Avaliação de abordagens semi-supervisionadas aplicadas a redes neurais convolucionais

Cristiano N. de O. Bassani[1], Priscila T. M. Saito[1,2], Pedro H. Bugatti[1],

[1]Departamento de Computação,
Universidade Tecnológica Federal do Paraná - Cornélio Procópio (UTFPR)
Avenida Alberto Carazzai, 1640, 86300-000, Cornélio Procópio, PR, Brazil

[2]Departamento de Computação,
Universidade Federal de São Carlos (UFSCar)
Rodovia Washington Luís, km 235, 13565-905, São Carlos, SP, Brasil

crisbassani94@gmail.com, priscilasaito@ufscar.br, pbugatti@utfpr.edu.br

Abstract. *In this work we focus on mitigating the general problem of sample labeling in image datasets used in convolutional neural networks. To do so, we apply the semi-supervised paradigm to convolutional neural networks. We proposed the comparation of two semi-supervised techniques from the literature with this type of deep learning networks and analyzed their behavior. Our experiments, with three public data sets, testify that our proposed aggregation obtained better results, gains of up to 88% in accuracy, when compared to the supervised paradigm.*

Resumo. *O presente trabalho foca em mitigar o problema geral de rotulação de amostras em bases de imagens utilizadas em redes neurais convolucionais. Para isso, foi aplicado o paradigma semi-supervisionado às redes neurais convolucionais. Foi proposta a comparação de duas técnicas semi-supervisionadas da literatura com este tipo de rede de aprendizado profundo e analizado o seu comportamento. Os experimentos, com três conjuntos de dados públicos, atestam que a agregação proposta obteve melhores resultados, ganhos de até 88% em acurácia, quando comparado com o paradigma supervisionado.*

1. Introdução

Existem várias abordagens de aprendizado de máquina na literatura, cada uma é melhor aplicável do que outras em diferentes problemas e situações. Cada paradigma de aprendizado de máquina possui uma gama de aplicabilidade e usabilidade [van Engelen and Hoos 2019, Rosenberg et al. 2005].

Neste trabalho, foca-se no paradigma semi-supervisionado aplicado às redes neurais convolucionais (CNNs). Tais redes vêm apresentando ótimos resultados no processo de classificação de imagens. No entanto, as CNNs requerem grande volume de dados rotulados para treinar um modelo de aprendizado, fato este que prejudica em grande parte sua aplicabilidade em diversos contextos onde há carência de amostras rotuladas.

Nesse sentido, o presente trabalho visa realizar análises entre duas principais abordagens semi-supervisionadas. De acordo com os experimentos realizados, pode-se notar que amostras pseudo-rotuladas podem gerar resultados equivalentes (em termos de

acurácia) ao paradigma de aprendizado supervisionado (o qual, no entanto, requer 100% de amostras rotuladas). Foram avaliados diferentes tipos de métricas como acurácia, precisão, revocação e f1-score para obter uma análise ampla e profunda das CNNs agregadas com as abordagens semi-supervisionadas.

Alguns trabalhos da literatura, como o de [Chen et al. 2020], realizam um pré-treinamento das CNNs com pseudorótulos e em seguida aplicam os pesos aprendidos a uma segunda CNN. A questão é que isso pode gerar um impacto maior ou até mesmo possíveis problemas de overfitting. Outra diferença é que, neste trabalho foi utilizada a abordagem de *transfer learning* por meio do ImageNet [Deng et al. 2009, Zhuang et al. 2019], para obter análises sem acréscimo de custo computacional.

As **contribuições** do presente trabalho referem-se a três aspectos: (i) realização de uma avaliação das abordagens semi-supervisionadas de self-training e co-training quanto ao seu comportamento agregado às CNNs; (ii) aplicação da abordagem de *transfer learning* no processo para verificar sua generalização sob o paradigma semi-supervisionado; (iii) análises considerando diferentes métricas e três conjuntos de dados públicos.

