Adversarially Robust Supervised Contrastive Learning for More Interpretable Computer-Aided Diagnosis

Cindy Lin* and Ryan Wang*
Massachusetts Institute of Technology
Cambridge, MA
{cindylin, ryanwang}@mit.edu

Abstract

Computer-aided diagnosis out-performs physicians, however, ML adoption in the medical field is very limited. We focus on addressing two of the major limitations in medical ML adoption, brittleness and interpretability. We verify that supervised contrastive pre-training outperforms baseline cross-entropy for classification with histopathological data. We show that adversarially trained models are less brittle and more interpretable using Integrated Gradients, however are sensitive to high and low contrast medical imaging. We show that fine-tuning adversarially pre-trained models results in superior generalization performance and interpretability.

1. Introduction

Computer vision and deep learning in medical image analysis is becoming increasingly pivotal to improving the effectiveness and accuracy of medical diagnoses. Deep learning algorithms have been utilized to build effective systems for cancer detection, diagnostic object detection in X-rays, and synthetic medical image generation [11]. While these algorithms prove promising to the advancement of medicine, they are not without flaws, resulting in the lack of adoption by practitioners.

This study tackles two major drawbacks in medical image classification: the brittleness of medical classifiers and the lack of interpretability of most medical imaging models, as many latch on to spurious features. A *brittle* neural network is one that quickly changes it's output based on a small change of input. This can be shown through *adversarial attacks* on a network, where noise is added to an input that maximizes the chance of predicting the wrong output. Improving model robustness against adversarial inputs may prevent this brittleness [7], increasing practitioner confidence. It has also been shown that adversarially robust models can result in more interpretable feature contributions to predictions [4]. Furthermore, alternate training methods such as supervised contrastive learning result in more robust models that even outperform standard cross-entropy [9], however interpretability with this learning method has not been fully explored. Understanding what contributes to a prediction while achieving high performance may help physicians or patients understand why assistive medical technologies make a given prediction and allow integration of their prior knowledge into final life-changing decisions.

1.1. Our Contribution

Adversarially robust cross-entropy training builds robust models and yet, tends to have poor validation performance as shown experimentally in Alexander Madry's 2018 NIPS tutorial. We build upon existing methods to explore a potential novel method of *improving model interpretability* while achieving *high performance* through by combining adversarial training with supervised contrastive learning. We evaluate our model on two datasets described in §3.1, §3.2 for domain-specific and domain-generalizable performance, respectively. We evaluate classification performance comparing our optimization process to baseline cross entropy, supervised contrastive, and adversarial training combinations. We investigate model interpretability and behavior under various adversarial attacks.

2. Related Work

2.1. Adversarial Robustness

The most common method to combat adversarial attacks is through *adversarial training* [2]. These often involve training on not only observed data, but adversarial data [17]. Madry [17] achieves this through extending Fast Gradient Sign Method (FGSM) to multi-step FGSM for an ℓ_∞-bounded adversary. *Adversarial Regularization* is another method that attempts to regularize loss functions using adversarial examples [2]. These methods are computationally

*Equal contribution, also affiliated with Harvard T.H. Chan SPH

expensive and other directions aim at improving *Efficient Adversarial Training*, for example Free-AT [22].

Adversarial robustness has previously been shown to have more interpretable saliency maps [4]. We expand on this by investigating Integrated Gradients [18], a common interpretability method for feature attribution of predictions.

2.2. Contrastive Learning

Contrastive learning is a representation learning framework that keeps *similar* observations together and pushes *dissimilar* observations apart [14]. These have been very successful in both visual representation learning and in multi-modal learning [3], [20].

Supervised Contrastive Learning (SupCon) was proposed by [9] to incorporate label information into the contrastive learning framework, out-performing cross-entropy. This method considers pushing observations of the same label and their augmentations (data *views*) together, while separating observations of different labels.

2.3. Similar Work

Li [15] takes a mixture of cross-entropy and adversarial contrastive loss in order to perform adversarially robust supervised contrastive training with natural language data. Kim [10] investigates self-supervised contrastive adversarial training, using adversarial data as a data augmentation.

Our work distinguishes from similar work by directly using supervised contrastive loss, rather than taking a loss mixture. We train on adversarially attacked views every training iteration in addition to training on the original data. Furthermore, we fine-tune the robust supervised contrastive representations with cross-entropy for classification.

3. Method

3.1. Breast Ultrasounds

Data Description The Breast Ultrasounds dataset [1] is a collection of breast ultrasound images of 600 women between the ages of 25 to 65 from a single hospital. The primary purpose of the ultrasounds is to identify breast cancer and thus, each image is categorized into non-malignant or malignant.

Pre-processing The images were loaded as size (224x224) images. For training involving SupCon Loss, we use two random augmentations per image in order to create two views for input.

Baseline Network A baseline network built on ResNet50 [5] was utilized for preliminary classification on this dataset. We choose this common network architecture as we believe this training algorithm to be model-agnostic and it was also used in the original SupCon [9]. The last fully-connected layer of the default ResNet50 model was removed for SupCon. For fine-tuned models, we take this feature encoder and attach a 1-layer MLP classifier head with dropout 0.2 and batch norm [6] (re-initializing for adversarial CE fine-tuning).

Training We minimize CE and follow the algorithms in §3.3. We use the Adam [12] optimizer with a learning rate of 0.001 for CE and 0.0003 for SupCon. The model was trained and validated on 20 epochs with batch sizes of 10 for CE and 64 for SupCon. We select the model with the best validation accuracy.

3.2. WILDS: Camelyon17

Data Description Stanford's WILDS' Camelyon17 dataset [19] was curated for domain adaptation and include tumour histopathology data from three hospital sites in the training set. The outcome is no tumor or tumor presence. The train, validation, and test sets are all from disjoint hospitals. This data will be used to evaluate generalization performance of our robust training method.

Pre-processing Images are upscaled to 224x224. We augment using horizontal flips, color jitter, grayscale, and normalization. For SupCon Loss, we use two random augmentations per image in order to create two views. We use 5000 label-balanced (50/50) samples for training, 5000 for validation, and 10000 for testing. We use this subset of $\geq 300,000$ samples available due to limited compute.

