±0.493
VGG16	Bayes	74.66±1.27	84.11±2.80	59.33±1.90	58.07±2.84	35.59±4.74	0.048±0.038	0.077±0.024
	MLP	89.47±1.45	88.21±1.82	86.88±1.83	87.47±1.74	75.07±3.46	31.354±4.076	0.116±0.057
	Nearest Neighbors	87.83±0.52	87.40±0.64	83.43±0.80	85.01±0.70	70.72±1.33	0.052±0.008	10.039±0.536
	Random Forest	86.14±0.44	84.79±0.70	81.99±0.65	83.15±0.52	66.72±0.99	13.143±1.720	5.684±1.489
	SVM Linear	88.82±0.97	86.69±1.34	87.70±0.85	87.12±1.00	74.38±1.97	4.623±0.359	2.191±0.121
	SVM Polynomial	53.73±18.68	26.86±9.34	50.00±0.00	33.92±8.45	0.00±0.00	10.814±0.655	5.624±0.192
	SVM RBF	92.20±0.79	91.77±0.60	89.75±1.36	90.65±1.03	81.49±1.95	8.148±0.585	3.245±0.268
VGG19	Bayes	80.38±1.24	85.14±2.37	69.32±1.72	71.64±2.06	52.08±3.79	0.047±0.030	0.068±0.019
	MLP	90.66±1.61	89.94±1.72	87.88±2.31	88.77±2.03	77.78±3.94	21.890±1.504	0.116±0.028
	Nearest Neighbors	86.74±0.50	87.56±0.80	80.79±0.54	83.10±0.60	68.01±1.25	0.044±0.003	9.848±1.696
	Random Forest	87.43±1.21	87.84±1.59	82.13±1.62	84.19±1.59	69.73±3.05	12.120±1.257	8.292±2.995
	SVM Linear	89.37±1.37	87.30±1.50	88.23±1.86	87.71±1.63	75.51±3.27	4.235±0.394	2.032±0.076
	SVM Polynomial	53.73±18.68	26.86±9.34	50.00±0.00	33.92±8.45	0.00±0.00	10.381±0.843	5.611±0.143
	SVM RBF	92.35±1.53	92.33±1.78	89.55±2.02	90.75±1.90	81.82±3.73	7.830±0.676	2.938±0.194

Table 5. Metrics (%), train time (ms) and predict time (ms) obtained by each combination of feature extraction and classifier for the 200X magnification factor. In bold, the best results per metric and the best combination method.

Fig. 4. Comparison between accuracies (%) reached by our method and by the others methods from related works per magnification factor.



apenas em nível de acurácia média, pois não foram informados os valores de outras métricas pelos autores dos trabalhos em comparação.

5.2. Comparação da melhor combinação com os métodos da literatura para o dataset completo

Um último experimento foi feito, com todas as imagens do dataset, com classificação supervisionada entre as classes benigno e maligno, dessa vez sem distinção de magnificação entre as imagens. O modelo gerado e escolhido por este estudo, com base nas Tabelas 3 a 6, foi também o que usa a rede DenseNet201 com o classificador SVM RBF.

Os resultados obtidos foram de 94.88±0.57 de acurácia, 94.64±0.49 de precisão, 93.97±0.70 de recall, 93.38±0.89 de F1-Score e 88.00±1.35 para o Coeficiente de Correlação de Matthews (MCC). Em geral, os resultados são equiparáveis aos resultados obtidos individualmente para cada fator de magnificação. Se comparados, por exemplo, com a Tabela 5, que mostra os resultados dos modelos para as imagens de 200X, houve uma queda nas métricas de avaliação. Em contrapartida, para os resultados das imagens de 400X, mostrados na Tabela 6, houve uma melhora de 2% na acurácia média, com diminuição do desvio-padrão, e uma melhora ainda mais notável para as demais métricas, culminando em um MCC quase 6% maior e com desvio-padrão mais baixo cerca de 50%.

Esses resultados se dão, muito provavelmente, pelo modelo discernir muito melhor as imagens com magnificação na faixa de 200X e começar a confundir as imagens com magnificação mais alta, de 400X. Dessa forma, muitos autores (citados acima) ainda defendem a separação do dataset em questão em diferentes fatores de magnificação para encontrar a melhor maneira de classificá-lo.