2. Conceitos Relacionados

Relacionado ao paradigma semi-supervisionado, foram consideradas as abordagens de self-training [Amorim et al. 2019] e co-training [Chen et al. 2020], dado que são amplamente utilizadas na literatura.

A abordagem self-training, uma das mais básicas e conhecidas do paradigma semi-supervisionado [Amorim et al. 2019], faz uso do processo de pseudo-rotulagem para classificar as amostras não rotuladas e, em seguida, tais amostras são adicionadas incrementalmente ao modelo de treinamento. A abordagem co-training também utiliza o processo de pseudo-rotulagem. No entanto, a mesma utiliza o treinamento de vários modelos de aprendizado nas mesmas amostras para validar o processo de rotulagem.

Em relação às redes neurais convolucionais, neste trabalho, as mesmas foram utilizadas como um processo end-to-end agregado às abordagens semi-supervisionadas. Para isso, foram consideradas duas arquiteturas de última geração chamadas Xception [T.R. et al. 2019] e ResNet152_V2 [Elshennawy and Ibrahim 2020]. Em [Zhang et al. 2017], as CNNs também foram usadas para classificar células cervicais.

3. Metodologias Propostas

Associou-se o paradigma semi-supervisionado com CNNs para classificar imagens de 3 conjuntos de dados públicos. Para isso, foram utilizadas as arquiteturas CNN Xception e ResNet152_V2. Foram definidos os mesmos hiperparâmetros para ambas as arquiteturas (por exemplo, taxas de aprendizado, otimizador, entre outros), redimensionamento de imagem, regularização e normalização de dados.

Primeiro, considerou-se a abordagem de self-training semi-supervisionado com a arquitetura Xception. O self-training seleciona uma parte das amostras rotuladas para treinar a CNN. Em seguida, um modelo de aprendizado é gerado para classificar as amostras não rotuladas do conjunto de treinamento. Cada uma dessas amostras, rotuladas com pseudo-rótulos, são então agregadas incrementalmente ao conjunto rotulado para retreinar a CNN. A Figura 1a mostra o pipeline da metodologia de self-training proposta, bem como ilustra o processo de geração de pseudo-rótulos para amostras não rotuladas.

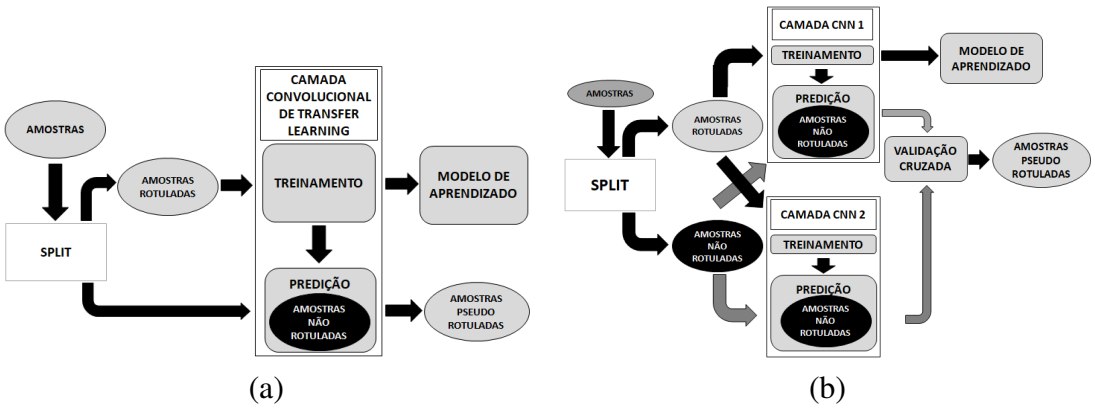


Figura 1. Arquiteturas semi-supervisionadas utilizadas. (a) pseudorotulagem no processo de self-training usando tranfer learning. (b) pseudorotulagem no processo de co-training usando transfer learning.

