# Uncertainty quantification for adversarial samples in medical imaging

Arttu Häkkinen, Vishnu Raj

February 2, 2022

## 1 Introduction

Deep learning (DL) has proven its potential in the past few years to be able to perform well in practical real world computer vision (CV) classification tasks such as in medical imaging. The diagnostic performance of Deep Neural Networks (DNNs) have been proved to be equivalent with medical professionals [1]. In 2018, the first steps were taken towards the future of DL systems replacing human experts in medical imaging domain: The U.S. Food and Drug Administration approved the first CV model to be utilized in medical diagnosis without any input of a human surveillance [2]. When it comes to clinical imaging diagnosis, fully replacing the expensive and error prone human employees with highly accurate DL systems is be tempting. Despite the high theoretical disease classification performance of DNNs, DL applications have not had their major breakthrough yet in practical medical diagnosis. The complex structure and the black-box nature of the DL classifiers yields to lack of detailed explainability and interpretability of the DNN models.

Additionally, the recent discovery of adversarial attacks [3] has raised concern and doubts against DL classifiers. The adversarial attacks enable the perturbation of the DL systems' input images in a mathematically sophisticated way which will cause any state-of-art DL classifier to misclassify with high confidence. Understanding adversarial attacks has been a hot topic in the research community since their discovery, as they pose a critical threat to cybersecurity in the practical real world DL applications. For example, a harmful entity may try to benefit themselves by manipulating their examination results to fool insurance companies or claim reimbursement from their medical treatments, or a malicious doctor may seek to profit by manipulating test reports which would lead to unnecessary surgeries, and the list goes on. Not to even mention the harmful actions that could be done by the hacker groups in a digital warfare between nations. Obviously, as long as there is even the slightest chance to do any kind of harm whenever a DL system is present in a high-stake decision making process such as medical diagnosis, some sort of exploitation of the system should be expected. Even by combining DL classifiers with surveillance of human health-care professionals, these adversarial methods can be dauntingly deceptive: In most cases they produce such small perturbations in the original images that they are left undetected even by human eye. The evaluation of the robustness as well as inventing potential defense methods against these attacks is highly necessary, since the automated diagnosis systems would otherwise be really beneficial in medical diagnosis.

One possible solution to handle the predictive uncertainty caused by the adversarial attacks, is to put more emphasis on quantifying and communicating the uncertainty present in the predictions of the medical DL classifiers. The term predictive uncertainty refers to taking into account that predictions of the DL model can be unreliable and we should seek to estimate the stability of the prediction. The capability to abstain predictions on samples for which high uncertainty is present would definitely be a step into right direction when it comes to enabling the DL systems in clinical decision making. For example, the most widely cited medical DL models lack the mechanism for abstention when they are uncertain with their predictions. By introducing some caution to clinical DL classifiers would allow safer clinical deployment and more trust within the healthcare workers, since it would be possible to know when the model

is not confident with its prediction. There are two sources of uncertainty: Aleatoric refers to noise in the data, and epistemic is a combination of the uncertainty caused by selection of the model and its parameters. [4]

Applying Bayesian methods to DNNs has been a widely successful solution to quantify the uncertainty of the model predictions [5] [6] [7]. Utilizing Bayesian methods in uncertainty quantification relies on approximating the posterior predictive distribution of the measure of interest [4]. The posterior predictive distribution is approximated since the analytical form is intractable. Variational inference (VI) is a popular method in performing the necessary tractable approximations, especially in Bayesian neural networks (BNNs) [8] [9]. From the estimated posterior predictive distribution of interest, it is straightforward to calculate the predictive uncertainty. Bayesian methods are applied to neural networks by replacing some of the deterministic layers with their Bayesian counterparts in order to observe the epistemic uncertainty present in the network parameters. In Bayesian layers the layer weights are not just some scalars or vectors, but rather it is assumed that the weights follow some (usually Gaussian) probability distribution. Thus, in BNNs, instead of directly optimizing the weights of the Bayesian layers we aim to optimize the parameters of the distributions they are sampled from.

In this paper, we train two deterministic state-of-art DL classifiers to perform binary classification for diabetic retinopathy and chest X-ray pneumonia datasets. We introduce some of the state-of-art adversarial attacks, apply them to perturb the two datasets to demonstrate how this leads to misclassifications in the predictions of the two deterministic models. Afterwards, we add Bayesian layer to both of our deterministic models in order to quantify the uncertainty of the predictions made on the adversarial samples. We compare the performance between the deterministic and the Bayesian models against the adversarial attacks. Our main goal is to see whether the addition of the uncertainty quantification yields more robust models against the adversarial attacks in the field of medical imaging.

## 2 Review of related literature

In this section we cover some of the relevant literature related to the research topic of adversarial attacks, DL based medical imaging systems and uncertainty quantification. These papers together serve as the motivation and inspiration for conducting our study.

Apostolidis et al. (2021) [10] have put together a comprehensive literature survey on robustness of DL based systems in medical image analysis. In their paper they summarize thoroughly the existing methods for adversarial attacks, their detection methods as well as some defense methods that have been developed by the research community. Finlayson et al. (2019) [12] focus on emphasizing practically that the medical imaging systems can be harmed by the adversarial attacks. They build their own representatives of the current state-of-art medical diagnosis DL systems which reflect the architectures already actively deployed into real world clinical processes. Furthermore, they demonstrate that the outputs of these representative systems can be efficiently manipulated by the use of adversarial attacks. In addition to these technical contributions, they provide some insights and concrete hypotheses on the questions of how and why adversarial attacks could cause harm against healthcare systems as a whole.