Basline Network We use the same network as in §3.1.

Training We monitor training behaviour to choose training hyperparameters. We use a batch size of 64, and use the AdamW [16] optimizer with a learning rate of 0.0003. We apply exponential learning rate scheduling with a factor of 0.85 for all algorithms. Appendix §8.2 for training details.

3.3. Our algorithm

3.3.1 Baseline Adversarial Training

First, we consider standard cross entropy adversarial training. Training iteration t is given by CROSSENTROPY-ADV-TRAIN.

Algorithm 1 CROSSENTROPY-ADV-TRAIN

$\theta^{t'} \leftarrow \theta^t - \alpha \nabla_{\theta^t} L(h_\theta(x), y)$ ▷ Standard Epoch
for m in $0 : M-1$ **do** ▷ Multi-step adversary
  $\delta^{m+1} \leftarrow \mathcal{P}_\infty \left(\delta^m + \alpha_\delta \nabla_{\delta^m} CE(h_{\theta^{t'}}(x+\delta^m), y)\right)$
end for
$\theta^{t+1} \leftarrow \theta^{t'} - \alpha \nabla_{\theta^{t'}} L\left(h_{\theta^{t'}}(x) + \delta^M, y\right)$ ▷ Adv. Ep

Here, CE is Cross-Entropy Loss and $h_\theta(\cdot) = z.$ is a model parameterized by θ. $\mathcal{P}_\infty$ denotes M-step projected gradient descent onto the ℓ_∞ ball, where the resulting adversarial noise update is given by [17]

$$\delta^{m+1} \leftarrow \delta^m + \alpha_\delta \text{sign}(\nabla_{\delta^m} L(h_\theta(x+\delta^m), y))$$

where, the update to be added to δ^m is clamped in $[-\epsilon, \epsilon]$, corresponding to $\mathcal{P}_\infty(\cdot)$. This general framework is described in more detail in Madry's NIPS tutorial.

3.3.2 Supervised Contrastive Adversarial Training

Next, we detail our proposed method, **adversarially robust supervised contrastive training**. Suppose we are on epoch t of training with the supervised contrastive objective [9],

$$L = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\{z_i \cdot z_p / \tau\}}{\sum_{a \in A(i)} \exp\{z_i \cdot z_a / \tau\}}$$

where $h_\theta(x_i) = z_i$ is the representation for observation i. I is the index of a sample, $P(i)$ denotes the set of positive indices (same label and one other augmentation of the current datapoint), $A(i)$ denotes the samples other than i. Training iteration t is given by SUPCON-ADV-TRAIN.

Algorithm 2 SUPCON-ADV-TRAIN

$\theta^{t'} \leftarrow \theta^t - \alpha \nabla_{\theta^t} L([h_\theta(x_1)||h_\theta(x_2)], y)$ ▷ Standard Epoch
for m in $0 : M-1$ **do** ▷ Multi-step adversary
  $g_1^m \leftarrow \nabla_{\delta_1^m} L(h_{\theta^{t'}}(x_1 + \delta_1^m), y)$ ▷ Adv. View 1
  $g_2^m \leftarrow \nabla_{\delta_2^m} L(h_{\theta^{t'}}(x_2 + \delta_2^m), y)$ ▷ Adv. View 2
  $\delta_1^{m+1} \leftarrow \mathcal{P}_\infty(\delta_1^m + \alpha_{\delta_1} g_1^m)$
  $\delta_2^{m+1} \leftarrow \mathcal{P}_\infty(\delta_2^m + \alpha_{\delta_2} g_2^m)$
end for
$h_{\theta^{t'}, \delta^M}(x) \leftarrow [h_{\theta^{t'}}(x_1 + \delta_1^M)||h_{\theta^{t'}}(x_2 + \delta_2^M)]$
$x \leftarrow [x_1||x_2]$
$\theta^{t+1} \leftarrow \theta^{t'} - \alpha \nabla_{\theta^{t'}} L\left(h_{\theta^{t'}, \delta^M}(x), y\right)$ ▷ Adv. Ep

Here, we take in two views of x corresponding to two random augmentations, x_1 and x_2. Next, we add M-step adversarial noise with respect to each individual augmentation in parallel. Finally, we concatenate the two adversarial inputs and update the model based on this attack.

Adversarial Hyperparameters We use hyperparameters as specified in Madry's NIPS tutorial ($\epsilon = 0.2$, $\alpha_\delta = 0.01$), however only apply 5-10 steps of adversarial noise optimization. This is because of computational limitations as the inclusion of an M-step adversary changes training runtime complexity from $\mathcal{O}(N)$ to $\mathcal{O}(NM)$ where N is the number of epochs.

3.4. Model Interpretability

We use the Integrated Gradient framework implemented by [13] in order to determine feature attribution to final predictions [18]. Integrated Gradients is defined as

$$IG_i(f, x, x') = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial f(x' + \alpha(x - x'))}{\partial x_i} d\alpha$$

This can be interpreted as the feature-attribution along dimension i of input x interpolated from some baseline input x' which represents the *absence* of that input [23]. This method provides a pixel-level local interpretation allowing for more granularity than other interpretability methods, and it's simple implementation makes it our first choice.

4. Results and Discussion

4.1. Classification Performance

4.1.1 Breast Ultrasound Dataset

Ultrasound	Test Accuracy %	Test AUC
CE	85.86	**0.6424**
Adv. CE	**93.94**	0.4675
SupCon FT	37.37	0.3072
Adv. SupCon FT	37.40	0.3089

Table 1. Accuracy and AUC on Breast Ultrasound Dataset, trained on 20 epochs each. CE = Cross-Entropy Minimization. Adv. CE = CROSSENTROPY-ADV-TRAIN. SupCon FT = SupCon then CE fine-tuning. Adv. SupCon FT = SUPCON-ADV-TRAIN then CE fine-tuning.