Em relação à abordagem de co-training, duas CNNs diferentes (Xception e ResNet152_v2) são consideradas para gerar os pseudo-rótulos, juntamente com o processo de validação mútua dos rótulos previstos por ambas as CNNs. Foram utilizadas as amostras rotuladas igualmente pelas duas arquiteturas. No caso do co-training, as mesmas amostras são submetidas a duas CNNs simultaneamente. Em seguida, são gerados dois conjuntos de pseudo-rótulos, que passam por uma validação mútua para atribuir um pseudo-rótulo a uma determinada amostra não rotulada. Uma vez finalizada a pseudo-rotulagem, as seguintes redes, previamente treinadas, são recarregadas e o novo conjunto de treinamento (conjunto de pseudo-rótulo) é submetido para treinar a CNN.

Inicialmente, considerando ambas as abordagens, todas as CNNs foram inicializadas com pesos (transfer learning) do ImageNet (pré-treinados). A Figura 1b mostra o processo de pseudo-rotulagem realizado pela abordagem de co-training proposta. O processo de retreinamento é idêntico ao do self-training.

Experimentalmente, dividiu-se cada conjunto de imagens em um conjunto de treinamento e um conjunto de teste, considerando 80% e 20%, respectivamente. Em seguida, o conjunto de treinamento foi dividido em conjuntos rotulados e não rotulados. Para melhor analisar as metodologias propostas foram consideradas diferentes divisões de conjuntos rotulados (10%, 20%, 30%, 40% e 50%).

4. Experimentos

4.1. Descrição dos conjuntos de dados

Os experimentos foram realizados considerando os conjuntos de dados públicos Papsmear [Zhang et al. 2017], MAMMOSET [Oliveira et al. 2017] e WHOI-Plankton [Orenstein et al. 2015]. O conjunto de dados Papsmear, obtido a partir do banco de dados da plataforma Kaggle, é baseado em um conjunto de imagens de células categorizadas em diferentes subclasses de células anormais e normais.

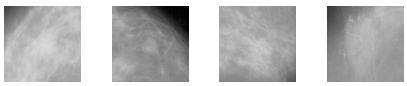
O conjunto de dados MAMMOSET compreende regiões de interesse (ROIs) obtidas a partir de exames mamográficos e é composto por 4 classes relacionadas a regiões de calcificação e massa, considerando nódulos benignos ou malignos. O terceiro conjunto de dados, denominado WHOI-Plankton, apresenta 103 classes. É composto por milhões

de imagens de diferentes tipos de plâncton marinho. No entanto, devido a limitações de espaço e custo computacional, para o conjunto WHOI-Plankton, foi utilizado um subconjunto aleatório de 14 classes.

No conjunto de dados do Papsmear, as imagens têm dimensões que variam de 53x129 pixels a 252x28 pixels, além de quantidades de amostras balanceadas. As imagens estão no espaço de cores RGB com profundidade de 8 bits. Por outro lado, o conjunto de dados MAMMOSET possui imagens (ROIs) em escala de cinza, com dimensões iguais de 1000x1000 pixels. Por fim, o conjunto de dados WHOI-Plankton apresenta imagens em escala de cinza com dimensões variadas. As Figuras 2 a 4 mostram exemplos de imagens de cada classe considerando os conjuntos de dados Papsmear, MAMMOSET e WHOI-Plankton, respectivamente.

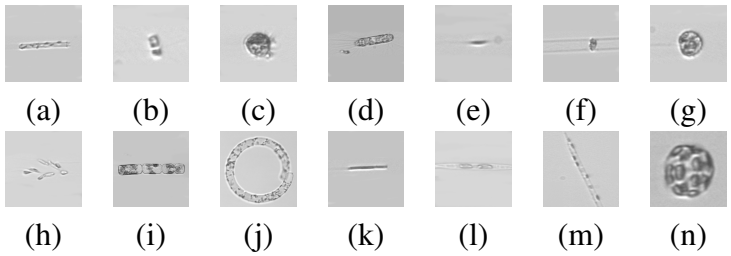
(a) (b) (c) (d) (e) (f) (g)