Ma et al. (2019) [11] put more emphasis on comparing the effectiveness of adversarial attacks on medical imaging DL models and DL models for natural images. They find out that medical DL models are more vulnerable against adversarial attacks. This is partly explained by the rich biological textures of medical images which sometimes lead to distracting the DNN model into paying too much attention to image regions unrelated to diagnosis classification. Consequently, this enables any adversarial attack method to target larger or more regions. Another explanation they give is that the medical diagnosis DNNs are often simply overparameterized, especially for simple medical diagnosis tasks. Surprisingly, it turns out in their study that adversarial attacks against medical images are also easier to detect. The deep feature based detection method introduced in their paper reached over 98% detection rate AUC against some of the state-of-art adversarial attacks. This was explained by the fact that the

features in perturbed images created by adversarial attacks are linearly separable form original features.

Scholars have put some effort on development of adversarial attacks since their discovery to be able to better understand them. The most simple and common adversarial attacks are the firstly proposed Fast Gradient Sign Method introduced by Goodfellow et al. (2015) [13] including its iterative extensions Basic Iterative Method introduced by Kurakin et al. (2017) [14] and Projected Gradient Descent introduced by Madry et al. (2019) [15]. Defensive distillation was one proposed technique to increase the DL model's robustness against some simple adversarial attacks. However, this technique was quickly invalidated by the highly effective and computationally intensive adversarial attack introduced by Carlini et al. (2016) [16]. Hirano et al. (2020) [21] focused on investigating more realistic Universal Adversarial Perturbation attack method introduced by Moosavi-Dezfooli et al. (2017) [20]. Their results suggest that adversarial retraining which is widely considered as an effective adversarial defense method against adversarial attacks is only efficient to certain extent by increasing the robustness of medical imaging systems only in limited cases. Lastly, all of the previously mentioned adversarial attacks require information about the model architecture and parameters which in most real world scenarios would not be available. Therefore, Uesato et al. (2018) [17] provide an alternative perturbation method to maliciously modify the input images without knowing any details about the classifier it is attacking.

Apostolidis et al. [10] enlist some of the current mitigation methods against the adversarial attacks: In addition to the previously mentioned adversarial pretraining and defensive distillation, methods like randomization, denoising, weight-sparse DNNs, detection-rejection methods, KNN-based defenses, consistency-based defenses and Bayesian inference-based defenses are mentioned. Furthermore, Kompa et al. (2021) [4] propose a new kind of approach to address the problem of adversarial attacks as an obstacle for fully integrating DL systems to clinical processes. They suggest that the possibility to quantify and communicate uncertainties of the individual predictions would increase the transparency and trust of the healthcare professionals when it comes to novel medical DL systems. Additionally, they provide a survey on different methods for quantifying predictive uncertainty, among which the BNNs – neural networks reinforced with Bayesian methods – are mentioned. Meanwhile, in the domain of natural image classification, Liu et al. (2018) [24] had already reached some promising results by increasing robustness against adversarial attacks through the use of BNN architecture.

Abdar et al. (2021) [22] published an extensive review on various uncertainty quantification techniques and applications for DL systems. According to their study, uncertainty quantification methods have already been studied for medical image analysis tasks – at least to some extent. For instance, Araújo et al. (2020) [23] designed a novel DL system to support its individual diagnoses with interpretable estimation of how uncertain it is about a given prediction.

An interesting study that combines Bayesian methods, DNNs and potential to increase adversarial robustness against a natural image dataset was conducted by Alemi et al. (2016) [25]. In their paper, they introduce a method called Deep Variational Information Bottleneck (DVIB), which is a variational approach to approximate the posterior distributions of the model weights. In their approach, combine a DL model with the VIB as the its final layer to demonstrate to be able to outperform deterministic models of similar architecture in robustness against a state-of-art adversary. Particularly, trying whether their approach improves the adversarial robustness of the state-of-art representatives of medical DL diagnosis models, acts as the main motivation for our study.

# 3 Deterministic models and data

In this section we introduce the two datasets and the deterministic baseline models used in our experiments. We explain the retrieval and preprocessing of the data, and the delve deeper in the details of the training process of deterministic models to perform classification on the data. Finally, we evaluate the diagnosis performance of the trained models on the two medical image datasets.

### 3.1 Datasets

We chose two datasets for our experiments: Diabetic retinopathy (DR) multi-label classification dataset from [26] and chest X-ray (CXR) pneumonia binary classification dataset from [27]. The two main reasons for selecting these datasets was the public availability and the successful usage in another medical DL experiments [12].

The downloaded DR dataset contained 3662 training samples and 1928 test samples. We didn't need that many images for our experiments so we only used the training images as our dataset. Rather than predicting the grade of the DR, we simplified our classification tasks by relabeling the data to positive examples, if DR grade was greater 0, and to negative examples otherwise. The proportion of the positive examples in the new preprocessed dataset is 50.71%, so minimal imbalance is present with this dataset.

The CXR dataset was downloaded as three separate training, test and validation sets. We combined these datasets into one, since the premade split didn't properly take into account the class imbalance present in the dataset. The proportion of positive samples is 72.97%, and we wanted our training, test and validation sets to have similar class imbalances.

Finally, both of the datasets were split into three: Training set containing 80% of the data, validation set containing 5% of the data and test set containing 15% of the data. As stated earlier, the split was done in a way that the class imbalances between the three datasets were similar. The data was always normalized before feeding as input to the models. The training set was shuffled and divided into batches of 32 samples. The test and validation sets were divided into batches of 5 samples, and no shuffling was done.

### 3.2 Our approach: InceptionV3 and ResNet50

The model architectures were also selected based on the experiments of Finlayson et al. (2019) [12] where they use DL models which – according to them – reflect the state-of-art representatives of clinical DL systems. In their publicly available code [28] they experiment with two different ImageNet architectures: InceptionV3 (GoogleNetv3) and ResNet50.

In our experiments, software package PyTorch [29] was utilized to train these two model architectures to perform classification on the medical image datasets. The images were resized to 299 x 299 pixels for InceptionV3 and 224 x 224 pixels for ResNet50, according to the suggestions given in their documentations. Since class imbalance was present especially in the CXR dataset, we trained two versions of both of the models too see whether having a weighted loss function had any effect on their performance. Thus, in total we trained four models for each dataset. The four models were trained with 200 epochs by optimizing Cross Entropy (CE) loss function via Stochastic Gradient Descent (SGD) optimizer with momentum of 0.9 and learning rate of $1^{-4}$. On each epoch, the CE loss was additionally computed on the validation set, and the model architecture with the lowest validation loss was selected as the deterministic baseline model.

