深度学习期末报告

基于No-sign方法的COVID-X光片对抗攻击

2019级人工智能班

魏楚扬 320190939281

一、摘要

近年来，COVID-19流行病以其极强的传染性在全球范围内蔓延。COVID-19的诊断在预防、控制和治疗程序中至关重要。基于深度学习的图像分类模型已被证明对使用胸部X光图像的肺炎分类有效，帮助医生更有效地诊断和治疗该疾病。然而，研究人员通过注入人类无法察觉的微小扰动，证实了深度神经网络的脆弱性。这些对抗性样本成为医疗安全系统的一个主要威胁，特别是在疾病检测领域。在这方面，我们做了一些实验来攻击胸部X射线图像，以研究两类具有符号和无符号算子的攻击方法的效率。我们选择了快速梯度符号法（FGSM）、基本迭代法（BIM）和投影梯度下降法（PGD），并将其转化为无符号攻击方法，分析了它们在白盒和黑盒测试中的有效性。我们确认替代的无符号攻击方法更有效率。在扰动强度边界内，白盒攻击比黑盒攻击更有效。PGDNM可以将VGG16的分类精度拉低到6.89%。

二、简介

与核酸检测相比，对冠状胸片和CT图像的筛查对COVID的鉴别和诊断更为有效[3]。它为医生和放射科医生提供了视觉证据，有助于证明肺炎检测和康复的准确性。然而，COVID肺炎、普通肺炎和正常人之间的胸部X光图像的视觉差异太过细微，难以分辨。为了克服这一瓶颈，对计算机辅助诊断系统存在着巨大的需求，其中机器学习技术发挥了作用。深度神经网络（DNNs）可以分析数以千计的图像，并以较高的准确性识别异常，推翻或支持诊断，从而帮助放射科医生有把握地解释放射学图像，并节省时间。

值得注意的是，医疗结果的安全性和可靠性对医生和患者都非常重要。然而，最近的研究发现，最先进的DNN很容易受到对抗性攻击，如FGSM、I-GSM、MI-GSM、C&W、PGD、UAP等。[4]-[10]. 这些攻击可以通过在人眼无法察觉的物体上添加噪音，轻松骗取分类模型的错误结果。特别是在医疗领域，DNN的脆弱性会大大降低社会信任度，从而降低诊断的效率，影响这一技术的创新和应用。

以前对医学图像攻击的研究[11]大多集中在基于梯度的算法上，如FGSM、I-FGSM和PGD与基于符号的方法(SM)。然而，也有一些解释和应用致力于无符号梯度方法(NM)[12]，这已被证明是更有效的。基于胸部X光数据集上的COVID分类任务，我们比较了这两类算法的有效性，即符号和无符号攻击。并凭借这个三重分类任务，进一步分析了对抗性攻击方法的性质。在这项研究中，我们选择VGG16、InceptionV3和DenseNet201作为我们的基线分类模型。我们采用了包括FGSM、FGNM、PGD在内的SM攻击，并在这些模型上做了相应的NM变体，以考察其鲁棒性。对攻击方法的研究有助于我们对DNNs的全面认识，这为我们在未来的工作中设计和优化针对攻击的防御算法提供了基础，从而提高COVID检测的鲁棒性。经过SM和NM的比较，应直接对梯度防御给予更多关注，在提高网络稳定性的同时，必须保护原始图像的梯度信息。更重要的是，在疫情发生期间，越来越多的胸腔X光数据采集支持了我们研究的可靠性和典型性。

我们的贡献总结如下：

- 我们使用无符号算子重写了三种常用的符号攻击方法，并从理论角度分析了它们的特性。
- 我们在最新的COVID胸部X射线数据集上使用三个预训练的DNN，然后再加上几个适应层，成功实现了三重分类任务。在基线模型的支持下，通过对扰动强度和测试性能的剖析，对SM和NM进行了比较和一些实验。
- 我们进一步分析了攻击的有效性以及通过白盒和黑盒攻击对特定医学图像的DNNs的鲁棒性。