Figura 2. Classes do conjunto de dados Papsmear. (a) carcinoma. (b) light. (c) moderate. (d) normal columnar. (e) normal intermediate. (f) severe dysplastic. (g) normal superficial



(a) (b) (c) (d)

Figura 3. Classes do conjunto de dados MAMMOSET. (a) malignant mass. (b) benign mass. (c) malignant calcification. (d) benign calcification



(a) (b) (c) (d) (e) (f) (g)

(h) (i) (j) (k) (l) (m) (n)

Figura 4. Classes do conjunto de dados WHOI-Plankton. (a) Cerataulina. (b) Chaetoceros. (c) Ciliate-mix. (d) Corethron. (e) Cyliandrotheca. (f) Dactyliosolen. (g) Dino30. (h) Dinobryon. (i) Guinardia delicatula. (j) Guinardia striata. (k) Leptocylindrus. (l) Pseudonitzschia. (m) Rhizosolenia. (n) Thalassiosira.

4.2. Resultados

Para analisar as metodologias (self-training e co-training), foram realizadas comparações com o paradigma supervisionado, considerando os três conjuntos de dados mencionados na Seção 4.1. A Tabela 1 mostra os resultados obtidos considerando o conjunto de dados Papsmear. As metodologias propostas apresentaram melhores resultados (todas as

métricas), considerando 50% das amostras rotuladas quando comparadas com a supervisionada (100% das amostras rotuladas). Por exemplo, a abordagem co-training obteve um ganho de até 15% de acurácia.

Esse mesmo comportamento foi observado ao analisar os conjuntos de dados Mammoset e WHOI-Plankton (Tabelas 2 e 3), respectivamente. Considerando os resultados obtidos a partir do dataset Mammoset (Tabela 2), a abordagem co-training obteve um ganho de até 19% de acurácia com apenas 40% de amostras rotuladas, quando comparadas com o supervisionado. Em relação à precisão e ao F1-score, os ganhos foram ainda maiores (45% e 38%, respectivamente). A abordagem self-training também apresentou melhores resultados que o paradigma supervisionado. Analisando os resultados do conjunto de dados WHOI-Plankton, o self-training alcançou os melhores resultados ($88,71\%$ de acerto) com 40% de amostras rotuladas, enquanto o supervisionado obteve uma acurácia de $84,86\%$ (com 100% de amostras rotuladas).

Tabela 1. Resultados por abordagem para cada porcentagem do conjunto rotulado (10%, 20%, 30%, 40%, 50% e 100%) considerando as métricas supracitadas na base de dados Papsmear

Rotulado	Abordagem	Acurácia	Precisão	Revocação	F1-Score
10%	Co-Training	37.71%±0.255	42.03%±0.258	39.74%±0.267	38.69%±0.253
	Self-Training	33.43%±0.317	41.51%±0.365	37.29%±0.351	33.11%±0.297
20%	Co-Training	38.86%±0.262	40.91%±0.253	40.91%±0.276	39.09%±0.248
	Self-Training	33.43%±0.305	37.57%±0.319	35.26%±0.321	32.43%±0.282
30%	Co-Training	38.14%±0.280	38.71%±0.262	40.20%±0.294	38.34%±0.270
	Self-Training	35.86%±0.314	38.89%±0.315	37.77%±0.329	34.91%±0.294
40%	Co-Training	40.14%±0.273	41.54%±0.262	42.26%±0.287	40.00%±0.253
	Self-Training	37.14%±0.350	40.49%±0.364	40.17%±0.379	35.83%±0.341
50%	Co-Training	**45.00%±0.276**	**44.83%±0.261**	**47.37%±0.291**	**44.83%±0.273**
	Self-Training	39.86%±0.345	44.43%±0.332	42.00%±0.363	38.46%±0.306
100%	Supervisionado	38.86%±0.290	42.03%±0.311	44.00%±0.322	40.66%±0.303