### 3.3 Results: Medical image diagnosis with deterministic models

The evaluation of the deterministic baseline models was carried out on the test sets of the both datasets, separately. Mainly, we wanted to see whether high diagnosis performance on medical images could be achieved, and whether using weighted loss functions in the training proved useful or not. The results and their compact interpretations on the two datasets are presented in tables 1 and 2 of Appendix A.

## 4 Adversarial attacks

In this section we focus on the adversarial attacks. Firstly, we go through the theoretical aspects behind the adversarial attack methods used in our experiments. Thereafter, we present our own experiments on launching the attacks on the two medical image datasets. Lastly, we

demonstrate that the adversarially perturbed images are misclassified with high confidence by the deterministic DL models introduced in section 3.

## 4.1 Theoretical foundations behind adversarial attacks

In this more detailed subsection, we cover four common state-of-art methods used for perturbing the images. The comparison of the methods is provided by delving deeper into the mathematics behind them. Additionally, some discussion about the effect of the hyperparameter selection on the perturbation intensity is provided. The attacks can be categorized to white- and black-box attacks based on whether the gradients of the model are accessible when applying the adversarial method. Furthermore, another possible categorization for adversarial attacks are targeted and untargeted attacks. In the former the goal is to cause misclassification to a predetermined class and in the latter the goal is only to cause any kind of misclassification.

### 4.1.1 Fast Gradient Sign Method

Fast Gradient Sign Method (FGSM) is a non-iterative white-box perturbation method. The main idea behind the method is to generate adversarial examples based on the original images by changing each pixel by a certain magnitude $\epsilon$ towards the direction of the gradient of its loss function. This is executed by maximizing the loss of the model by modifying data, since the weights of the model are already fixed to minimize the loss. As long as the added noise by magnitude $\epsilon$ is small enough, the changes in the resulting image are undetectable by human eye. Hence, the optimal magnitude hyperparameter $\epsilon$ in order to cause maximal perturbation while keeping the perturbations undetected to human eye would be as small as possible while causing the model to misclassify with high enough confidence. Another hyperparameter that needs to be decided is the norm that constrains the perturbation. Most commonly used norm is the max norm $||\cdot||_{inf}$ since it is robust to datasets with high dimensionality. The adversarial examples $\boldsymbol{X_{adv}}$ using FGSM attack can be generated as

$$\boldsymbol{X_{adv}} = \boldsymbol{X} + \epsilon \cdot sign(\nabla_{\boldsymbol{X}} J(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{y})), \tag{1}$$

where $\boldsymbol{X}$ is the clean input image, $\boldsymbol{\theta}$ are the model weights, $\boldsymbol{y}$ are the true class labels, $\epsilon$ is the maximum magnitude of the max norm of the perturbation and $sign(\nabla_{\boldsymbol{X}} J(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{y}))$ is the sign (+/-) of the gradient of the loss function w.r.t. input image. [13]

### 4.1.2 Projected Gradient Descent

Projected Gradient Descent (PGD) is one of the iterative extensions of FGSM. It is basically a generalized version of the other iterative extension of FGSM called the Basic Iterative Method (BIM) [14]. Essentially, the difference is that in PGD the initial point for iteration process is decided randomly within the $\ell_{inf}$-bounded ball around the pixel. The hyperparameter selection centers around the number of iteration steps $T$ and the iteration perturbation magnitude $\alpha$. By increasing either $T$ or $\alpha$, the misclassification confidence of the model increases. The iterative process applied to every pixel of the image to achieve the PGD adversial image is

$$\boldsymbol{X_t} = \Pi_{\boldsymbol{X_{t-1}}+\mathcal{S}}(\boldsymbol{X_t} + \alpha \cdot sign(\nabla_{\boldsymbol{X}} J(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{y}))), \tag{2}$$

where $t = 1, 2, ..., T$, $\boldsymbol{X}_0$ is the clean input image, $\mathcal{S}$ is a set of allowed perturbations bounded typically by $\ell_{inf}$-ball, $\alpha$ is the step size for one attack iteration and $\Pi_{\boldsymbol{X}+\mathcal{S}}(\cdot)$ is the projecting function. [15]

#### 4.1.3 Carlini & Wagner

The Carlini & Wagner (CW) method is a highly effective white-box adversarial attack method. It was demonstrated to be able to produce adversarial examples that caused models trained with defensive distillation technique to misclassify with high confidence. The main objective of CW is to minimize the norm between the original image $\boldsymbol{x}$ and the perturbed image $\boldsymbol{x} + \boldsymbol{\delta}$. There are three options for the distance metric selection: $\ell_0$, $\ell_2$ and $\ell_{inf}$. We will use distance metric $\ell_2$ later, so we will focus on that version of the CW attack here. The formula for creating CW perturbed images with $\ell_2$-norm is

$$\min ||\frac{1}{2}(tanh(\boldsymbol{w}) + 1) - \boldsymbol{x}||_2^2 + c \cdot f(\frac{1}{2}(tanh(\boldsymbol{w}) + 1)), \tag{3}$$

where $c$ is the suitability chosen constant decided with method described in [16] and $\boldsymbol{w}$ is just a reparametrization technique to smooth the clipped gradient descent by setting $\delta_i = \frac{1}{2}(tanh(w_i)+1) - x_i$ such that instead of optimizing over $\boldsymbol{\delta}$ we optimize over $tanh(\boldsymbol{w})$ resulting in valid solution where $0 \leq x_i + \delta_i \leq 1$. The function f in equation 3 is defined as

$$f(x) = \max(\max Z(\boldsymbol{x})_i : i \neq t - Z(\boldsymbol{x})_t, -\kappa), \tag{4}$$

where $Z(\boldsymbol{x})$ are the logits of the model output, $t$ is the target class, and $\kappa$ is the parameter with which the confidence of the misclassification can be adjusted. [16]