Entre os autores citados, apenas Mehra *et al.* [19] e Celik *et al.* [17] fizeram um estudo com todas as imagens, generalizando as escalas de 40X, 100X, 200X e 400X em um único dataset, sendo esse último restrito a detecção de invasive ductal carcinoma (IDC), como mencionado na Seção 2, e por isso não se enquadra em comparação com os nossos resultados. Já Mehra *et al.* testou um treinamento total ou par-

Feature Extractor	Classifier	Accuracy	Precision	Recall	F1-Score	Matthews	Train Time	Predict Time
DenseNet121	Bayes	73.29±1.41	79.34±3.71	59.47±2.01	58.12±2.98	33.27±5.57	0.064±0.053	0.071±0.038
	MLP	89.23±1.91	87.87±2.43	87.46±2.02	87.64±2.16	75.33±4.36	140.998±5.360	0.373±0.183
	Nearest Neighbors	88.74±1.27	88.32±1.42	85.46±1.63	86.65±1.54	73.71±3.00	0.087±0.012	15.624±0.697
	Random Forest	86.92±0.88	86.04±1.01	83.49±1.49	84.52±1.18	69.47±2.19	11.939±1.040	6.299±0.245
	SVM Linear	89.18±0.81	87.55±1.08	87.87±0.60	87.69±0.83	75.42±1.63	7.269±0.156	3.961±0.035
	SVM Polynomial	53.48±17.35	26.74±8.67	50.00±0.00	33.96±7.83	0.00±0.00	15.096±0.184	9.344±0.315
	SVM RBF	92.42±0.64	92.28±0.84	90.22±0.89	91.12±0.77	82.47±1.50	11.703±0.353	4.953±0.378
DenseNet169	Bayes	75.60±1.32	80.17±2.06	63.49±2.00	63.97±2.66	40.28±3.99	0.083±0.028	0.064±0.018
	MLP	90.06±1.01	89.28±1.36	87.81±1.34	88.43±1.18	77.06±2.30	146.267±5.852	0.571±0.242
	Nearest Neighbors	90.33±1.28	89.52±1.09	88.10±2.02	88.73±1.64	77.60±3.11	0.111±0.013	24.793±0.462
	Random Forest	87.37±2.04	85.93±2.30	84.90±2.58	85.34±2.41	70.81±4.77	7.231±0.218	6.666±0.267
	SVM Linear	90.11±1.60	88.43±1.70	89.32±2.14	88.82±1.85	77.74±3.76	11.047±0.114	6.557±0.127
	SVM Polynomial	53.48±17.35	26.74±8.67	50.00±0.00	33.96±7.83	0.00±0.00	24.843±0.219	16.189±0.172
	SVM RBF	92.42±1.72	91.88±1.76	**90.63±2.30**	91.19±2.05	82.49±4.02	13.303±1.016	7.689±0.199
DenseNet201	Bayes	82.53±1.98	86.02±2.03	74.21±2.95	76.69±3.15	59.00±5.00	0.099±0.021	0.113±0.026
	MLP	90.71±1.27	89.90±1.34	88.65±1.77	89.21±1.56	78.53±3.03	84.052±8.840	0.352±0.129
	Nearest Neighbors	90.82±1.20	90.63±0.87	88.15±2.14	89.17±1.60	78.72±2.93	0.146±0.020	27.270±0.638
	Random Forest	89.07±1.93	89.52±2.03	85.12±2.69	86.79±2.47	74.50±4.65	19.350±1.236	9.986±0.552
	SVM Linear	90.71±0.90	89.36±0.98	89.45±1.21	89.40±1.07	78.81±2.13	10.822±0.217	6.512±0.101
	SVM Polynomial	53.48±17.35	26.74±8.67	50.00±0.00	33.96±7.83	0.00±0.00	26.524±0.138	17.781±0.120
	SVM RBF	**92.64±1.14**	**92.41±1.48**	90.60±1.33	**91.41±1.34**	**82.98±2.70**	21.835±0.409	10.720±0.385
VGG16	Bayes	72.25±1.22	79.07±2.96	57.72±1.90	55.29±3.10	29.78±4.59	0.065±0.039	0.088±0.024
	MLP	86.65±1.23	85.16±1.47	83.96±1.68	84.48±1.52	69.11±2.97	17.253±1.703	0.108±0.057
	Nearest Neighbors	84.28±1.16	82.54±1.28	80.88±1.68	81.57±1.46	63.39±2.82	0.042±0.002	9.316±0.315
	Random Forest	85.82±1.33	84.65±1.03	82.24±2.34	83.20±1.89	66.83±3.42	9.322±0.250	6.547±0.507
	SVM Linear	86.54±1.02	84.38±1.21	85.66±1.10	84.92±1.10	70.03±2.18	4.984±0.175	2.443±0.038
	SVM Polynomial	53.48±17.35	26.74±8.67	50.00±0.00	33.96±7.83	0.00±0.00	9.471±0.179	5.108±0.041
	SVM RBF	88.52±1.27	87.09±1.60	86.54±1.34	86.79±1.42	73.62±2.86	5.469±0.259	2.419±0.158
VGG19	Bayes	77.31±1.11	81.62±1.19	66.22±1.81	67.51±2.37	45.22±2.96	0.046±0.036	0.072±0.031
	MLP	87.64±2.03	86.12±2.22	85.40±2.62	85.72±2.41	71.51±4.77	40.986±1.050	0.156±0.061
	Nearest Neighbors	85.71±1.55	85.42±1.74	81.13±2.35	82.68±2.09	66.39±3.86	0.042±0.003	9.172±0.274
	Random Forest	86.98±1.33	87.07±1.49	82.65±2.32	84.22±1.85	69.55±3.30	14.121±0.705	7.529±0.572
	SVM Linear	87.69±1.22	85.88±1.61	86.65±1.79	86.11±1.35	72.50±2.84	4.406±0.190	2.112±0.106
	SVM Polynomial	53.48±17.35	26.74±8.67	50.00±0.00	33.96±7.83	0.00±0.00	9.090±0.178	4.838±0.045
	SVM RBF	89.67±1.00	88.18±1.57	88.60±1.19	88.28±1.05	76.75±2.17	4.721±0.096	2.435±0.053