Tabela 2. Resultados por abordagem para cada porcentagem do conjunto rotulado (10%, 20%, 30%, 40%, 50% e 100%) considerando as métricas supracitadas na base de dados Mammoset

Rotulado	Abordagem	Acurácia	Precisão	Revocação	F1-Score
10%	Co-Training	31.56%±0.115	33.70%±0.113	33.25%±0.121	32.55%±0.104
	Self-Training	30.03%±0.330	28.85%±0.286	31.55%±0.348	24.60%±0.267
20%	Co-Training	34.75%±0.118	36.50%±0.106	36.50%±0.125	35.88%±0.108
	Self-Training	26.52%±0.328	14.45%±0.240	27.95%±0.346	17.80%±0.252
30%	Co-Training	35.83%±0.121	39.10%±0.132	37.95%±0.128	37.60%±0.114
	Self-Training	35.08%±0.305	36.05%±0.281	36.90%±0.321	31.10%±0.239
40%	Co-Training	**37.55%±0.128**	**39.60%±0.113**	**39.40%±0.135**	**38.65%±0.110**
	Self-Training	36.64%±0.352	33.10%±0.167	38.50%±0.371	31.45%±0.216
50%	Co-Training	36.59%±0.166	39.15%±0.137	38.50%±0.175	37.30%±0.134
	Self-Training	37.00%±0.319	38.70%±0.215	38.85%±0.336	32.60%±0.200
100%	Supervisionado	30.49%±0.365	21.80%±0.251	33.75%±0.405	23.95%±0.265

Para melhor explicitar os resultados obtidos, ilustra-se nas Tabelas 4-6 todas as métricas para cada classe dos conjuntos de dados Papsmear, Mammoset e WHOI-Plankton, respectivamente, considerando os melhores casos. Considerando a Tabela 5 (conjunto MAMMOSET), a abordagem co-training com 40% de amostras rotuladas apresenta melhores resultados para as classes 0, 1 e 3. Vale ressaltar que a self-training obteve para a classe 0 uma precisão 4,4 vezes melhor que a supervisionada. Por fim, ao analisar a Tabela 6 (conjunto WHOI-Plankton), as metodologias propostas apresentaram os melhores resultados para quase todas as classes.

Tabela 3. Resultados por abordagem para cada porcentagem do conjunto rotulado (10%, 20%, 30%, 40%, 50% e 100%) considerando as métricas supracitadas na base de dados WHOI-Plankton

Rotulado	Abordagem	Acurácia	Precisão	Revocação	F1-Score
10%	Co-Training	82.43%±0.124	87.70%±0.109	86.73%±0.131	86.49%±0.100
	Self-Training	78.93%±0.172	85.84%±0.137	83.09%±0.181	82.79%±0.131
20%	Co-Training	84.00%±0.115	88.67%±0.086	88.43%±0.122	88.09%±0.091
	Self-Training	85.71%±0.113	90.87%±0.085	90.17%±0.119	89.94%±0.084
30%	Co-Training	85.29%±0.106	90.39%±0.076	89.70%±0.113	89.49%±0.073
	Self-Training	87.79%±0.095	93.29%±0.082	92.36%±0.100	92.21%±0.069
40%	Co-Training	86.29%±0.106	91.31%±0.067	90.76%±0.112	90.50%±0.068
	Self-Training	**88.71%±0.095**	**94.17%±0.073**	**93.36%±0.100**	**93.13%±0.065**
50%	Co-Training	87.00%±0.102	91.91%±0.068	91.53%±0.108	91.29%±0.072
	Self-Training	88.71%±0.110	94.01%±0.065	93.39%±0.116	92.97%±0.076
100%	Supervisionado	84.86%±0.082	94.70%±0.060	94.17%±0.090	94.01%±0.060

Tabela 4. Resultados por abordagem para cada classe do conjunto de dados Papsmear considerando as métricas supracitadas