#### 4.1.4 Simultaneous Perturbation Stochastic Approximation

Simultaneous Perturbation Stochastic Approximation (SPSA) adversary is an optimization method particularly useful when the gradients of the model do not point in any useful directions – so in other words, when the model is non-differentiable. This also makes SPSA a black-box attack which doesn't require any information about the model it is attacking against. One of the main motivations for developing SPSA adversary was to have a truly transparent model to help in designing adversarial defenses. For example, the other adversaries (such as PGD) are local optimization techniques which only allow the computation of the approximation of the lower bounds on the true adversarial risk, while the transparent models could theoretically reach the true risk. The gradient-free optimization of SPSA is based on approximations of the difference estimates in random directions. It is demonstrated in [19] that if the conditions are set appropriately, the introduced stochasticity by sampling the perturbations allows SPSA to converge to the global minimum. [17] [18]

The inputs to iterative process of the SPSA adversary are the loss function $f$ to be minimized, the initial image $x_0 \in \mathbb{R}^D$, perturbation magnitude $\delta$, perturbation step size $\alpha > 0$, number of iterations $T$ and batch size $n$. As stated earlier, knowledge about the classifier to be attacked is not needed by SPSA.

**Algorithm 1** SPSA adversary

**for** $t = 1, 2, \ldots, T$ **do**
  Sample $v_1, \ldots, v_n \sim \{-1, 1\}^D$ (Rademacher distribution)
  Define $v_i^{-1} = [v_{i,1}^{-1}, \ldots, v_{i,D}^{-1}]$
  Derive $g_i = \frac{v_i^{-1}}{2\delta}(f(x_{t-1} + \delta v_i) - f(x - \delta v_i)))$, where $x$ is a random perturbation within set $N_{\ell_{\text{inf}}}(x_0)$ which is bounded by $\ell_{\text{inf}}$-ball around $x_0$
  Set $x'_{t-1} = x_{t-1} - \frac{\alpha}{n} \sum_{i=1}^{n} g_i$
  Project $x_t = \arg\min_{x \in N_{\ell_{\text{inf}}}(x_0)} ||x'_{t-1} - x_0||$
**end for**

### 4.2 Our approach: Adversarial attacks via CleverHans

In our experiments, the adversarial attacks were implemented by using an open source software package called CleverHans [30]. CleverHans is developed to benchmark the vulnerability of the DL systems to adversarial examples, and help research community on the journey of developing DL systems that are adversarially robust. Conveniently for us, CleverHans supports the DL framework PyTorch used in our model training. Hence, launching the adversarial attacks on the medical image datasets by utilizing CleverHans was a natural choice.

We experimented with the four adversarial attack methods. The mathematical details of these methods are covered thoroughly in section 4.1. For FGSM and PGD, we fixed the hyperparameter $\epsilon$ to 0.02, since this was also used in the experiments of Finlayson et al. [12]. Furthermore, the required hyperparameter of allowed step size per attack iteration in the PGD attack, was set to 0.01. For PGD, we experimented with 1, 2 and 5 number of iterations. $\ell_{inf}$-norm is used in FGSM, PGD and SPSA. In CW, $\ell_2$-norm is used for convenience, since CleverHans had only this version of the attack available in their open source library. For SPSA, we fixed the number of iterations to 5, since increasing it further did not have too much effect on the ability to degrade the model accuracy, but significantly increased the time it took to execute the attack. On the other hand, with SPSA, we experimented with set of $\{0.02, 0.05, 0.1, 0.5, 1\}$ when it comes to hyperparameter $\epsilon$. Interestingly the images perturbed with SPSA, already upwards from $\epsilon = 0.1$ the differences between the perturbed and original images become clearly distinguishable to human eye.

### 4.3 Results: Adversarial attacks versus deterministic models

We performed the evaluation of the deterministic model robustness by launching four different adversarial attacks methods on the test sets of both medical image datasets. Most importantly, we wanted to convince ourselves that these attacks degrade the classification accuracy of both models and see in practice, how different hyperparameter selections for adversarial attacks affect the results. The respective results for these experiments on the two datasets can be read from tables 3 and 4 of Appendix B. Short interpretation of the results is provided in the table captions.

## 5 Bayesian Neural Networks

In this section we cover the concept of Bayesian neural networks (BNNs). To begin with, theory behind BNNs is introduced in order to better understand the benefits of embedding Bayesian inference methods to DL systems. Afterwards, we present our approach of implementing Bayesian extensions of the deterministic models presented in section 3. Finally, we launch adversarial attacks against our Bayesian DL models, and see whether the adversarial robustness increases.

### 5.1 Theoretical foundations behind Bayesian neural networks

An extensive introductory survey of BNNs was conducted by Goan et al. (2020) [31], and hence, most of this section is based on their paper, enhanced with fragments from the paper by Abdar et al. [22]. Despite the success of deterministic neural networks (NNs), their frequentist setting withholds the possibility to reason about the uncertainty in their predictions. Combining Bayesian statistics in NNs has been a widely popular solution to address this issue, since they allow natural reasoning about the uncertainty in predictions, and can provide transparency on how the model ended up with these predictions. They key distinction between BNNs and deterministic NNs is that in BNNs the model weights $\boldsymbol{w}$ are treated as latent random variables. Instead of trying to optimize them to some fixed values, training process of BNNs strives to learn the distribution over $\boldsymbol{w}$ conditioned on what we can observe in the data $\mathcal{D}$. This distribution is called the posterior distribution $p(\boldsymbol{w}|\mathcal{D})$, and is defined by

$$p(\boldsymbol{w}|\mathcal{D}) = p(\boldsymbol{w})p(\mathcal{D}|\boldsymbol{w}). \tag{5}$$

