Proposta submetida ao edital N^o 04/2022 ProPes – Iniciação Científica UFABC

Projeto de Pesquisa

Estudo de Interpretabilidade de Redes Neurais Profundas para Classificação de Imagens Médicas

Resumo

Nos últimos anos, estudos de ferramentas de inteligência artificial para auxílio ao diagnóstico médico têm reportado valores de acurácia comparáveis aos de profissionais médicos. No entanto, vários outros desafios existem para a adoção de tais ferramentas na prática clínica. Muitos dos avanços recentes devem-se a modelos *opacos* de redes neurais, em que a mecânica de decisão não pode ser observada. Isso contribui para diminuir a confiança sobre os diagnósticos apontados. No presente trabalho, nós propomos uma investigação de modelos de redes neurais *interpretáveis* para auxílio ao diagnóstico, isto é, modelos em que um ser humano tenha a possibilidade de compreender o processo de decisão do algoritmo. O uso de métodos interpretáveis tem como potenciais vantagens o aumento da confiança por parte de profissionais médicos, uma maior transparência do processo de diagnóstico e a possibilidade de investigar quais são as características das imagens que influenciam a saída do algoritmo de auxílio ao diagnóstico.

Santo André, junho 2022

1 Introdução

Na última década, modelos de redes neurais artificiais (RNAs) se estabeleceram como o estado da arte em aplicações como classificação [KSH12, ALE+16] e segmentação de imagens [HDW+17]. Essa nova geração de redes neurais artificias é caracterizada pela profundidade (existência de múltiplas camadas), número grande de parâmetros ajustáveis (passando de milhões em alguns casos) e uso de bases de treinamento com número grande de exemplos (centenas de milhares ou mesmo milhões). Em estudos realizados com número pequeno de exemplos de treinamento, os sistemas gerados podem sofrer de baixa capacidade de *generalização*. Isto é, quando transportados para outros cenários, o sistemas de classificação não atingem desempenho tão bom quanto no estudo original. Modelos de redes neurais para classificação de imagens são considerados *opacos*: via de regra, não é possível inspecionar a mecânica de decisão do modelo ou compreender quais atributos da imagem de entrada resultaram na saída observada. Essas características: necessidade de grandes quantidades de dados, possibilidade de baixa generalização e falta de transparência, são desafios para a adoção de modelos de redes neurais em aplicações de imagens médicas, em geral marcadas por escassez de dados, necessidade de alta confiabilidade e transparência.

Quanto à transparência de modelos, Rudin [Rud19] estabelece uma distinção entre modelos explicados e modelos interpretáveis. No primeiro caso, uma *explicação* é produzida para a saída do modelo. Isso traz potenciais disvantagens, entre elas a possibilidade de que a explicação não seja fiel ao mecanismo que o modelo usou para chegar a sua resposta. Por sua vez, os modelos interpretáveis são aqueles em que um ser humano seria capaz de entender (em algum grau) o mecanismo de decisão. Rudin [Rud19] argumenta que não é possível definir uma noção de interpretabilidade aplicável a qualquer domínio, sendo necessário criar definições específicas para cada aplicação (por exemplo, a partir de justificativas que os especialistas de domínio fornecem para suas próprias decisões). A preferência por modelos explicados teria causa em uma ideia mal justificada [Rud19], de que a interpretabilidade seria obtida ao custo de uma menor acurácia. Nos últimos anos, alguns métodos foram propostos para criar explicações *a posteriori* para a saída de redes neurais profundas, porém estes podem ser pouco robustos, mesmo perturbações imperceptíveis das imagens podem gerar grandes variações de saída [GAZ18].

O presente texto propõe uma investigação sobre a interpretabilidade de redes neurais profundas para auxílio ao diagnóstico com imagens médicas. A investigação seria composta por duas tarefas principais:

- Uma pesquisa bibliográfica sobre modelos de redes neurais interpretáveis e sua aplicação em problemas de diagnóstico médico.
- A construção e avaliação de um modelo de rede neural em um problema de classificação de imagens médicas. Além do desempenho de classficação, a avaliação levará em conta aspectos como robustez e interpretabilidade do modelo. Nós pretendemos usar a metodologia das "fichas de modelo" (*model cards*)[MWZ+19] para documentar as características principais do modelo, como conjunto de treinamento, desempenho, possíveis problemas com o conjunto de dados de treinamento, limitações e características de interpretabilidade.