Classes	Métricas	Supervisionado	Co-Training 50%	Self-Training 50%
0	Acurácia	13.00%±0.178	5.00%±0.035	14.00%±0.204
	Precisão	10.20%±0.143	19.00%±0.122	12.00%±0.098
	Revocação	14.40%±0.198	5.20%±0.039	14.80%±0.217
	F1-Score	12.00%±0.166	8.20%±0.057	12.20%±0.135
1	Acurácia	16.00%±0.219	45.00%±0.145	34.00%±0.245
	Precisão	23.80%±0.223	35.40%±0.048	48.20%±0.158
	Revocação	31.00%±0.298	47.40%±0.154	35.80%±0.258
	F1-Score	26.40%±0.241	40.00%±0.073	36.40%±0.139
2	Acurácia	25.00%±0.145	20.00%±0.117	19.00%±0.129
	Precisão	26.60%±0.080	26.40%±0.178	48.40%±0.307
	Revocação	26.80%±0.174	21.00%±0.124	20.00%±0.136
	F1-Score	23.00%±0.052	23.00%±0.139	23.60%±0.114
3	Acurácia	37.00%±0.256	47.00%±0.044	14.00%±0.171
	Precisão	22.00%±0.020	39.60%±0.062	8.20%±0.096
	Revocação	30.20%±0.268	49.60%±0.050	14.80%±0.181
	F1-Score	21.20%±0.104	44.00%±0.051	10.20%±0.120
4	Acurácia	70.00%±0.252	83.00%±0.067	92.00%±0.044
	Precisão	73.80%±0.178	75.00%±0.067	67.00%±0.178
	Revocação	80.00%±0.294	87.40%±0.072	96.80%±0.048
	F1-Score	76.00%±0.241	80.60%±0.065	77.80%±0.150
5	Acurácia	74.00%±0.147	77.00%±0.057	70.00%±0.183
	Precisão	92.40%±0.107	88.20%±0.068	95.20%±0.074
	Revocação	79.00%±0.146	81.00%±0.057	73.80%±0.191
	F1-Score	84.00%±0.099	84.00%±0.019	81.00%±0.098
6	Acurácia	37.00%±0.130	38.00%±0.103	36.00%±0.386
	Precisão	45.40%±0.108	30.20%±0.053	32.00%±0.247
	Revocação	46.60%±0.201	40.00%±0.110	38.00%±0.407
	F1-Score	42.00%±0.102	34.00%±0.067	28.00%±0.230

5. Conclusões

Neste trabalho, foi proposta a agregação de técnicas semi-supervisionadas com redes neurais convolucionais. Os resultados comprovam vantagens significativas de tal agregação, considerando as técnicas de self-training e co-training. Além disso, mostrou-se que o processo de pseudo-rotulagem aliado à técnica de tranfer learning pode alcançar melhores resultados quando comparado com o paradigma supervisionado em diferentes contextos de imagem. Alcançou-se uma precisão 4,4 vezes melhor ao analisar cada classe dos conjuntos de dados e com apenas metade (ou seja, 50%) das amostras rotuladas. Assim, a pseudo-rotulagem de amostras não rotuladas pode fornecer maior eficácia (ou seja, melhor acurácia, entre outras métricas) e eficiência (menos amostras rotuladas). Em trabalhos futuros, pretende-se propor a junção das duas metodologias propostas.

Tabela 5. Resultados por abordagem para cada classe do conjunto de dados Mammoset considerando as métricas supracitadas