Thus, in order to compute the posterior we must define the prior $p(\boldsymbol{w})$ and likelihood $p(\mathcal{D}|\boldsymbol{w})$. Our prior should reflect our prior beliefs of the distribution of $\boldsymbol{w}$ before seeing any data. Since the NN weights typically are low magnitude values centered around zero it is a common approach to use zero mean Gaussian with small variance as prior. Our likelihood must be defined depending on the model architecture and loss function. Theoretically, after defining the prior and likelihood, applying Bayes' theorem results in the posterior $p(\boldsymbol{w}|\mathcal{D})$. In most cases, performing this step requires computing of the marginal likelihood $p(\mathcal{D})$, which is analytically intractable for models with non-linear latent variables or non-conjugate prior for likelihood. Neural networks represent the former and hence, only an approximation of the true posterior can be computed. This representation of $p(\boldsymbol{w}|\mathcal{D})$ is defined by

$$p^*(\boldsymbol{w}|\mathcal{D}) = \frac{p(\boldsymbol{w})p(\mathcal{D}|\boldsymbol{w})}{p(\mathcal{D})}. \tag{6}$$

This approximation of the posterior $p^*(\boldsymbol{w}|\mathcal{D})$, allows us to extract various quantities of interest behind the model outputs as expectations over the posterior:

$$\mathbb{E}_{p^*}[f] = \int f(\boldsymbol{w})p^*(\boldsymbol{w}|\mathcal{D})d\boldsymbol{w}. \tag{7}$$

Depending on the quantity of interest, only the function $f(\boldsymbol{w})$ needs to be changed. The integration over $\boldsymbol{w}$ is generally referred as marginalization which enables us to interpret the generative process of a model, by forming representations of predictions as conditional probabilities. Thus, also uncertainty quantification of the outputs can be evaluated based on the quantities derived via marginalization. For instance, the predictive uncertainty can be assessed by computing the predictive variance.

Multiple methods to approximate $p^*(\boldsymbol{w}|\mathcal{D})$ have been presented by the research community over the years. We do not cover them all in this paper, but for those interested Abdar et al. provide an extensive review on them in [22]. Regardless, Variational Inference (VI) needs to be covered, since modern studies about BNNs focus on this approach. VI is an approximation method which allows framing the marginalization to a tractable optimization problem to learn the posterior over $\boldsymbol{w}$. This is especially convenient in case of DL, since VI problem can be optimized using similar backpropagation approach which is also used in the training process of deterministic NNs. For a single mini-batch $\mathcal{D}_i \subset \mathcal{D}$ of size $N$, the loss to be optimized in VI backpropagation is defined by

$$\mathcal{L}_{VI}(\boldsymbol{w}, \boldsymbol{\theta}) = -\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_q[\log\big(p(\mathcal{D}_i|\boldsymbol{w})\big)] + KL\big(q_{\boldsymbol{\theta}}(\boldsymbol{w})||p(\boldsymbol{w})\big). \tag{8}$$

Here, the first term is the mean negative log-likelihood and the second term is the Kullback-Leibler-divergence, which can be interpreted as a measure of dissimilarity between the approximated variational distribution $q_{\boldsymbol{\theta}}(\boldsymbol{w})$ and the prior $p(\boldsymbol{w})$. The variational parameters $\boldsymbol{\theta}$ are updated using backpropagation to reduce the dissimilarity between variational distribution and true posterior. The key assumption to simplify the VI optimization process into tractable form, is to restrict the variational distribution to a certain family of distributions with $\boldsymbol{\theta}$. Lastly, to prove that the combination of the two terms in the VI loss indeed measures the dissimilarity between variational distribution $q_{\boldsymbol{\theta}}(\boldsymbol{w})$ and true posterior $p(\boldsymbol{w}|\mathcal{D})$, and that the intractable

marginal log-likelihood term can be dropped from the equation when optimizing w.r.t $\boldsymbol{\theta}$, we convince ourselves with the following derivations:

$$\begin{aligned} KL\big(q_{\boldsymbol{\theta}}(\boldsymbol{w})||p(\boldsymbol{w}|\mathcal{D})\big) &= \int q_{\boldsymbol{\theta}}(\boldsymbol{w}) \log \frac{q_{\boldsymbol{\theta}}(\boldsymbol{w})}{p(\boldsymbol{w}|\mathcal{D})} \\ &= \mathbb{E}_q\big[\log \frac{q_{\boldsymbol{\theta}}(\boldsymbol{w})}{p(\boldsymbol{w})} - \log p(\mathcal{D}|\boldsymbol{w})\big] + \underbrace{\log p(\mathcal{D})}_{\text{independent w.r.t. } \boldsymbol{\theta}} \\ &= -\mathbb{E}_q[\log\big(p(\mathcal{D}|\boldsymbol{w})\big)] + KL\big(q_{\boldsymbol{\theta}}(\boldsymbol{w})||p(\boldsymbol{w})\big). \end{aligned} \tag{9}$$

## 5.2 Our approach: Deep Variational Information Bottleneck

Our approach on embedding Bayesian inference to the medical DL diagnosis models introduced in section 3, was mainly inspired by DVIB method introduced by Alemi et al. [25]. In their approach they demonstrated that by replacing the final layer of an adversarially vulnerable DL classifier with a VIB, and by retraining, robustness of the model against adversarial attacks can be improved. They achieved promising results against $\ell_2$-norm version of the CW attack launched on MNIST dataset with model architecture of pretrained InceptionV2 embedded with VIB. Thus, we wanted to see whether similar improvements in adversarial robustness could be reached by replacing the last layer of our deterministic medical DNNs with VIB. Additionally, even though the CW attack is widely considered as the benchmark adversarial perturbation method which produces highly effective adversarial examples, we also wanted to test our VIB models against the weaker attacks.