O restante do texto é organizado da seguinte maneira: na Seção 2, alguns conceitos fundamentais são apresentados e discute-se a relação entre os métodos interpretáveis e a adoção de ferramentas de auxílio ao diagnóstico. A Seção 3 detalha os objetivos da proposta atual. A Seção 4 descreve as etapas a serem cumpridas e apresenta o cronograma proposto para a execução do projeto.

2 Conceitos Fundamentais

O uso de inteligência artifical em aplicações de imagens médicas pode impactar significativamente os processos de diagnóstico. Modelos mais recentes de redes neurais têm reportado acurácia compatível com a de radiologistas [WPP+19], embora em certos casos o melhor resultado ainda seja obtido com uma combinação de diagnósticos, de um médico e do algoritmo[WPP+19]. O uso de ferramentas computacionais pode reduzir a necessidade da realização de tarefas repetitivas pelo profissional de diagnóstico (*p.ex.* observar inúmeros cortes de imagem em um único exame). Assim, um mesmo profissional poderia atender um número maior de pacientes, contornando limitações no número de médicos disponíveis e propiciando um atendimento especializado para mais pessoas.

Em uma pesquisa recente [SRM+21], médicos avaliam que avanços em inteligência artificial podem reduzir o tempo gasto em tarefas repetitivas e melhorar a acurácia do diagnóstico por imagens, embora temam uma desvalorização da *expertise* médica e a redução da participação de profissionais no processo de diagnóstico [SRM+21]. Além disso, médicos que realizam diagnóstico por imagens manifestam preocupação com a possibilidade de responsabilização devido a erros cometidos pelos algoritmos [SRM+21]. Para que as novas ferramentas de diagnóstico auxiliadas por inteligência artifical sejam adotadas, é necessário

que os profissionais possam entender o processo de decisão, aumentando sua confiança no processo.

Um traço comum aos modelos mais complexos, como redes neurais profundas, é a *subespecificação* [DHM+20], caracterizada pela existência de vários modelos diferentes (*i.e.* diferentes conjuntos de parâmetros aprendidos) que alcançam desempenho semelhante em treinamento. Ainda assim, esses modelos podem ter desempenho muito diferente quando levados para campo, resultando em baixa confiabilidade [DHM+20]. É fato conhecido que modelos de imagens médicas podem obter desempenho muito diferente quando implantados em ambientes distintos daquele onde foram obtidos os dados de treinamento [BZOR+18, MSG+20, DHM+20], estudos mais recentes tentam aplicar avaliações multi-instituição para diminuir este problema [MSG+20].

Outro fator de preocupação é que bases públicas de imagens médicas podem apresentar problemas relacionados à baixa qualidade das anotações [OR20] e/ou pela presença de subpopulações desconhecidas (*estratificação oculta*) [ORDCR20] que produzem distorções de treinamento e afetam a capacidade de generalização. Isso reforça a necessidade de modelos transparentes, em que a mecânica de decisão seja observável. De acordo com Megnan *et al* [DLH19], ainda não existe um procedimento bem estabelecido para avaliar a interpretabilidade de um modelo, devido à existência de diversas métricas que enfatizam diferentes aspectos da interpretabilidade.

Algumas tentativas de criar modelos com explicações ou interpretáveis para auxílio a diagnóstico por imagens médicas incluem:

- Criar modelos de similaridade entre imagens, mostradas ao profissional responsável pelo diagnóstico, em lugar de gerar uma única previsão [TH20].
- Gerar mapas de saliência, que mostrariam as regiões da imagem mais relevantes para o diagnóstico [SWP+20].
- Gerar automaticamente um texto (parecer) que acompanha o diagnóstico e pode ser interpretado por um profissional [GORC+19].

Alguns dos métodos propostos são baseados em gerar uma explicação para a saída de uma rede neural profunda, por exemplo, construindo um mapa de regiões mais relevantes para a classificação. Porém, esses métodos ainda precisam ser melhor validados, e ferramentas semelhantes têm se revelado pouco robustas, criando mapas muito diferentes para imagens

perceptualmente semelhantes [GAZ18, DAA+19]. Em alguns casos, os próprios médicos relatam pouca disposição para usar ferramentas de visualização [GORC+19].

3 Objetivos e metas

A presente proposta tem dois objetivos principais:

- Realizar um levantamento bibliográfico sobre modelos interpretáveis de redes neurais em aplicações de diagnóstico médico por imagens.
- Construir e avaliar um modelo de rede neural profunda para classificação em uma tarefa de auxílio ao diagnóstico, utilizando uma base de imagens públicas. O modelo será avaliado por métricas de desempenho de classificação (por exemplo, acurácia, precisão, *recall*) e também quanto às suas características de interpretabilidade.