Classes	Métricas	Supervisionado	Co-Training 40%	Self-Training 50%
0	Acurácia	14.20%±0.317	29.40%±0.068	62.60%±0.377
	Precisão	12.80%±0.286	44.80%±0.019	46.40%±0.162
	Revocação	15.80%±0.353	30.80%±0.072	66.00%±0.398
	F1-Score	14.20%±0.318	36.40%±0.050	43.60%±0.249
1	Acurácia	15.74%±0.161	26.38%±0.055	16.60%±0.166
	Precisão	24.20%±0.230	21.60%±0.038	22.00%±0.182
	Revocação	17.20%±0.177	27.60%±0.056	17.20%±0.174
	F1-Score	20.00%±0.198	23.80%±0.037	19.20%±0.176
2	Acurácia	46.60%±0.442	40.60%±0.064	35.60%±0.347
	Precisão	24.80%±0.269	44.00%±0.054	31.60%±0.315
	Revocação	51.60%±0.490	42.60%±0.064	37.40%±0.365
	F1-Score	31.20%±0.296	43.20%±0.050	34.00%±0.335
3	Acurácia	45.40%±0.437	53.80%±0.091	33.20%±0.302
	Precisão	25.40%±0.284	48.00%±0.040	54.80%±0.389
	Revocação	50.40%±0.486	56.60%±0.099	34.80%±0.317
	F1-Score	30.40%±0.288	51.20%±0.025	33.60%±0.307

6. Agradecimentos

O presente trabalho teve o apoio do CNPq, CAPES, Fundação Araucária, SETI, UTFPR.

Referências

Amorim, W., de Rosa, G., Thomazella, R., Castanho, J. E., Dotto, F., Júnior, O., Marana, A., and Papa, J. (2019). Semi-supervised learning with connectivity-driven convolutional neural networks. *Pattern Recognition Letters*.

Chen, J., Feng, J., Sun, X., and Liu, Y. (2020). Co-training semi-supervised deep learning for sentiment classification of mooc forum posts. *Symmetry*, 12(1).

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Elshennawy, N. M. and Ibrahim, D. M. (2020). Deep-pneumonia framework using deep learning models based on chest x-ray images. *Diagnostics*, 10(9).

Oliveira, P., de Carvalho Scabora, L., Cazzolato, M., Bedo, M., Traina, A., and Jr, C. (2017). Mammoset: An enhanced dataset of mammograms. In *Dataset Showcase Workshop - DSW at the Brazilian Symposium on Databases*, pages 1–11.

Orenstein, E., Beijbom, O., Peacock, E., and Sosik, H. (2015). Whoi-plankton- a large scale fine grained visual recognition benchmark dataset for plankton classification. *Tech Report*, pages 1–2.

Rosenberg, C., Hebert, M., and Schneiderman, H. (2005). Semi-supervised self-training of object detection models. In *2005 Seventh IEEE Workshops on Applications of Computer Vision - Volume 1*, volume 1, pages 29–36.

T.R., S., Thomas, S. A., Kamath, V., and B., N. K. (2019). Hybrid xception model for human protein atlas image classification. In *IEEE 16th India Council International Conference*, pages 1–4.

van Engelen, J. E. and Hoos, H. H. (2019). A survey on semi-supervised learning. *Machine Learning*, 109:373–440.

Tabela 6. Resultados por abordagem para cada classe do conjunto de dados WHOI-Plankton considerando as métricas supracitadas