We show the results of all our trained classifiers in Table 1 for the Ultrasound dataset. The baseline classifier, which categorized breast ultrasound images into two classes: non-malignant and malignant, showed a decent high level of accuracy, with 85.86% test accuracy and an AUC of 0.6426 over 20 epochs. Though the adversarial cross-entropy had a high test accuracy of 93.94%, the AUC score was low, at 0.4675, suggesting the model performance is heavily impacted by threshold selection.

The performance of supervised contrastive loss was considerably poor. This may be due to our low batch size, lack of discriminant features, limited image-to-image variability of ultrasounds and cohort size, or perhaps the lack of informativeness of individual ultrasounds in clinical settings. Coupling mammography or biopsy images with ultrasounds may be a more clinically robust method for future classification tasks. Further investigation is needed to generalize our method for organ-level medical images.

4.1.2 Camelyon17 Histopathological Dataset

We show the Camelyon17 domain generalization results of all our trained classifiers in Table 2 and use CE as a baseline since it is the standard classification training procedure. CROSSENTROPY-ADV-TRAIN alone results in the worst test accuracy and test AUC, but as later be explored in §4.2, has the most detailed interpretability map. Fine-tuning this robust classifier results in the best performing model. Close behind was both the CE-fine-tuned SupCon

Camelyon17	Test Accuracy %	Test AUC
CE	87.81	0.9444
Adv. CE	71.45	0.7828
Adv. CE FT	**90.21**	**0.9667**
SupCon FT	89.37	0.9596
Adv. SupCon FT (**ours**)	88.40	0.9517

Table 2. Accuracy and AUC on the Camelyon17 dataset after 30 epochs for CE, 20 epochs for SupCon, and 30 epochs for fine-tuning on top of the previous. Adv. CE FT = CROSSENTROPY-ADV-TRAIN, then CE fine-tuning. CE, Adv. CE, SupCon FT, Adv. SupCon FT defined in Table 1

model and then our CE-fine-tuned robust SupCon model. These suggest plain adversarial training have poor generalization performance, since these test data came from a different hospital than validation and training sets. However, fine-tuning on top of a pre-training procedure (adversarial CE, SupCon, and adversarial SupCon) seems to benefit generalization performance the most.

Our method does slightly worse than standard SupCon, however outperforms cross-entropy. This is interesting as it also results in more robust feature maps as §4.2 will discuss, with only a minor drop in performance had we used SupCon. The worse performance of our method compared with supervised contrastive learning matches cross-entropy results examined in Madry's NIPS tutorial. The robust classifier outperforms our proposed method since it may have a better starting point due to pre-training on the same task.

4.2. Interpretability

Figure 1 shows the results of integrated gradient maps on the Cameylon17 dataset for the following regimen: standard cross entropy, adversarial cross-entropy, adversarial cross-entropy fine-tuned, standard supervised contrastive, adversarial supervised contrastive. Comparing the plots generated from each of the five regimes, fine-tuned adversarial CE is most robust to adversarial attack while, non-fine-tuned adversarial CE has the most recognizable features, however lacks detail in the edge cells. Fine-tuned adversarial CE, SupCon, and adversarial SupCon are activated by more meaningful parts of the images (the cells in the corners as opposed to the adipose cell in the center), however only fine-tuned adversarial CE retains the correct prediction after attack (results are mixed in other examples, see §8.4). Noticeably, the spurious activations in the fat increase as we attack the image, and while the SupCon and standard CE models lose feature map definition, the adversarially trained models retain the general shape and feature attribution locations, even after non-adversarial fine-tuning. This illustrates the effectiveness of adversarially trained models in learning more relevant feature attributions for clinical data.

In contrast, interpretability plots were less ideal for breast ultrasound images, perhaps due to the low-contrast and overall murkiness of image details (see Appendix §8.4).

5. Discussion

Our study demonstrated that incorporating adversarial training methods with perturbed inputs resulted in less brittle models that may be better suited to handle the large heterogeneity of medical images and inspire confidence in practitioners. While the fine-tuned robust CE model outperformed our proposed method in classification, our robust SupCon pre-training can allow for more general task-agnostic representations, rather than being specified to a classification task, while still performing relatively well. Additionally, these fine-tuned representations enable our models to identify discriminating features and image representations that are critical for effective medical diagnosis on **histopathological images** and ultimately, improved patient outcomes.

With Integrated Gradients, we verified that higher-level image features, although not clinically relevant, were contributing to final predictions, primarily with the non-finetuned robust CE model. We showed that fine-tuned adversarially robust methods could capture the clinically relevant features in **histopathological images**. This may help address the skepticism among physicians regarding technological assistants. Clear and interpretable plots illustrating salient features in medical images, upon which deep learning models heavily rely on for making classifications, can significantly contribute to the enhancement of physicians' understanding, prior knowledge integration, and trust in black box models for medical diagnoses.

5.1. Limitations

Despite our improved interpretability and our proposed methods relatively high accuracy, this study is not without its limitations. One key limitation is the technique's poor performance and suboptimal interpretability on ultrasound images. Ultrasound images tend to be relatively low-contrast with some images being extremely dark and can result in greater difficulty displaying meaningful features. Lack of image-to-image variability, high levels of noise, and limited definition in ultrasound images may also contribute to poor performance in the SupCon model. Consequently, our method may falter with other low-contrast, organ-level medical images such as X-rays.

Another large constraint of this project is compute, which limited our ability to further test and evaluate model performance on a full set of adversarial inputs. Contrastive learning generally requires large batch size (for example the original SupCon paper had a batch size of 6144 [9]), and so more compute could result in better performing models.