三、算法与模型准备

1. 数据解释

策划的数据集是两个公开的数据集的组合，COVIDx-CRX2[19]和Chest X-Ray[23]，总共包含2358张COVID-19、15574张正常和4273张肺炎胸片图像。所有的图像都被归一化为0到1的范围，并将其尺寸重塑为224×224×3的尺寸。由于实验的目的是研究攻击方法的效率，所以随机选择了一个平衡的数据集，以充分利用三重分类任务的计算资源。从每个类别中分别提取了1500张图片。

2. 对抗攻击方法

在实验中，FGSM、BIM和PGD作为基本的基于符号的攻击方法(SM)被实现，并根据Y. Cheng等人[12]提出的算子改写成无符号攻击方法(NM)。 我们将这些无符号攻击方法分别称为FGNM、BINM和PGDNM。

2.1 sign 符号攻击方法

公式（1）给出了FGSM的公式。其中，x^{adv}为攻击结果，x为输入图像，y为x的标签，θ是模型的权重，$J(\theta, x, y)$是模型的损失函数，∇_x是x的梯度符号算子，ϵ为扰动强度。

$$x^{adv} = x + \epsilon \cdot \operatorname{sign}\left(\nabla_x J(\theta, x, y)\right) \tag{1}$$

BIM的攻击方法通过迭代FGSM生成，如公式（3）所示。其中，攻击在样本x上更新了T次，每个小步长$\alpha = \epsilon/T$。夹子运算符表示输出的裁剪，将其约束在半径为 ϵ 的球中。内部公式由FGSM演变而来，因此显示出类似的形式。

$$\begin{aligned} x_0^{adv} &= x \\ \operatorname{Clip}_{x,\epsilon}\left(x^{adv}\right) &= \min\left(\max\left(x^{adv}, x-\epsilon\right), x+\epsilon\right) \end{aligned} \tag{2}$$

$$x_{t+1}^{adv} = \operatorname{Clip}_{x,\epsilon}\left\{x_t^{adv} + \alpha \cdot \operatorname{sign}\left(\nabla_x J\left(\theta, x_t^{adv}, y\right)\right)\right\} \tag{3}$$

PGD定义为公式（4）。其中，$\operatorname{Proj}_{x+s}$代表产生扰动后的投影步骤，$S$代表样本周围的$L_\infty$或$L_2$邻域。

$$x_{t+1}^{adv} = \operatorname{Proj}_{x+s}\left\{x_t^{adv} + \alpha \cdot \operatorname{sign}\left(\nabla_x J\left(\theta, x_t^{adv}, y\right)\right)\right\} \tag{4}$$

2.2 No-sign 无符号攻击方法

无符号算子的定义如公式（5）所示。噪声δ_t通过对g_t的线性放缩，不仅保留了梯度的方向，同时也将幅值统一成$\|\operatorname{sign}(g_t)\|$。为了重写和简化公式，Y. Cheng简化了公式，令$g_t = \nabla_x J\left(\theta, x_t^{adv}, y\right)$，并将$\operatorname{sign}(g_t)$替换为了$\frac{g_t}{\|g_t\|}\|\operatorname{sign}(g_t)\|$。

$$\zeta = \frac{\|\operatorname{sign}(g_t)\|}{\|g_t\|}, \quad \delta_t = x_{t+1}^{adv} - x_t^{adv} = \alpha \cdot \zeta \cdot g_t \tag{5}$$

因此无符号攻击表达式FGNM、BINM和PGDNM可以定义如下：

$$x^{adv} = x + \epsilon \cdot \frac{g_t}{\|g_t\|} \cdot \|\operatorname{sign}(g_t)\| \tag{6}$$

$$x_{t+1}^{adv} = \operatorname{Clip}_{x,\epsilon}\left\{x_t^{adv} + \alpha \cdot \frac{g_t}{\|g_t\|} \cdot \|\operatorname{sign}(g_t)\|\right\} \tag{7}$$

$$x_{t+1}^{adv} = \operatorname{Proj}_{x+S}\left\{x_t^{adv} + \alpha \cdot \frac{g_t}{\|g_t\|} \cdot \|\operatorname{sign}(g_t)\|\right\} \tag{8}$$

2.3 理论分析

为了以最快的方式接近最佳状态，攻击的幅度和其步长的方向都是关键的组成部分。新方法的优点是不仅保留了符号法的大小，而且使攻击方向与梯度方向保持一致。因此，从理论分析的角度来看，这种NM算法的言论更加合理，值得更好的表现。在实验部分，将实现SM和NM对COVID数据集的攻击，并对比了对抗结果的效率。