Com esses objetivos, nós pretendemos alcançar os seguintes resultados: em primeiro lugar, contribuir para a formação da bolsista em iniciação científica em uma área aberta de pesquisa e que tem alto potencial de impacto, habilitando-a a prosseguir em estudos mais avançados posteriormente. Em segundo lugar, estabelecer fundamentos para outros estudos futuros. Com o levantamento bibliográfico, esperamos mapear oportunidades de pesquisa na área de modelos neurais interpretáveis. O estudo de um modelo interpretável ajudará a criar *expertise* em nosso grupo de pesquisa, fornecendo um modelo que sirva de base de comparação para estudos futuros. Os avanços que surgirem na área de modelos interpretáveis em aplicações médicas podem contribuir para acelerar a adoção de ferramentas de auxílio ao diagnóstico, aumentar a confiança nessas ferramentas e aumentar a transparência de processos de diagnóstico médico.

4 Materiais e Métodos

O projeto empregará bases públicas de imagens médicas. Inicialmente, vamos trabalhar com o conjunto de dados *Lung and Colon Cancer Histopathological Image Dataset* (LC25000) [BBT+19], que contém imagens de lâminas de microscópio de tecido de pulmão e do intestino. A tarefa relacionada ao conjunto de dados é o diagnóstico de tumores. A base de dados é formada por 25.000 imagens coloridas

Bases públicas de imagens médicas podem sofrer como problemas como baixa qualidade de rótulos [OR20] e *estratificação oculta* [ORDCR20], em que um subconjunto de exemplos pode apresentar características muito distintas do resto da amostra, distorcendo as medidas de desempenho de modelos. Nós pretendemos estudar as técnicas propostas na literatura [OR20, ORDCR20] para diagnóstico visual e auditoria de modelos, avaliando a viabilidade da aplicação ao nosso projeto.

Para o treinamento dos modelos, serão usados recursos computacionais dos laboratórios da UFABC. A implementação dos classificadores será baseada em bibliotecas de software livre, como *Numpy* [Oli15] e *Tensorflow* [AAB+16].

No início do projeto, uma pesquisa bibliográfica preliminar será conduzida, para construir um panorama do estado da arte em modelos interpretáveis na área de auxílio ao diagnóstico médico por imagens. Serão incluídos nessa revisão artigos mais gerais sobre interpretabilidade em aprendizado de máquina [Lip17, DVK17, GBY+19, MCB20, RCC+21], métodos específicos que resultam em modelos interpretáveis [CLT+19, HCLR19, LZH+20], artigos que analisam questões como confiança em resultados de modelos de aprendizado de máquina [RU18, PSGH+21] e artigos que abordam a interpretabilidade de redes neurais em diagnóstico médico [SCM21, GLMA21, GPdSV+21, PNT22]. Como produto desse estudo, será gerado um texto de revisão, que servirá de fundamento para as etapas seguintes do projeto e poderá orientar também futuras propostas de estudo.

Inicialmente, nós pretendemos investigar alguns modelos de redes neurais profundas que se utilizam de *protótipos*, regiões da imagem que concentram características visuais distintivas da classe correspondente [CLT+19, HCLR19].

Em paralelo à construção do modelo, será elaborada uma "ficha de modelo" (*model card*) [MWZ+19]. Uma ficha de modelo é um documento curto, descrevendo características do modelo treinado, com o propósito de aumentar sua transparência. Alguns aspectos relevantes para modelos voltados para imagens médicas e que devem ser documentados na ficha de modelo são: possível viés dos dados de treinamento, procedimentos de avaliação, desempenho de treinamento e em um conjunto de validação, medidas quantitavivas e/ou qualitativas da interpretabilidade do modelo,