Furthermore, a limitation of image-based models is that it excludes other factors that may contribute to diagnoses,

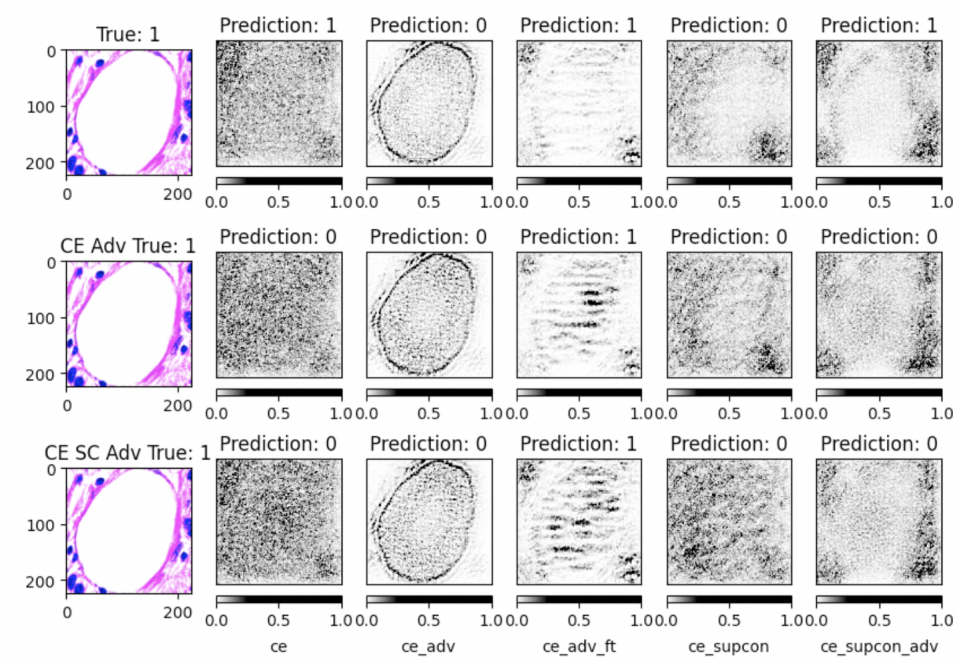

Figure 1. Example of comparisons between the various model predictions and Integrated Gradient maps. The first row is the original data. The second row is data corresponding to a 10 epoch adversarial attack on the CE non-fine-tuned model. The third row is data corresponding to a 10 epoch adversarial attack on the CE fine-tuned SupCon model. `ce` is the cross-entropy trained model, `ce_adv` is the adversarial robust CE model, `ce_adv_ft` is the CE fine-tuned robust model, `ce_supcon` is the CE fine-tuned SupCon model, and `ce_supcon_adv` (**ours**) is the CE fine-tuned robust SupCon model.

including but not limited to: patient health records and/or demographic information, lab test results, and even patient conversations, all of which are information physicians would likely have. Consequently, there is much information that we cannot capture with just 2D image data, and may lead to reduced performance on other tasks.

6. Conclusion

6.1. Future Directions

In the future, we hope to bring this algorithm to scale with more compute, allowing experimentation with other model architectures/robust optimization methods. This would allow for a more rigorous evaluation, especially of the Camelyon17 dataset where we only used a fraction of the total dataset. We also hope to further explore the potential of our robust supervised contrastive framework by applying them to other downstream tasks, such as segmentation, 3D imaging, or in multi-modal applications. In addition, we may opt to freeze the adversarially robust feature encoders and only fine-tune a classifier head in order to compare to fine-tuning the entire encoder. Another avenue of research could be to investigate what feature contributions from an input image contribute to the supervised contrastive loss, rather than the fine-tuned classifier, in order to understand how robust representations behave. Furthermore, to improve the interpretability plots across all 2D medical imaging modalities, including the aforementioned low-contrast, organ-level ultrasound or x-ray images, there is opportunity to explore alternate interpretability methods such as GradCAM or XRAI [21] [8].

6.2. Concluding remarks

Adversarially robust supervised contrastive learning outperforms cross entropy in generalization scenarios, however, is limited by batch size. With a slight decrease in performance from it's non-robust counterpart, we observe less spurious feature attributions. Overall, for a medical task-specific framework like classification, fine-tuning a CE robust classifier is likely the best option. For task-agnostic frameworks, adversarially trained supervised contrastive representation learning seems like a promising method moving forward. Together, these results can potentially improve not only diagnostic performance, but also prediction interpretability to allow for more confidence in ML model adoption in the clinical field.

7. Contributions

- **Cindy** Built baseline ResNet classifier, modified algorithm and evaluated them for Breast Cancer Ultrasound dataset.
- **Ryan** Conceived study, implemented training/interpretability algorithms and evaluated them for the Camelyon17 dataset.

References

[1] Gomaa M. Khaled H. AlDhabyani, W. and A Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 2020. 2

[2] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. *IJCAI*, 2021. 1

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *PMLR*, 2020. 2

[4] Christian Etmann, Sebastian Lunz, Peter Maass, and Carola-Bibiane Schönlieb. On the connection between adversarial robustness and saliency map interpretability, 2019. 1, 2

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 2

[6] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015. 2

[7] Umrao S. Chang E. Choi R. Yang D. X. Duncan J. S. Omuro A. Herbst R. Krumholz H. M. Joel, M. Z. and S. Aneja. Using adversarial images to assess the robustness of deep learning models trained on diagnostic images in oncology. *JCO Clinical Cancer Informatics*, 2022. 1

[8] Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viégas, and Michael Terry. Xrai: Better attributions through regions, 2019. 5

[9] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *NIPS*, 2020. 1, 2, 3, 4

[10] Minseon Kim, Jihoon Tack, and Sung Ju Hwang. Adversarial self-supervised contrastive learning, 2020. 2

[11] Yun J. Cho Y. Shin K. Jang R. Bae H. Kim, M. and N. Kim. Deep learning in medical imaging. *Neurospine*, 2019. 1

[12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 2

[13] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch, 2020. 3

[14] Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. Contrastive representation learning: A framework and review. *IEEE*, 2020. 2

[15] Weidong Li, Bo Zhao, Yang An, Chenhan Shangguan, Minzi Ji, and Anqi Yuan. Supervised contrastive learning for robust text adversarial training. *Neural Computing and Applications*, 2023. 2

[16] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 2

[17] Aleksander Madry, Aleksandar Makelov, Dimitris Tsipras Ludwig Schmidt, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *ICLR*, 2018. 1, 2

[18] Qiqi Yan Mukund Sundararajan, Ankur Taly. Axiomatic attribution for deep networks. *PMLR*, 2017. 2, 3