Table 6. Metrics (%), train time (ms) and predict time (ms) obtained by each combination of feature extraction and classifier for the 400X magnification factor. In bold, the best results per metric and the best combination method.

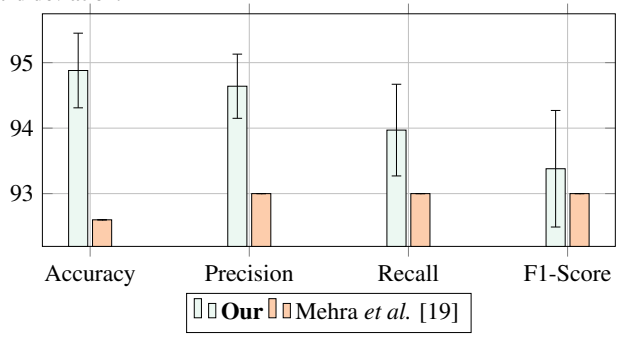
Table 7. Mean accuracies (%) and their respective standard deviations of our best method (DenseNet201 + SVM RBF) in comparison to other methods from related works for each magnification factor. The other autors only presented accuracy as evaluation metric.

Method	Magnification Factors			
	40X	100X	200X	400X
Our best method	**94.94±0.87**	94.18±1.34	**95.38±0.40**	**92.64±1.14**
Song *et al.* [20].	87.00±2.60	86.20±3.70	85.20±2.10	82.90±3.70
Zhi *et al.* [18].	93.30±2.30	**94.60±2.20**	94.80±3.20	88.40±4.10
Deniz *et al.* [21].	90.96±1.59	90.58±1.96	91.37±1.72	91.30±0.74

Table 8. Resultado do teste de Kolmogorov-Smirnov Para o melhor método proposto por este estudo e métodos encontrados na literatura com um α de 1%.

Method	Magnification Factors			
	40X	100X	200X	400X
Our X Song *et al.* [20].	≠	≠	≠	≠
Our X Zhi *et al.* [18].	≠	≠	≠	≠
Our X Deniz *et al.* [21].	≠	≠	≠	≠

Fig. 5. Comparison between our method and the other method from related works per metric. The results below is from using of all dataset, without magnification factor distinction. The other author did not specify the stardard deviation.



cial de redes neurais convolucionais (modelos VGG16, VGG19 e ResNet50) também com o dataset BreaKHis, mas de forma balanced and augmented. A Tabela 9 faz a comparação do nosso melhor método (transfer learning com DenseNet201 + SVM RBF) com o melhor método dentre os usados pelo autor citado (fazendo um fine-tuning com o modelo de VGG16, com regressão logística).

Table 9. Performance of our best method (DenseNet201 + SVM RBF) in comparison to other from related works for a classification experiment using all BreaKHis dataset, without magnification factor distinction. The other author did not present MCC or time values.

Method	Métricas			
	Accuracy	Precision	Recall	F1-Score
Our best method	**94.88±0.57**	**94.64±0.49**	**93.97±0.70**	**93.38±0.89**
Mehra *et al.* [19].	92.60	0.93	0.93	0.93

Da mesma forma, a Figura 5 ilustra a comparação entre os métodos, mostrada em números na Tabela 9. Pela análise da figura, percebe-se que, mesmo sem treinar a rede por meio de um fine-tuning e sem balancear ou aumentar o dataset, os resultados obtidos pelo nosso método de transfer learning são superiores aos resultados do autor para as métricas em questão. Os desvios-padrões do método em comparação não foram especificados pelo autor, deixando uma lacuna referente a validação dos resultados e aos possíveis altos níveis de desvios-padrões comprometendo os reais valores obtidos pelo estudo apresentado.