Before continuing to detailed description of our VIB training, we must explain the main difference between VIB loss and standard VI loss covered in section 5.1. In VIB loss we control the trade-off between the negative log-likelihood term and the KL-divergence term with a Lagrange multiplier $\beta \geq 0$. Introducing $\beta$ allows us to control the trade-off between predictive power and dependency on training data. More intuitively, the first term encourages to be accurate in our classifications, and the second term encourages us to care about training data as little as possible. Hence, the VIB objective is defined by

$$\mathcal{L}_{VIB}(\boldsymbol{w}, \boldsymbol{\theta}) = -\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_q[\log\big(p(\mathcal{D}_i|\boldsymbol{w})\big)] + \beta KL\big(q_{\boldsymbol{\theta}}(\boldsymbol{w})||p(\boldsymbol{w})\big). \tag{10}$$

Since the deterministic models with weighted CE loss proved themselves better in sections 3 and 4, we continued to VIB training only with these versions of deterministic baseline models. The training of a single VIB involved replacing the last fully connected layer of a baseline model with the VIB. The weights of the VIB were initialized using the Xavier initialization scheme, and the biases were initialized to zero. For the encoder part of the VIB, 256 dimensional Gaussian embeddings were used to create latent distribution parameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$. Additionally for the $\boldsymbol{\sigma}$ values, ReLu activation and Softplus transformation with bias of $-5$ were applied in the respective order. This additional procedure was executed in order to avoid computational issues. After this, the latent samples were sampled from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\boldsymbol{I})$ and forwarded through the linear decoder layer to produce the output logits. Then the weights of the VIB were updated by backpropagating through the VIB loss. The log-likelihood term of the VIB loss was set as weighted CE loss because of the classification setting, and a weakly informative prior $p(\boldsymbol{w}) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ was selected. For both deterministic model architectures, we trained 10 variants of these VIBs with differing $\beta$ values ranging from $1^{-10}$ to $1^{-1}$. We used 20 epochs for training of a single VIB, since the training loss converged quickly. Adam optimizer with initial learning rate of $10^{-4}$ and hyperparameters ($\beta_1 = 0.5, \beta_2 = 0.999$) was used. Furthermore, an exponential weight decay with $\gamma = 0.97$ was applied every 2 epochs.

### 5.3 Results: Adversarial attacks versus Bayesian models

In order to see whether embedding VIB layer to our baseline models had an effect on the adversarial robustness, we launched similar experiments as in section 4.3 on each VIB model trained with different $\beta$. When evaluating the VIB models, the mean output derived from 12 different latent samples was used as the model output. Results on both datasets are presented in the tables 5, 6, 7 and 8 of Appendix C. Further details and interpretations of the results are written in the captions of the respective tables.

## 6 Concluding remarks

High classification performance of DL models in CV tasks has raised optimism to automate parts of medical imaging processes. This has a lot of potential to decrease the cost and increase the efficiency in these processes. Unfortunately, the recent discovery shows that it is possible to adversarially perturb input images of a DL system, with an intention to cause the system to misclassify with high confidence. Naturally, this has raised concerns within the healthcare community, insurance companies and ordinary patients.

In this paper, we approached the issue of increasing the robustness of medical DL model against adversarial attacks by quantifying the uncertainty behind the model predictions. This was done by embedding Bayesian inference, which has proven itself to be an efficient way to provide natural insights about the uncertainties of the model predictions, to the DL models. First, we trained two state-of-art representatives of the deep medical imaging models, to perform binary classification on two medical disease image datasets. Thereafter, we launched an array of common adversarial attacks on the datasets, resulting in severe degradation of the classification performance of the models. Then, inspired by Alemi et al. [25] we replaced the final layers of these models with VIB variants, and trained these VIB variants separately on the same datasets. Lastly, we evaluated the DVIB models by launching the same adversarial attacks against them, in order to assess whether the the adversarial robustness increased in medical setting.

The biggest improvements in adversarial robustness were reached against the highly effective CW attack, which was expected based on the results of Alemi et al. [25]. Additionally, we tested the VIB method against three other common adversarial attacks, and although some slight improvements in robustness against these were seen, the results are at most encouraging. Hence, embedding VIB to the medical DL model can be seen as a step into right direction, but it must be noted that we are still far from achieving high enough adversarial robustness to be able to safely incorporate these DL systems in practical high-stake decision making processes such as clinical diagnosis. There are multiple further study topics regarding the issue of increasing adversarial robustness of medical DL classifiers. To enlist a few, more emphasis could be put to optimize $\beta$, VIB could be experimented with other model architectures, and experiments on combining state-of-art adversarial detection or defense methods (e.g. adversarial training) with VIB could be conducted.

# References


[1] Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Heal. 2019;1:e271–97. doi:10.1016/S2589-7500(19)30123-2. https://www.researchgate.net/publication/336068436_A_comparison_of_deep_learning_performance_against_health-care_professionals_in_detecting_diseases_from_medical_imaging_a_systematic_review_and_meta-analysis

[2] 2018. Press Announcements ¿ FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems. https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm604357.htm.

[3] Kathrin Grosse, Nicolas Papernot, Praveen Manoharan, Michael Backes, and Patrick McDaniel. 2017. Adversarial examples for malware detection. In European Symposium on Research in Computer Security. Springer, 62–79. https://arxiv.org/abs/1610.04256.

[4] Kompa B, Snoek J, Beam A. Second opinion needed: communicating uncertainty in medical machine learning. Digital Medicine (2021) 4:4 ; https://doi.org/10.1038/s41746-020-00367-3.

[5] Olivier A, Shields M, Graham-Brady L. Bayesian neural networks for uncertainty quantification in data-driven materials modeling. https://doi.org/10.1016/j.cma.2021.114079.

[6] Mosser L, Naeini E. Calibration and Uncertainty Quantification of Bayesian Convolutional Neural Networks for Geophysical Applications. https://arxiv.org/abs/2105.12115.

[7] Kwon Y, Won J, Kim B, Paik M. Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation. https://doi.org/10.1016/j.csda.2019.106816.

[8] Blundell C, Cornebise J, Kavukcuoglu K, Wiersta D. Weight Uncertainty in Neural Networks. https://arxiv.org/abs/1505.05424.