[19] Henrik Marklund Sang Michael Xie Marvin Zhang Akshay Balsubramani Weihua Hu Michihiro Yasunaga Richard Lanas Phillips Irena Gao Tony Lee Etienne David Ian Stavness Wei Guo Berton A. Earnshaw Imran S. Haque Sara Beery Jure Leskovec Anshul Kundaje Emma Pierson Sergey Levine Chelsea Finn Percy Liang Pang Wei Koh, Shiori Sagawa. Wilds: A benchmark of in-the-wild distribution shifts. *ICML*, 2021. 2

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arxiv*, 2021. 2

[21] Abhishek Das Ramakrishna Vedantam Devi Parikh Dhruv Batra Ramprasaath R. Selvaraju, Michael Cogswell. Grad-cam: Visual explanations from deep networks via gradient-based localization. *IEEE*, 2017. 5

[22] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S. Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *NIPS*, 2019. 2

[23] Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. *Distill*, 2020. https://distill.pub/2020/attribution-baselines. 3

8. Appendix

8.1. Open-sourced code

The code for our project can be found on GitHub.

8.2. Training Curves

8.2.1 Camelyon17

The training for `ce_adv` and `ce_adv_ft` was re-run as we found a bug previously. `ce_adv` was selected at epoch 15 of training with validation accuracy $\approx 63.00\%$. `ce_adv_ft` was selected at epoch 11 with validation accuracy $\approx 78.10\%$. The other models were selected using the highest accuracy from the above plot Figure 2. Supcon was evaluated in Figure 3.

8.2.2 Breast Ultrasound

Figure 4 shows the loss and accuracy of each model over 20 training epochs. Note that with supervised contrastive loss, the model is stuck at a local extrema, a potential result of limited discriminative information from low-contrast ultrasound images.

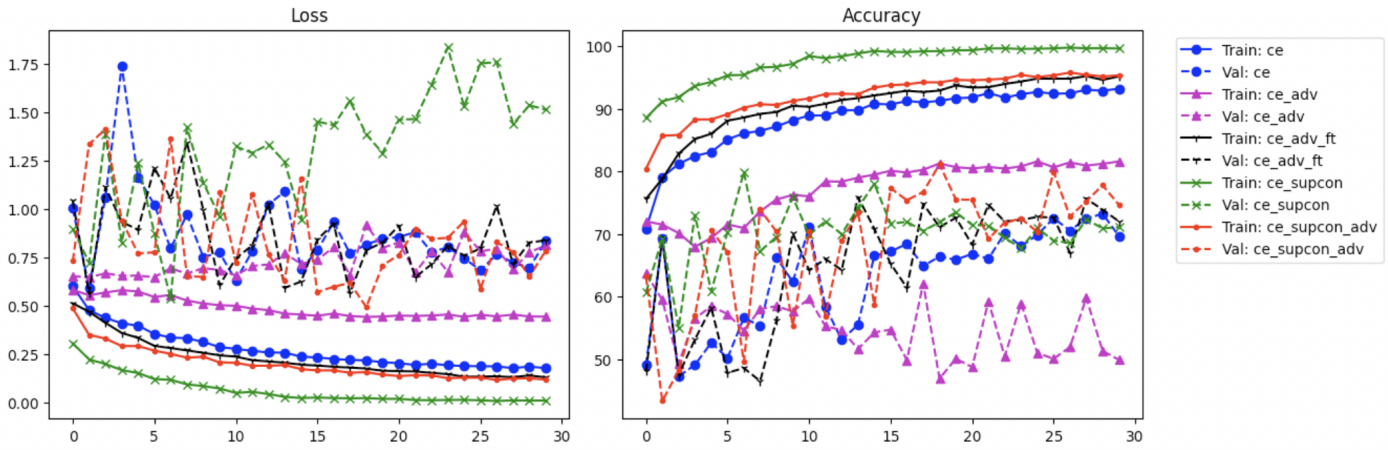

Figure 2. Loss and accuracy of training and validation on the Camelyon17 dataset over 30 epochs.

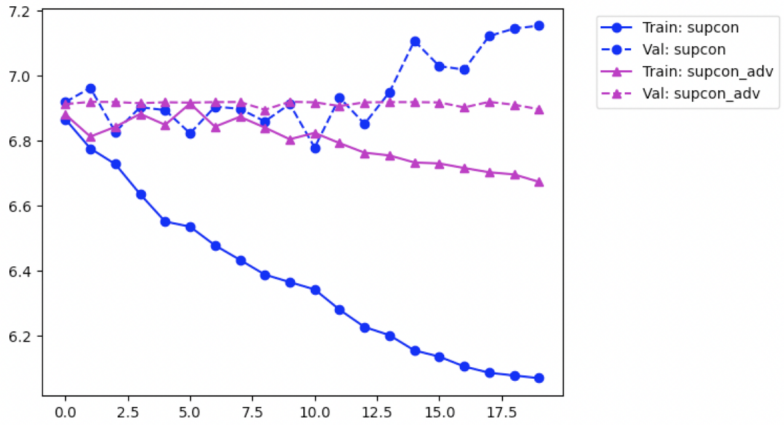

Figure 3. Training curves for SupCon algorithms on the Camelyon17 dataset over 20 epochs.

8.3. Supervised Contrastive Latent Space

As shown in Figure 5, adversarial training adds a bit more noise to the latent space principal components and may have improved separation, however, there is still much overlap. This is potentially due to our very small batch size of 64.

8.4. Interpretability Figures

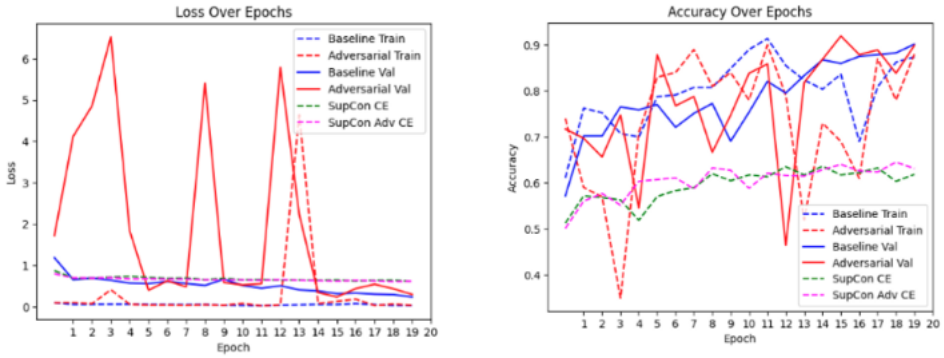

Figure 4. Loss and accuracy of training and validation on the Breast Cancer Ultrasound dataset over 20 epochs.