6. Conclusions and Future Work

Este estudo apresenta um método para classificar imagens de tecidos compostas de exames histopatológicos em quatro escalas binárias entre benigno e maligno. A abordagem é dividida em duas etapas: I) extração de características usando CNNs por meio da técnica de transferência de aprendizagem, e II) classificação automática de imagens usando métodos de aprendizado de máquina.

Os resultados mostram que as CNNs, combinadas com a técnica de transferência de aprendizagem e métodos de aprendizagem de máquinas podem ser usados como extratores de recursos para este problema. Assim, Podemos afirmar que os experimentos deste estudo alcançaram baixo custos computacional, dispondo de uma acurácia de 95.38% com o extrator DenseNet201 aliado ao classificador SVM RBF na classe de 200x entre malígno e benígno, com uso da base de dados BreaKHis, além de F1-Score 93.69% e 10 ms no tempo de teste médio, como monstra a tabela 5. Nesse sentido, o método proposto com uso do melhor modelo (Modelo:) trás ganhos significativos para aplicações médicas para auxílio ao pré-diagnóstico com o objetivo de identificar de tecidos em exames histopatológicos com excelentes resultados.

Portanto, o estudo tem como foco, intensificar as análises, buscando parâmetros que viabilizem cada vez mais um melhores resultados. Para trabalhos, futuros, é tido como proposta, aplicação do modelo deste estudo para diferente imagens de tecidos, como melanoma e diferentes tipos de doenças de pele, afim de avaliar a generalização do modelo. para diferentes tipos de problemas e dataset.

Acknowledgments

Acknowledgments should be inserted at the end of the paper, before the references, not as a footnote to the title. Use the unnumbered Acknowledgements Head style for the Acknowledgments heading.

References

[1] Syed, L, Jabeen, S, Manimala, S. Telemammography: a novel approach for early detection of breast cancer through wavelets based image processing and machine learning techniques. In: Advances in Soft Computing and Machine Learning in Image Processing. Springer; 2018, p. 149–183.

[2] Organization, WH. Breast cancer. ???? URL: `https://www.who.int/news-room/fact-sheets/detail/breast-cancer#:~:text=In%202020%2C%20there%20were%202.3,the%20world%27s%20most%20prevalent%20cancer.`

[3] De Vrieze, T, Nevelsteen, I, Thomis, S, De Groef, A, Tjalma, WA, Gebruers, N, et al. What are the economic burden and costs associated with the treatment of breast cancer-related lymphoedema? a systematic review. Supportive Care in Cancer 2020;28(2):439–449.

[4] Nuciforo, S, Fofana, I, Matter, MS, Blumer, T, Calabrese, D, Boldanova, T, et al. Organoid models of human liver cancers derived from tumor needle biopsies. Cell reports 2018;24(5):1363–1376.

[5] Pugliese, N, Di Perna, M, Cozzolino, I, Ciancia, G, Pettinato, G, Zeppa, P, et al. Randomized comparison of power doppler ultrasonography-guided core-needle biopsy with open surgical biopsy for the characterization of lymphadenopathies in patients with suspected lymphoma. Annals of hematology 2017;96(4):627–637.

[6] Komura, D, Ishikawa, S. Machine learning methods for histopathological image analysis. Computational and structural biotechnology journal 2018;16:34–42.

[7] Qi, Q, Li, Y, Wang, J, Zheng, H, Huang, Y, Ding, X, et al. Label-efficient breast cancer histopathological image classification. IEEE journal of biomedical and health informatics 2018;23(5):2108–2116.

[8] Bhise, V, Rajan, SS, Sittig, DF, Morgan, RO, Chaudhary, P, Singh, H. Defining and measuring diagnostic uncertainty in medicine: a systematic review. Journal of general internal medicine 2018;33(1):103–115.

[9] Robertson, S, Azizpour, H, Smith, K, Hartman, J. Digital image analysis in breast pathology—from image processing techniques to artificial intelligence. Translational Research 2018;194:19–35.

[10] Liakos, KG, Busato, P, Moshou, D, Pearson, S, Bochtis, D. Machine learning in agriculture: A review. Sensors 2018;18(8):2674.

[11] Chen, J, Ran, X. Deep learning with edge computing: A review. Proceedings of the IEEE 2019;107(8):1655–1674.