3. 分类模型

在实验中，我们选择了三个不同的预训练的深度学习模型作为基线模型。VGG16、InceptionV3和DenseNet201通常用于医学图像分类任务[11]，我们选择它们来研究对对抗性攻击准确性的影响。表Ⅰ显示了三种DNN架构的总可训练参数和通过ImageNet数据训练的前1、前5的准确率。更重要的是，这三个预训练的模型被加强和微调以适应肺部X射线三分类任务。以预训练模型后接了一个适应模块，由一个平均池化层、一个flatten 层、一个批量正则化层和两个全连接层组成。

表Ⅰ：可训练参数的数量和前1名前2名的准确率（在ImageNet验证数据集上的预训练模型性能）

模型选择	参数个数	Top-1 Accuracy	Top-5 Accuracy
VGG16	14,747,715	71.3%	90.1%
InceptionV3	21,934,307	77.9%	93.7%
DenseNet201	18,452,803	77.3%	93.6%

四、实验与结果分析

1. 实验流程

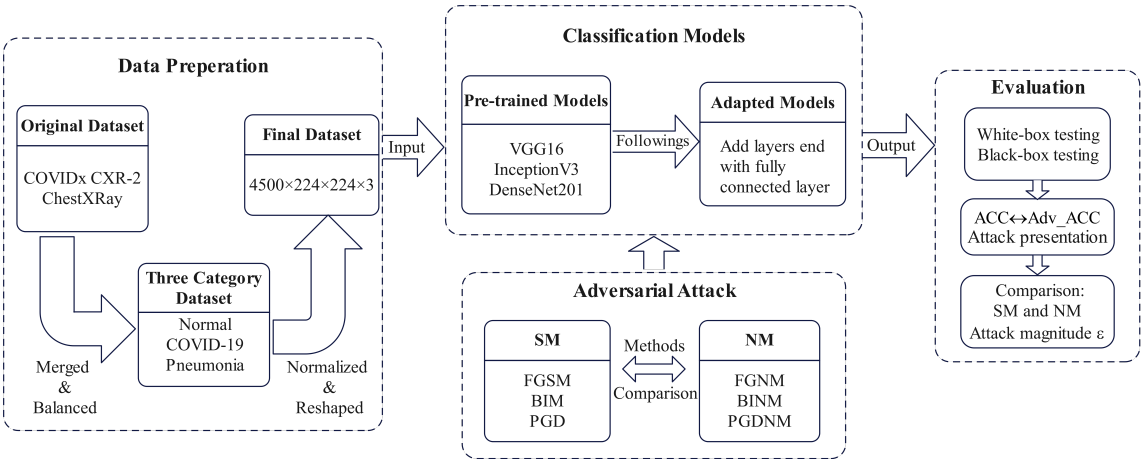

图1：实验流程汇总

该实验分为四个部分，如图1所示。第一部分是数据准备。对收集到的数据进行合并、平衡、归一化和重塑，得到一个有4500张图像的数据集，其中3600张用于训练模型，900张作为测试集的长度。在训练集中，提取10%的图像（360幅）用于验证过程。 然后将VGG16、Inception V3和具有适应层的DenseNet201首先与预训练模型的冻结参数进行拟合。而后我们解冻所有那些无法训练的参数，并以相对较低的学习率对模型进行微调。首先获得了清洁测试集的纯分类精度。训练模型过程中使用的超参数列举如下：

- Loss function: Categorical Cross-Entropy.
- Optimizer: Adam.
- Batch size: 64.
- Decay-Learning rate: 5e-4 (decay = 5e-6).
- Worker number: 15.