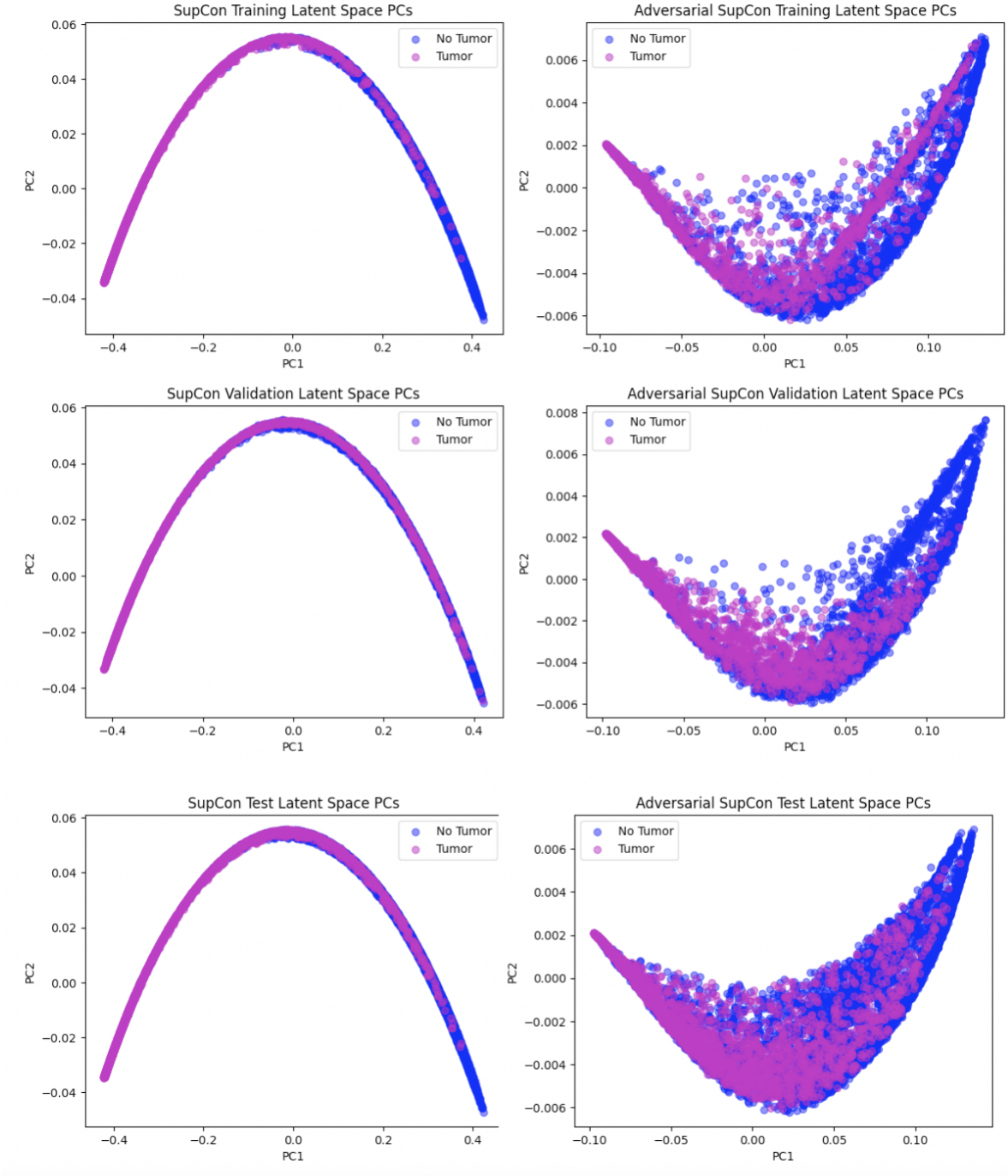

Figure 5. Camelyon17 feature PCA. Adversarial training adds more noise to the latent space.

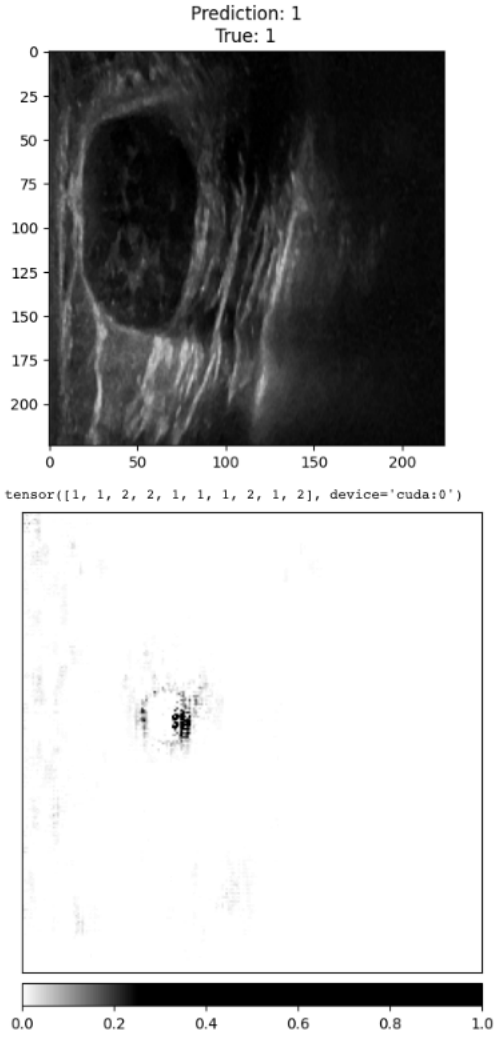

Figure 6. Integrated Gradient Plots for Ultrasound. Due to low-contrast of the image, the plot lacks clear feature visualizations