[9] Graves A. Practical Variational Inference for Neural Networks. In Advances in Neural Information Processing Systems 24 (eds. Shawe-Taylor J, Zemel R, Bartlett P, Pereira F, Weinberger K) 2348–2356 (Curran Associates, Inc., 2011).

[10] Apostolidis K, Papakostas G. A Survey on Adversarial Deep Learning Robustness in Medical Image Analysis. 2021. https://doi.org/10.3390/electronics10172132

[11] Xingjun Ma, Yuhao Niu, Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, Feng Lu. Understanding Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems. 2019. https://doi.org/10.1016/j.patcog.2020.107332.

[12] Samuel G. Finlayson, Hyung Won Chung, Isaac S. Kohane, Andrew L. Beam. Adversarial Attacks Against Medical Deep Learning Systems. 2019. https://arxiv.org/abs/1804.05296.

[13] Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy. 2015. Explaining and harnessing adversarial examples. https://arxiv.org/pdf/1412.6572.pdf.

[14] Alexey Kurakin, Ian J. Goodfellow & Samy Bengio. 2017. Adversarial examples in the physical world. https://arxiv.org/pdf/1607.02533.pdf.

[15] Alexander Madry, Aleksandar Makelov & Ludwig Schmidt. 2019. Towards Deep Learning Models Resistant to Adversarial Attacks. https://arxiv.org/pdf/1706.06083.pdf.

[16] Nicholas Carlini & David Wagner. 2017. Towards Evaluating Robustness of Neural Networks. https://arxiv.org/pdf/1608.04644.pdf.

[17] Jonathan Uesato, Brendan O'Donoghue, Aaron van der Oord & Pushmeet Kohli. 2018. Adversarial Risk and the Dangers of Evaluating Against Weak Attacks. https://arxiv.org/pdf/1802.05666.pdf.

[18] James Spall. 1992. Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.19.4562&rep=rep1&type=pdf.

[19] John Maryak & Daniel Chin. 2001. Global Random Optimization by Simultaneous Perturbation Stochastic Approximation. https://www.jhuapl.edu/spsa/PDF-SPSA/Maryak_Global_Random.pdf

[20] Moosavi-Dezfooli, S.-M.; Fawzi, A.; Fawzi, O.; Frossard, P. Universal Adversarial Perturbations. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 86–94.

[21] Hokuto Hirano, Akinori Minagi, Kazuhiro Takemoto. Universal adversarial attacks on deep neural networks for medical image classification. 2020. https://doi.org/10.21203/rs.3.rs-70727/v1.

[22] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. 2021. https://doi.org/10.1016/j.inffus.2021.05.008

[23] Teresa Araújo, Guilherme Aresta, Luís Mendonça, Susana Penas, Carolina Maia, Ângela Carneiro, Ana Maria Mendonça, Aurélio Campilho. DR—GRADUATE: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images. 2020. https://doi.org/10.1016/j.media.2020.101715.

[24] Xuanqing Liu, Yao Li, Chongruo Wu, Cho-Jui Hsieh. Adv-BNN: Improved Adversarial Defense through Robust Bayesian Neural Network. https://arxiv.org/abs/1810.01279.

[25] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, Kevin Murphy. Deep Variational Information Bottleneck. 2016. https://arxiv.org/abs/1612.00410.

[26] APTOS 2019 Blindness Detection. https://www.kaggle.com/c/aptos2019-blindness-detection/data.

[27] Daniel Kermany, Kang Zhang, Michael Goldbaum. Labeled optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification. Mendeley Data, v2. http://dx.doi.org/10.17632/rscbjbr9sj.2. https://www.kaggle.com/aviadl/starter-chest-x-ray-images-pneumonia-d498de7c-3/data.

[28] Publicly available code for medical DL imaging system experiments of Finlayson et al. https://github.com/sgfin/adversarial-medicine.

[29] Paszke, A. et al., 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Advances in Neural Information Processing Systems 32. Curran Associates, Inc., pp. 8024–8035. Available at: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[30] Nicolas Papernot and Fartash Faghri and Nicholas Carlini and Ian Goodfellow and Reuben Feinman and Alexey Kurakin and Cihang Xie and Yash Sharma and Tom Brown and Aurko Roy and Alexander Matyasko and Vahid Behzadan and Karen Hambardzumyan and Zhishuai Zhang and Yi-Lin Juang and Zhi Li and Ryan Sheatsley and Abhibhav Garg and Jonathan Uesato and Willi Gierke and Yinpeng Dong and David Berthelot and Paul Hendricks and Jonas Rauber and Rujun Long. Technical Report on the CleverHans v2.1.0 Adversarial Examples Library. https://arxiv.org/abs/1610.00768.

[31] Ethan Goan, Clinton Fookes. Bayesian Neural Networks: An Introduction and Survey. 2020. https://arxiv.org/abs/2006.12024

# Appendix A: Medical image diagnosis with deterministic DL models

| model | accuracy | precision | recall | F1-score | AUC | TP | FP | FN | TN |
|---|---|---|---|---|---|---|---|---|---|
| InceptionV3* | 94.76% | .96 | .96 | .96 | .99 | 618 | 23 | 23 | 214 |
| InceptionV3 | 94.65% | .96 | .97 | .96 | .99 | 620 | 26 | 21 | 211 |
| ResNet50* | 96.01% | .96 | .98 | .97 | .98 | 628 | 23 | 13 | 214 |
| ResNet50 | 94.19% | .94 | .98 | .96 | .98 | 631 | 41 | 10 | 196 |

Table 1: Performance evaluation metrics of the four trained models on the CXR Pneumonia dataset. The models trained with weighted CE loss are marked with (*). As expected, the models trained with weighted CE loss outperformed their counterparts, as serious class imbalance was present in the CXR Pneumonia dataset. There isn't significant difference between the performances of the two model architectures. Overall, the performance of the models is respectable.

| model | accuracy | precision | recall | F1-score | AUC | TP | FP | FN | TN |
|---|---|---|---|---|---|---|---|---|---|
| InceptionV3* | 94.00% | .97 | .91 | .94 | .98 | 255 | 9 | 24 | 262 |
| InceptionV3 | 95.09% | .97 | .93 | .95 | .98 | 260 | 8 | 19 | 263 |
| ResNet50* | 93.09% | .94 | .93 | .93 | .98 | 259 | 18 | 20 | 253 |
| ResNet50 | 92.00% | .92 | .92 | .92 | .98 | 257 | 22 | 22 | 249 |

Table 2: Performance evaluation metrics of the four trained models on the DR dataset. Again, the models trained with weighted CE loss are marked with (*). Here we don't see any significant difference between two versions of the same model architectures trained with weighted and non-weighted CE loss. Of course, there is almost zero class imbalance present with the DR dataset. On the other hand, InceptionV3 seems to slightly outperform ResNet50 on the DR dataset as an architecture. Overall, the performance of the models is good.