第三部分是关于攻击的。我们使用VGG16作为白盒模型计算梯度，我们在测试集上随机选取64张图片为一组，使用NM和SM方法计算对抗性样本。生成对抗性样本后，白盒分类准确率由VGG16测试，黑盒分类准确率由Inception V3和DenseNet201测试。在表Ⅱ和表Ⅲ中，白盒测试准确率用*标记，其余为黑盒测试结果。攻击算法参考了Cleverhans[29]开源库。最后，为了全面比较两类方法的性能，我们进一步讨论了DNN在特定扰动下的脆弱性。分析和结论来自于表格中记录的分类精度降低的比较。

2. 结果分析

VGG16、InceptionV3和DenseNet201在原始胸部X光数据集上的三分类准确率分别为92.22%、90.11%和96.78%。FGSM和FGNM的比较结果见表Ⅱ，表格记录了扰动强度从0.0005增加到0.03时三种网络的分类准确率。由于VGG16是白盒模型，一个小的扰动就足以对原始分类器造成巨大破坏。显然，在一定的扰动强度范围内，白盒攻击通常比黑盒攻击更有效，这种结果符合理论上的预期。在 $\epsilon = 0.01$，0.02 时，最有效的FGSM、FGNM白盒攻击将准确率分别拉低到7.67%和7.33%。当 $\epsilon \in (0.001, 0.005)$时，NM的优势非常明显，NM的表现比SM好，准确率拉低了30%。随着干扰强度的增加，NM的优点变得狭窄。在0.006和0.03之间的区间内，对FGSM和FGNM的白盒攻击都很有效，使分类精度降低了约80%。而NM更微妙，因为较小的扰动强度会触发更好的白盒攻击。更重要的是，随着扰动强度的增加，黑盒攻击会逐渐变强。这是可以接受的，因为黑箱测试本质上是不依赖已知梯度的，因此削弱了其效率。然而，由于人眼很容易分辨出医学图像上的扰动，攻击强度不宜过大。此外，我们在0.03处切断了ϵ，不再增加扰动强度。关于扰动强度约束的细节将在第讨论部分解释。总而言之，对比SM和NM，NM的强度在0.01的阈值下更有效。

表Ⅱ：FGSM和FGNM攻击方法在白盒和黑盒测试中的定量比较

FG	ϵ	VGG	Inv3	Den	ϵ	VGG	Inv3	Den
SM	0	92.22	90.11	96.78	6e-3	11.89	79.89	83.22
NM		92.22	90.11	96.78		**8.11**	74.44	75.78
SM	5e-4	88.44	89.67	96.78	8e-3	9.22	74.44	76.00
NM		**78.44**	89.78	96.22		**7.78**	70.67	67.67
SM	1e-3	77.89	89.22	96.22	1e-2	8.11	68.56	68.89
NM		**55.67**	88.78	95.00		**7.67**	66.00	57.78
SM	2e-3	54.22	87.89	94.22	2e-2	**7.33**	54.89	51.56
NM		**24.44**	86.67	92.00		7.78	58.89	41.00
SM	4e-3	22.22	85.00	88.89	3e-2	7.89	38.67	33.78
NM		**10.11**	81.22	84.22		13.89	48.56	33.44

表Ⅲ记录了PGD、BIM和PGDNM以及BIMNM在迭代扰动强度范围在$(0.0005，0.003)$时的比较。迭代总次数设定为$T = 10$，用L_∞规范球将邻域从$x - 1$截断到$x + 1$。很明显，扰动强度对分类精度的影响不是线性的。值得注意的是，NM的攻击效率普遍高于SM，白盒测试的最佳性能达到6.89%。另一方面，与表Ⅱ所示的非迭代方法相比，在实施控制总攻击幅度相同的情况下，迭代攻击方法比非迭代攻击更有效率。此外，PGD并没有从嘈杂的初始点中获益多少，因为PGD和BIM的性能几乎相同。迭代的ϵ在0.003的边缘被切断，因为较大的扰动强度是无法察觉的，这是没有意义的。

表Ⅲ：PGD、BIM和PGDNM以及BIMNM攻击方法在白盒和黑盒测试中的定量比较

ϵ	BIM	VGG	Inv3	Den	PGD	VGG	Inv3	Den
5e-5	SM	87.89	89.89	96.67	SM	87.89	89.89	89.89
	NM	76.22	89.89	96.56	NM	**76.11**	96.67	96.56
1e-4	SM	76.00	89.33	96.33	SM	76.00	89.33	96.33
	NM	**46.78**	89.11	95.22	NM	46.89	89.11	95.22
2e-4	SM	45.56	88.45	94.67	SM	45.56	88.44	94.67
	NM	**12.22**	86.22	92.89	NM	**12.22**	86.22	93.00
3e-4	SM	23.44	87.78	93.00	SM	23.44	87.78	93.00
	NM	**7.22**	84.33	90.33	NM	**7.22**	84.56	90.45
6e-4	SM	7.22	82.78	86.56	SM	7.22	82.78	86.56
	NM	**7.00**	79.56	79.56	NM	**7.00**	79.56	79.33
1e-3	SM	7.00	74.56	76.22	SM	7.00	74.56	76.22
	NM	**7.00**	76.67	74.00	NM	**7.00**	78.00	74.22
3e-3	SM	7.00	51.11	43.89	SM	7.00	51.11	43.89
	NM	**6.89**	67.56	57.78	NM	**6.89**	67.89	57.89