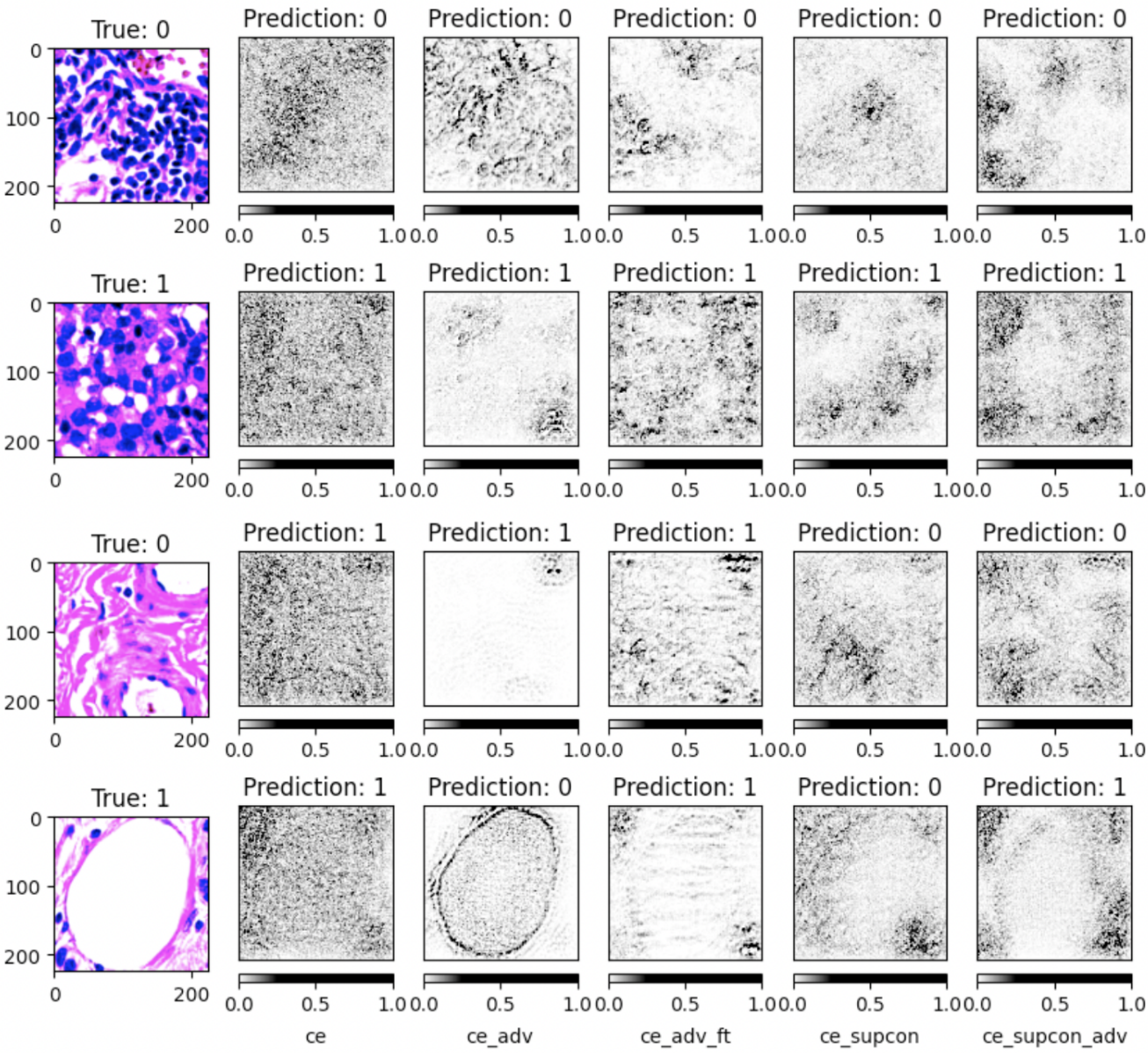

Figure 7. Original data predictions and integrated gradients.