[12] Kooi, T, Litjens, G, Van Ginneken, B, Gubern-Mérida, A, Sánchez, CI, Mann, R, et al. Large scale deep learning for computer aided detection of mammographic lesions. Medical image analysis 2017;35:303–312.

[13] Sharma, S, Mehra, R. Conventional machine learning and deep learning approach for multi-classification of breast cancer histopathology images—a comparative insight. Journal of digital imaging 2020;33(3):632–654.

[14] Benhammou, Y, Achchab, B, Herrera, F, Tabik, S. Breakhis based breast cancer automatic diagnosis using deep learning: Taxonomy, survey and insights. Neurocomputing 2020;375:9–24.

[15] Tsochatzidis, L, Costaridou, L, Pratikakis, I. Deep learning for breast cancer diagnosis from mammograms—a comparative study. Journal of Imaging 2019;5(3):37.

[16] Ohata, EF, das Chagas, JVS, Bezerra, GM, Hassan, MM, de Albuquerque, VHC, Reboucas Filho, PP. A novel transfer learning approach for the classification of histological images of colorectal cancer. The Journal of Supercomputing 2021;:1–26.

[17] Celik, Y, Talo, M, Yildirim, O, Karabatak, M, Acharya, UR. Automated invasive ductal carcinoma detection based using deep transfer learning with whole-slide images. Pattern Recognition Letters 2020;133:232–239.

[18] Zhi, W, Yueng, HWF, Chen, Z, Zandavi, SM, Lu, Z, Chung, YY. Using transfer learning with convolutional neural networks to diagnose breast cancer from histopathological images. In: International Conference on Neural Information Processing. Springer; 2017, p. 669–676.

[19] Mehra, R, et al. Breast cancer histology images classification: Training from scratch or transfer learning? ICT Express 2018;4(4):247–254.

[20] Song, Y, Zou, JJ, Chang, H, Cai, W. Adapting fisher vectors for histopathology image classification. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). IEEE; 2017, p. 600–603.

[21] Deniz, E, Şengür, A, Kadiroğlu, Z, Guo, Y, Bajaj, V, Budak, Ü. Transfer learning based histopathologic image classification for breast cancer detection. Health information science and systems 2018;6(1):1–7.

[22] Zhu, Z, Albadawy, E, Saha, A, Zhang, J, Harowicz, MR, Mazurowski, MA. Deep learning for identifying radiogenomic associations in breast cancer. Computers in biology and medicine 2019;109:85–90.
[23] Paul, R, Hawkins, SH, Balagurunathan, Y, Schabath, MB, Gillies, RJ, Hall, LO, et al. Deep feature transfer learning in combination with traditional features predicts survival among patients with lung adenocarcinoma. Tomography 2016;2(4):388.
[24] Orenstein, EC, Beijbom, O. Transfer learning and deep feature extraction for planktonic image data sets. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE; 2017, p. 1082–1088.
[25] Tajbakhsh, N, Shin, JY, Gurudu, SR, Hurst, RT, Kendall, CB, Gotway, MB, et al. Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE transactions on medical imaging 2016;35(5):1299–1312.
[26] Deng, J, Dong, W, Socher, R, Li, LJ, Li, K, Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. Ieee; 2009, p. 248–255.
[27] Simonyan, K, Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556 2014;.
[28] Huang, G, Liu, Z, Van Der Maaten, L, Weinberger, KQ. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, p. 4700–4708.
[29] Theodoridis, S, Koutroumbas, K. Pattern recognition, academic press. Burlington, MA[Google Scholar] 2008;.
[30] Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review 1958;65(6):386.
[31] Haykin, SS, et al. Neural networks and learning machines/simon haykin. 2009.
[32] Aha, DW, Kibler, D, Albert, MK. Instance-based learning algorithms. Machine learning 1991;6(1):37–66.
[33] Ho, TK. The random subspace method for constructing decision forests. IEEE transactions on pattern analysis and machine intelligence 1998;20(8):832–844.
[34] Vapnik, V. N.(1998). statistical learning theory. ????
[35] Xiao, Y. A fast algorithm for two-dimensional kolmogorov–smirnov two sample tests. Computational Statistics & Data Analysis 2017;105:53–58.
[36] Ho, J, Tumkaya, T, Aryal, S, Choi, H, Claridge-Chang, A. Moving beyond p values: data analysis with estimation graphics. Nature methods 2019;16(7):565–566.
[37] Spanhol, FA, Oliveira, LS, Petitjean, C, Heutte, L. A dataset for breast cancer histopathological image classification. Ieee transactions on biomedical engineering 2015;63(7):1455–1462.