3. 攻击表现

我们在图2和图3中展示了两组攻击。两组原始图像都选择了COVID胸部X光类。白盒测试模型将对抗性样本错误地分类为正常类。图2中利用的对抗性攻击算法是FGNM，图3中实现的攻击算法是PGDNM。我们有意选择了两种扰动强度来产生不同的对抗性样本。第一个扰动强度比第二个扰动强度小10倍。由于对抗性样本中的第二个扰动是人类可以感知的，而第一个扰动是不可感知的，所以第二个攻击失败，第一个攻击成功。由于迭代的特性，根据表Ⅱ和表Ⅲ，PGDNM选择的扰动强度比FGNM小10倍。为了更加直观，我们显示了扰动强度增加了一千倍。

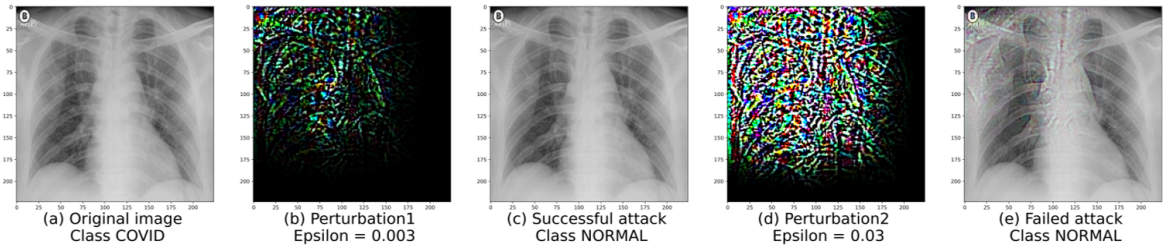

图2：FGNM攻击性能结果的比较

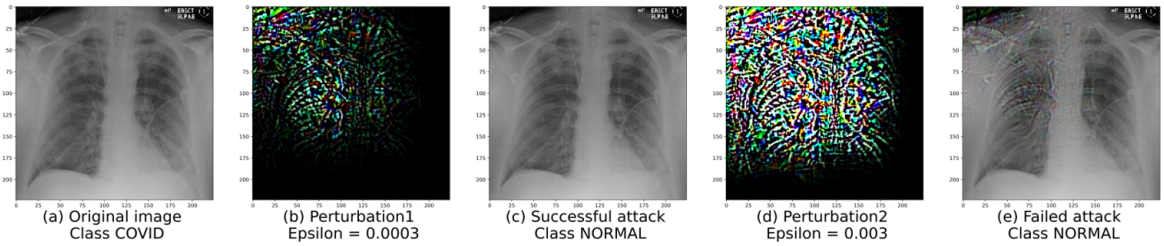

图3：PGDNM攻击性能结果的比较

五、讨论

1. NM表现好的原因

无符号攻击方法比符号攻击方法更有效，在图4中黄色曲线（NM）下降到蓝色曲线（SM）之前。受益于无符号算子的良好攻击可以在较低的攻击强度下实现，这与原始图像相比是相对难以察觉的。

无符号算子保持了符号法扰动的振幅，在方向上与梯度一致。基于梯度的攻击与网络训练时的梯度下降方向相悖，而在对抗性测试中理论上和实践上都表现得更好。横向比较图4中的两个图，迭代和非迭代的扰动强度几乎满足$\epsilon = \alpha \cdot T$。