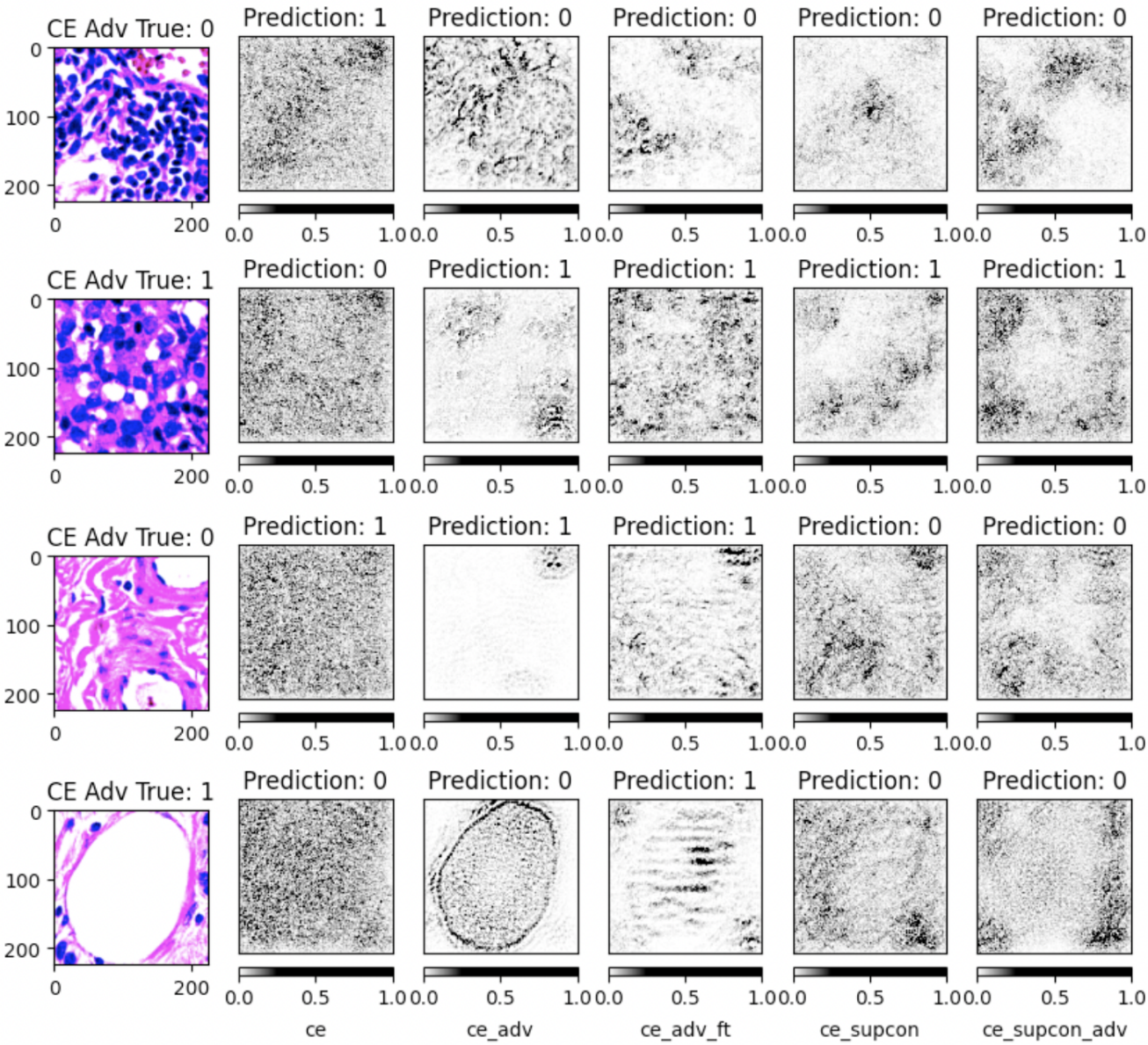

Figure 8. Cross-entropy model adversarially attacked data predictions and integrated gradients.

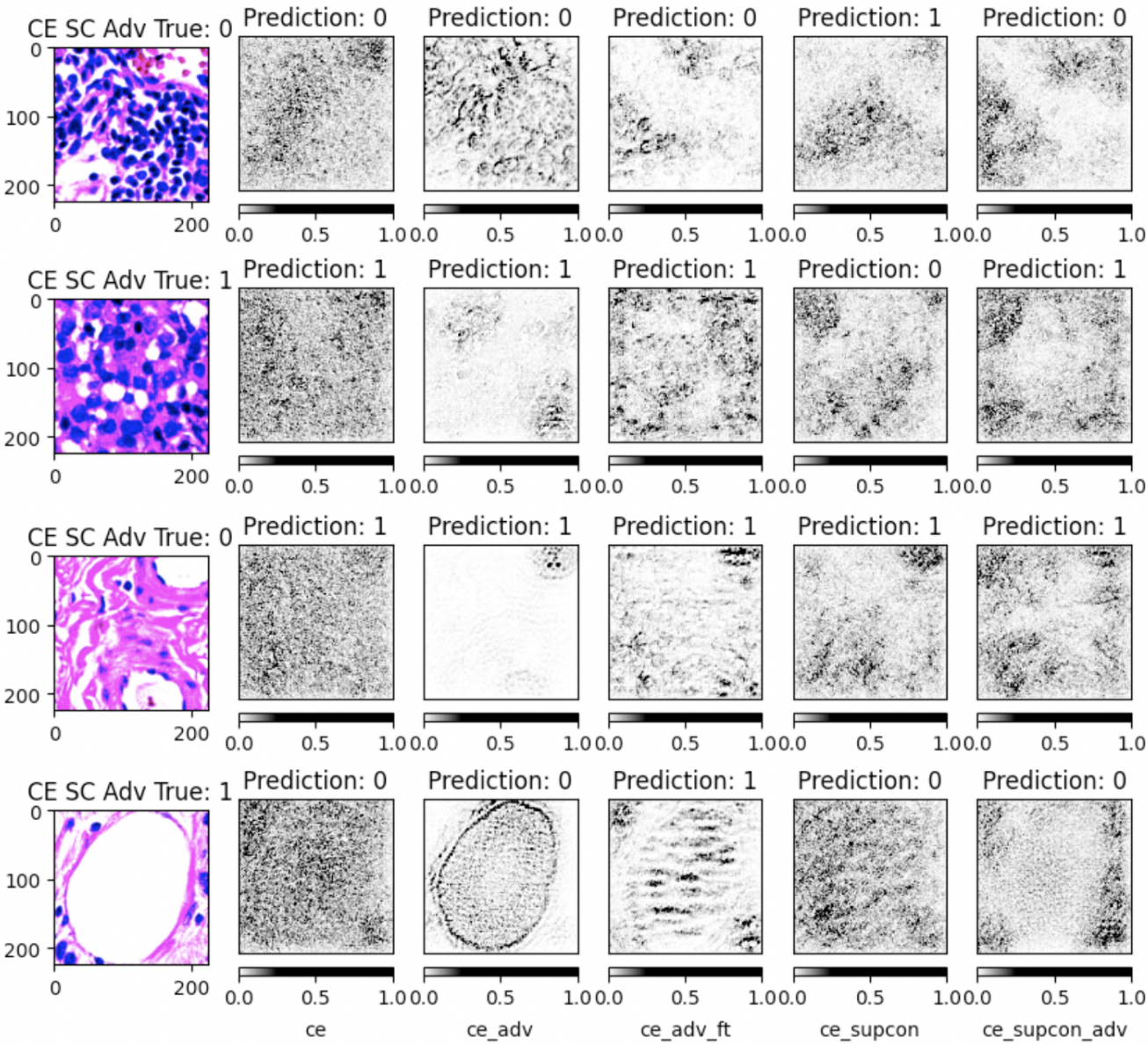

Figure 9. CE fine-tuned supervised contrastive model adversarially attacked data predictions and integrated gradients.