A Figura 1 apresenta o cronograma proposto para execução do projeto. O trabalho de revisão bibliográfica foi dividido em duas fases para melhor gerenciamento. Na fase inicial de revisão, durando 3 meses, os objetivos serão estudar os fundamentos de modelos interpretáveis, como avaliá-los e elencar alguns modelos de redes neurais que sejam viáveis

de implementar nas aplicações de auxílio a diagnóstico. Também serão estudados princípios de processamento de imagens médicas e extração de atributos. Durante os primeiros 3 meses também pretendemos definir o conjunto de bibliotecas de software, o modelo interpretável que será implementado e quais bases de imagens serão empregadas. Isso envolve estudar a base LC25000 e definir se existem outras bases disponíveis que possam agregar ao estudo do problema de classificação de tumores. A segunda fase de revisão será mais abrangente e visará construir um panorama da área de aprendizado de modelos interpretáveis e suas aplicações em processamento de imagens médicas. A partir do quarto mês, os esforços se concentrarão no desenvolvimento do modelo interpretável para o problema de classificação de tumores. Para efeito de avaliação, também serão implantados alguns modelos alternativos (p.ex. florestas aleatórias, redes neurais *opacas*) para comparação. Nos últimos quatro meses de projeto, prevê-se o refinamento do modelo interpretável e procedimentos de avaliação.

Etapa	Mês											
	1	2	3	4	5	6	7	8	9	10	11	12
Revisão Bibliográfica Inicial	■	■	■									
Estudo da metodologia de ficha de modelos	■	■	■									
Avaliação de Bibliotecas de software		■	■									
Definição de modelos a ser implementados		■	■	■								
Estudo bases de imagens públicas		■	■									
Revisão Bibliográfica, segunda fase				■	■	■	■	■	■	■		
Escrita de relatório parcial					■	■						
Implementação e teste modelo interpretável				■	■	■	■	■				
Implementação de modelos altenativos					■	■	■	■	■			
Elaborar ficha de modelo									■	■	■	
Refinar modelo e avaliação									■	■	■	
Escrita do relatório final									■	■	■	■

Figura 1: Cronograma de execução da proposta

Referências

[AAB+16] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[ALE+16] Michael David Abràmoff, Yiyue Lou, Ali Erginay, Warren Clarida, Ryan Amelon, James C. Folk, and Meindert Niemeijer. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Investigative Ophthalmology & Visual Science*, 57(13):5200–5206, 10 2016.

[BBT+19] Andrew A. Borkowski, Marilyn M. Bui, L. Brannon Thomas, Catherine P.

Wilson, Lauren A. DeLand, and Stephen M. Mastorides. Lung and colon cancer histopathological image dataset (lc25000), 2019.

[BZOR+18] Marcus A. Badgeley, John R. Zech, Luke Oakden-Rayner, Benjamin S. Glicksberg, Manway Liu, William Gale, Michael V. McConnell, Beth Percha, Thomas M. Snyder, and Joel T. Dudley. Deep Learning Predicts Hip Fracture using Confounding Patient and Healthcare Variables. *arXiv:1811.03695 [cs]*, November 2018. arXiv: 1811.03695.

[CLT+19] Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. This looks like that: Deep learning for interpretable image recognition, 2019.

[DAA+19] Ann-Kathrin Dombrowski, Maximilian Alber, Christopher J. Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame, 2019.

[DHM+20] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification Presents Challenges for Credibility in Modern Machine Learning. *arXiv:2011.03395 [cs, stat]*, November 2020. arXiv: 2011.03395.

[DLH19] Mengnan Du, Ninghao Liu, and Xia Hu. Techniques for interpretable machine learning. *Commun. ACM*, 63(1):68–77, dec 2019.

[DVK17] Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608 [cs, stat]*, March 2017. arXiv: 1702.08608.

[GAZ18] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile, 2018.

[GBY+19] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining Explanations: An Overview of Interpretability of Machine Learning. *arXiv:1806.00069 [cs, stat]*, February 2019. arXiv: 1806.00069.

[GLMA21] Mara Graziani, Thomas Lompech, Henning Müller, and Vincent Andrearczyk. Evaluation and comparison of cnn visual explanations for histopathology. In *XAI-AAAI-21*, 2021.

[GORC+19] William Gale, Luke Oakden-Rayner, Gustavo Carneiro, Lyle J. Palmer, and Andrew P. Bradley. Producing radiologist-quality reports for interpretable deep learning. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1275–1279, 2019.

[GPdSV+21] Mara Graziani, Iam Palatnik de Sousa, Marley M. B. R. Vellasco, Eduardo Costa da Silva, Henning Müller, and Vincent Andrearczyk. Sharpening local interpretable model-agnostic explanations for histopathology: Improved understandability and reliability. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 540–549, Cham, 2021. Springer International Publishing.

[HCLR19] Peter Hase, Chaofan Chen, Oscar Li, and Cynthia Rudin. Interpretable image recognition with hierarchical prototypes, 2019.