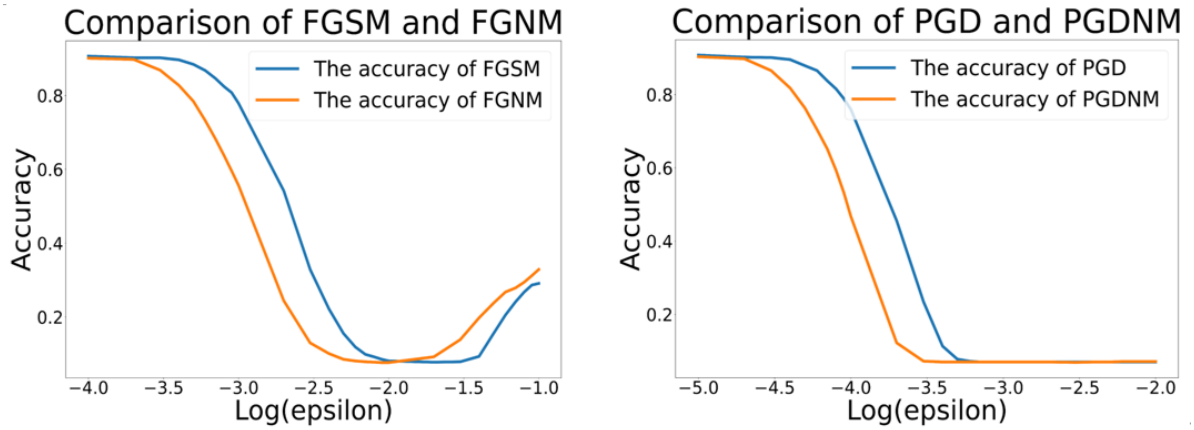

图4：攻击强度与效果对比

2. ϵ 边界

对抗性攻击应满足一个基本原则，即人眼无法区分对抗性样本和原始样本之间的差异。当扰动强度超过边界时，对抗性攻击将是人类可以感知的，这是没有意义的。因此，讨论扰动强度的上限是至关重要的。

图3(e) 和图4(e) 显示了攻击强度为0.003和0.03时FGNM和PGDNM的攻击结果。对抗性图像明显受到了污染。因此，在进行对抗性性能的比较分析时，我们在表Ⅱ和表Ⅲ中设置了扰动强度的上限。这样一来，成功的对抗性攻击的效率得到了保证，对抗性样本也保证了不被察觉。

在实验中，扰动强度的适当边界是通过绘制攻击图像人为选择的。这里有一个问题，如何能快速找到ϵ的边界。基于数字图像处理的认知，一个原始的想法是在一个固定的域上计算图像和它的灰度的卷积。不等式可以由某一图像的扰动低于其灰度的平均值的关系来构建。然后，通过解决相关的不等式，将得到 ϵ 边缘的数值解。我们打算在未来的工作中进一步研究这个想法。

3. 图像特异性讨论

在医疗领域，图像往往是单色的，这意味着它们容易受到攻击。也许它们可以被转换为单通道灰度图像。这样的转换混合了局部特征，尽管是以失去图像的某些特征为代价，这可能有助于提高医学图像的分割和分类精度。 对于对抗性攻击的研究，背景对原始图像特征提取的影响不能被强调。而目标分割和识别需要更强大的鲁棒性DNNs的支持。

六、总结

在本文中，从算法的角度比较了有符号和无符号攻击方法的性能。在对扰动强度进行了一系列的比较后，一般来说，NM比SM更有效率。基于COVID数据集，进一步分析了医学图像的脆弱性和特殊性。此外，还提出了一个相关问题，即攻击强度的量化边界。我们的工作对构建具有鲁棒性的深度神经网络系统有参考价值。

参考文献

[1] Worldometer, “COVID-19 CORONAVIRUS PANDEMIC.” https://www.worldometers.info/coronavirus/

[2] E. Mahase, “Covid-19: What do we know about the delta omicron recombinant variant?” British Medical Journal Publishing Group, 2022.

[3] Y. Fang et al., “Sensitivity of Chest CT for COVID-19: Comparison to RT-PCR,” Radiology, vol. 296, no. 2, pp. E115–E117, Aug. 2020, doi: 10.1148/radiol.2020200432.

[4] C. Szegedy et al., “Intriguing properties of neural networks,” arXiv preprint arXiv:1312.6199, 2013.

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” arXiv preprint arXiv:1412.6572, 2014.