Classes	Métricas	Supervisionado	Co-Training 50%	Self-Training 40%
0	Acurácia	61.00%±0.108	60.00%±0.061	64.00%±0.102
	Precisão	100.00%±0.000	90.40%±0.059	100.00%±0.000
	Revocação	67.80%±0.117	63.20%±0.065	67.40%±0.107
	F1-Score	80.20%±0.089	74.00%±0.035	79.80%±0.076
1	Acurácia	89.00%±0.022	78.00%±0.075	93.00%±0.044
	Precisão	100.00%±0.000	90.20%±0.092	96.20%±0.060
	Revocação	98.80%±0.027	81.80%±0.080	97.80%±0.049
	F1-Score	99.40%±0.013	85.60%±0.072	96.80%±0.032
2	Acurácia	84.00%±0.022	89.00%±0.022	89.00%±0.022
	Precisão	96.40%±0.033	94.00%±0.064	93.20%±0.053
	Revocação	93.00%±0.022	93.80%±0.027	93.80%±0.026
	F1-Score	94.60%±0.025	93.60%±0.038	93.20%±0.021
3	Acurácia	84.00%±0.022	93.00%±0.027	81.00%±0.082
	Precisão	95.40%±0.047	94.00%±0.042	93.40%±0.049
	Revocação	93.00%±0.022	98.00%±0.027	85.20%±0.084
	F1-Score	94.20%±0.018	95.80%±0.011	88.80%±0.048
4	Acurácia	90.00%±0.000	95.00%±0.000	95.00%±0.000
	Precisão	100.00%±0.000	94.00%±0.042	100.00%±0.000
	Revocação	100.00%±0.000	100.00%±0.000	100.00%±0.000
	F1-Score	100.00%±0.000	96.80%±0.020	100.00%±0.000
5	Acurácia	90.00%±0.000	95.00%±0.000	95.00%±0.000
	Precisão	98.00%±0.027	97.00%±0.045	100.00%±0.000
	Revocação	100.00%±0.000	100.00%±0.000	100.00%±0.000
	F1-Score	98.80%±0.016	98.40%±0.023	100.00%±0.000
6	Acurácia	90.00%±0.000	93.00%±0.027	95.00%±0.000
	Precisão	97.20%±0.063	94.20%±0.062	98.00%±0.027
	Revocação	100.00%±0.000	98.00%±0.027	100.00%±0.000
	F1-Score	98.40%±0.036	96.00%±0.042	98.80%±0.016
7	Acurácia	88.00%±0.027	88.00%±0.057	92.00%±0.044
	Precisão	82.40%±0.051	85.60%±0.047	77.60%±0.058
	Revocação	97.60%±0.033	92.60%±0.062	96.80%±0.048
	F1-Score	89.40%±0.030	88.80%±0.034	86.00%±0.041
8	Acurácia	90.00%±0.000	95.00%±0.000	95.00%±0.000
	Precisão	94.00%±0.042	98.00%±0.027	89.80%±0.065
	Revocação	100.00%±0.000	100.00%±0.000	100.00%±0.000
	F1-Score	96.80%±0.020	98.80%±0.016	94.40%±0.037
9	Acurácia	85.00%±0.035	84.00%±0.074	91.00±0.054
	Precisão	89.00%±0.070	88.60%±0.097	96.00%±0.041
	Revocação	94.20%±0.039	88.20%±0.078	95.60%±0.060
	F1-Score	91.40%±0.038	88.00%±0.044	95.40%±0.015
10	Acurácia	90.00%±0.000	95.00%±0.000	91.00%±0.065
	Precisão	98.00%±0.027	89.80%±0.064	87.20%±0.096
	Revocação	100.00%±0.000	100.00%±0.000	95.80%±0.069
	F1-Score	98.80%±0.016	94.60%±0.036	90.40%±0.040
11	Acurácia	83.00%±0.027	86.00%±0.022	84.00%±0.108
	Precisão	93.20%±0.045	92.20%±0.029	97.80%±0.030
	Revocação	92.00%±0.027	90.20%±0.027	88.40%±0.116
	F1-Score	92.60%±0.013	91.40%±0.025	92.40%±0.066
12	Acurácia	85.00%±0.035	85.00%±0.050	89.00%±0.065
	Precisão	89.40%±0.032	85.00%±0.097	94.60%±0.052
	Revocação	94.20%±0.039	89.40%±0.055	93.60%±0.070
	F1-Score	91.80%±0.022	86.80%±0.050	94.00%±0.053
13	Acurácia	79.00%±0.054	82.00%±0.097	88.00%±0.027
	Precisão	92.80%±0.027	93.80%±0.065	94.60%±0.005
	Revocação	87.80%±0.059	86.20%±0.101	92.60%±0.032
	F1-Score	89.80%±0.045	89.40%±0.047	93.80%±0.016

Zhang, L., Lu, L., Nogues, I., Summers, R. M., Liu, S., and Yao, J. (2017). Deeppap: Deep convolutional networks for cervical cell classification. *IEEE Journal of Biomedical and Health Informatics*, 21:1633–1643.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. (2019). A comprehensive survey on transfer learning. *CoRR*, abs/1911.02685.