[HDW+17] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron C. Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, 35:18–31, 2017.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[Lip17] Zachary C. Lipton. The Mythos of Model Interpretability. *arXiv:1606.03490 [cs, stat]*, March 2017. arXiv: 1606.03490.

[LZH+20] Jimmy Lin, Chudi Zhong, Diane Hu, Cynthia Rudin, and Margo Seltzer. Generalized and Scalable Optimal Sparse Decision Trees. *arXiv:2006.08690 [cs, stat]*, August 2020. arXiv: 2006.08690.

[MCB20] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges. *arXiv:2010.09337 [cs, stat]*, October 2020. arXiv: 2010.09337.

[MSG+20] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S. Corrado, Ara Darzi, Mozziyar Etemadi, Florencia Garcia-Vicente, Fiona J. Gilbert, Mark Halling-Brown, Demis Hassabis, Sunny Jansen, Alan Karthikesalingam, Christopher J. Kelly, Dominic King, Joseph R. Ledsam, David Melnick, Hormuz Mostofi, Lily Peng, Joshua Jay Reicher, Bernardino Romera-Paredes, Richard Sidebottom, Mustafa Suleyman, Daniel Tse, Kenneth C. Young, Jeffrey De Fauw, and Shravya Shetty. International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788):89–94, January 2020. Number: 7788 Publisher: Nature Publishing Group.

[MWZ+19] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA, 2019. Association for Computing Machinery.

[Oli15] Travis E. Oliphant. *Guide to NumPy*. CreateSpace Independent Publishing Platform, USA, 2nd edition, 2015.

[OR20] Luke Oakden-Rayner. Exploring Large-scale Public Medical Image Datasets. *Academic Radiology*, 27(1):106–112, January 2020. Publisher: Elsevier.

[ORDCR20] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. *Proceedings of the ACM Conference on Health, Inference, and Learning*, 2020:151–159, April 2020.

[PNT22] Cristiano Patrício, João C. Neves, and Luís F. Teixeira. Explainable deep learning methods in medical imaging diagnosis: A survey, 2022.

[PSGH+21] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. Manipulating and Measuring Model Interpretability. *arXiv:1802.07810 [cs]*, January 2021. arXiv: 1802.07810.

[RCC+21] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges. *arXiv:2103.11251 [cs, stat]*, March 2021. arXiv: 2103.11251.

[RU18] Cynthia Rudin and Berk Ustun. Optimized Scoring Systems: Toward Trust in Machine Learning for Healthcare and Criminal Justice. *INFORMS Journal on Applied Analytics*, 48(5):449–466, October 2018. Publisher: INFORMS.

[Rud19] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019. Number: 5 Publisher: Nature Publishing Group.

[SCM21] Chetan L. Srinidhi, Ozan Ciga, and Anne L. Martel. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, 67:101813, 2021.

[SRM+21] Jane Scheetz, Philip Rothschild, Myra McGuinness, Xavier Hadoux, H. Peter Soyer, Monika Janda, James J. J. Condon, Luke Oakden-Rayner, Lyle J. Palmer, Stuart Keel, and Peter van Wijngaarden. A survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology and radiation oncology. *Scientific Reports*, 11(1):5193, March 2021. Number: 1 Publisher: Nature Publishing Group.

[SWP+20] Yiqiu Shen, Nan Wu, Jason Phang, Jungkyu Park, Kangning Liu, Sudarshini Tyagi, Laura Heacock, S. Gene Kim, Linda Moy, Kyunghyun Cho, and Krzysztof J. Geras. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *arXiv:2002.07613 [cs, eess, stat]*, February 2020. arXiv: 2002.07613.

[TH20] Johnson Thomas and Tracy Haertling. AIBx, Artificial Intelligence Model to Risk Stratify Thyroid Nodules. *Thyroid: Official Journal of the American Thyroid Association*, 30(6):878–884, June 2020.

[WPP+19] N. Wu, J. Phang, J. Park, Y. Shen, Z. Huang, M. Zorin, S. Jastrzebski, T. Févry, J. Katsnelson, E. Kim, S. Wolfson, U. Parikh, S. Gaddam, L. L. Y. Lin, K. Ho, J. D. Weinstein, B. Reig, Y. Gao, H. T. K. Pysarenko, A. Lewin, J. Lee, K. Airola, E. Mema, S. Chung, E. Hwang, N. Samreen, S. G. Kim, L. Heacock, L. Moy, K. Cho, and K. J. Geras. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Transactions on Medical Imaging*, pages 1–1, 2019.