[6] A. Kurakin, I. J. Goodfellow, and S. Bengio, “Adversarial examples in the physical world,” in Artificial intelligence safety and security, Chapman and Hall/CRC, 2018, pp. 99–112.

[7] Y. Dong et al., “Boosting adversarial attacks with momentum,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 9185–9193.

[8] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” arXiv preprint arXiv:1706.06083, 2017.

[9] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in 2017 ieee symposium on security and privacy (sp), 2017, pp. 39–57.

[10] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, “Universal adversarial perturbations,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1765–1773.

[11] S. Kaviani, K. J. Han, and I. Sohn, “Adversarial attacks and defenses on AI in medical imaging informatics: A survey,” Expert Systems with Applications, p. 116815, 2022.

[12] Y. Cheng, X. Zhu, Q. Zhang, L. Gao, and J. Song, “Fast Gradient Non-sign Methods,” arXiv preprint arXiv:2110.12734, 2021.

[13] A. S. Panayides et al., “AI in Medical Imaging Informatics: Current Challenges and Future Directions,” IEEE J. Biomed. Health Inform., vol. 24, no. 7, pp. 1837–1857, Jul. 2020, doi: 10.1109/JBHI.2020.2991043.

[14] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in International Conference on Medical image computing and computer-assisted intervention, 2015, pp. 234–241.

[15] A. Esteva et al., “Dermatologist-level classification of skin cancer with deep neural networks,” nature, vol. 542, no. 7639, pp. 115–118, 2017.

[16] D. S. Kermany et al., “Identifying medical diagnoses and treatable diseases by image-based deep learning,” Cell, vol. 172, no. 5, pp. 1122–1131, 2018.

[17] A. K. Jaiswal, P. Tiwari, S. Kumar, D. Gupta, A. Khanna, and J. J. Rodrigues, “Identifying pneumonia in chest X-rays: a deep learning approach,” Measurement, vol. 145, pp. 511–518, 2019.

[18] T. Rahman et al., “Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest X-ray,” Applied Sciences, vol. 10, no. 9, p. 3233, 2020.

[19] L. Wang and A. Wong, “COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images,” arXiv:2003.09871 [cs, eess], May 2020, Accessed: Apr. 18, 2022. [Online]. Available: http://arxiv.org/abs/2003.09871

[20] A. Nguyen, J. Yosinski, and J. Clune, “Deep neural networks are easily fooled: High confidence predictions for unrecognizable images,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 427–436.

[21] H. Kannan, A. Kurakin, and I. Goodfellow, “Adversarial logit pairing,” arXiv preprint arXiv:1803.06373, 2018.

[22] F. Chollet, “Keras Applications.” https://keras.io/api/applications/

[23] D. Kermany, K. Zhang, M. Goldbaum, and others, “Labeled optical coherence tomography (oct) and chest x-ray images for classification,” *Mendeley data*, vol. 2, no. 2, 2018.

[24] B. Wang, S. Qiu, and H. He, “Dual encoding u-net for retinal vessel segmentation,” in International conference on medical image computing and computer-assisted intervention, 2019, pp. 84–92.

[25] Y. Ikeda, K. Doman, Y. Mekada, and S. Nawano, “Lesion Image Generation Using Conditional GAN for Metastatic Liver Cancer Detection,” Journal of Image and Graphics, vol. 9, no. 1, 2021.

[26] Y.-C. Chen, D.-R. Chen, H.-K. Wu, and Y.-L. Huang, “Intra-operative Tumor Margin Evaluation in Breast-Conserving Surgery with Deep Learning,” Journal of Image and Graphics, vol. 7, no. 3, 2019.

[27] S. Kamath, K. Prasad, K. Rajagopal, and others, “Segmentation of breast thermogram images for the detection of breast cancer: a projection profile approach,” 2015.

[28] A. A. Dovganich, A. V. Khvostikov, Y. A. Pchelintsev, A. A. Krylov, Y. Ding, and M. C. Q. Farias, “Automatic Out-of-Distribution Detection Methods for Improving the Deep Learning Classification of Pulmonary X-ray Images,” Journal of Image and Graphics, vol. 10, pp. 56–63, 2022.

[29] N. Papernot *et al.*, “Technical Report on the CleverHans v2.1.0 Adversarial Examples Library,” *arXiv preprint arXiv:1610.00768*, 2018.