# Appendix B: Evaluation of deterministic model robustness against adversarial attacks

| | InceptionV3* % | InceptionV3 % | ResNet50* % | ResNet50 % |
|---|---|---|---|---|
| No attack | 95 | 95 | 96 | 94 |
| FGSM | 62 | 51 | 10 | 14 |
| CW | 5 | 5 | 4 | 6 |
| PGD1 | 53 | 49 | 8 | 12 |
| PGD2 | 26 | 22 | 4 | 6 |
| PGD5 | 9 | 8 | 4 | 6 |
| SPSA5, $\epsilon = 0.02$ | 95 | 95 | 96 | 94 |
| SPSA5, $\epsilon = 0.05$ | 95 | 93 | 95 | 94 |
| SPSA5, $\epsilon = 0.1$ | 92 | 76 | 95 | 93 |
| SPSA5, $\epsilon = 0.5$ | 27 | 56 | 80 | 79 |
| SPSA5, $\epsilon = 1.0$ | 27 | 27 | 58 | 27 |

Table 3: Evaluating adversarial robustness of the deterministic models on the CXR Pneumonia dataset. The models trained with weighted CE loss are marked with (*). Adversarial attacks degrade the classification accuracy of the deterministic models significantly. InceptionV3 model trained with weighted CE loss seems to be the most robust model against the adversarial attacks. Overall, InceptionV3 as an model architecture seems to be more robust against adversarial examples compared to ResNet50. Interestingly, performance against SPSA attacks with small $\epsilon$, the performance is high.

| | InceptionV3* % | InceptionV3 % | ResNet50* % | ResNet50 % |
|---|---|---|---|---|
| No attack | 94 | 95 | 93 | 92 |
| FGSM | 62 | 87 | 45 | 58 |
| CW | 6 | 5 | 7 | 8 |
| PGD1 | 70 | 89 | 35 | 30 |
| PGD2 | 55 | 79 | 20 | 11 |
| PGD5 | 39 | 62 | 9 | 8 |
| SPSA5, $\epsilon = 0.02$ | 92 | 94 | 93 | 93 |
| SPSA5, $\epsilon = 0.05$ | 74 | 91 | 94 | 92 |
| SPSA5, $\epsilon = 0.1$ | 55 | 71 | 94 | 92 |
| SPSA5, $\epsilon = 0.5$ | 49 | 49 | 79 | 50 |
| SPSA5, $\epsilon = 1.0$ | 49 | 49 | 51 | 57 |

Table 4: Evaluating adversarial robustness of the deterministic models on the DR dataset. The models trained with weighted CE loss are marked with (*). Adversarial attacks degrade the classification accuracy of the models. Again, against SPSA with small $\epsilon$ the performance is quite good. InceptionV3 seems to beat ResNet50 as an architecture of choice, providing high robustness against PGD1 and FGSM with its variant trained with non-weighted CE loss function. Otherwise, there isn't much difference between model versions trained with weighted and non-weighted CE loss.

# Appendix C: Evaluation of Bayesian model robustness against adversarial attacks

| | *Det* | $1^{-10}$ | $1^{-9}$ | $1^{-8}$ | $1^{-7}$ | $1^{-6}$ | $1^{-5}$ | $1^{-4}$ | $1^{-3}$ | $1^{-2}$ | $1^{-1}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| No attack | 95 | 95 | 95 | 95 | 95 | 95 | 95 | 95 | 96 | 94 | 94 |
| FGSM | 62 | 64 | 65 | 64 | 63 | 64 | 65 | 65 | 65 | 64 | 67 |
| CW | 5 | 6 | 5 | 6 | 23 | 5 | 6 | 15 | 42 | 47 | 39 |
| PGD1 | 53 | 55 | 56 | 55 | 53 | 57 | 57 | 57 | 58 | 54 | 56 |
| PGD2 | 26 | 27 | 29 | 29 | 25 | 31 | 32 | 32 | 32 | 25 | 27 |
| PGD5 | 9 | 10 | 11 | 11 | 9 | 10 | 11 | 11 | 11 | 10 | 11 |
| SPSA5, $\epsilon = 0.02$ | 95 | 95 | 95 | 95 | 95 | 96 | 96 | 95 | 96 | 94 | 94 |
| SPSA5, $\epsilon = 0.05$ | 95 | 95 | 96 | 95 | 94 | 96 | 95 | 94 | 95 | 95 | 94 |
| SPSA5, $\epsilon = 0.1$ | 92 | 92 | 92 | 93 | 90 | 93 | 93 | 93 | 92 | 91 | 91 |
| SPSA5, $\epsilon = 0.5$ | 27 | 27 | 27 | 27 | 27 | 27 | 27 | 27 | 27 | 27 | 27 |
| SPSA5, $\epsilon = 1.0$ | 27 | 27 | 27 | 27 | 27 | 27 | 27 | 27 | 27 | 27 | 27 |

Table 5: Classification accuracies (%): CXR Pneumonia dataset, InceptionV3 + VIB trained with different $\beta$ values evaluated against adversarial attacks. *Det* stands for 'deterministic'. Against the CW attacks, major increase in robustness is achieved. There is no significant increase in performance, only slight improvements, when it comes to other attacks.

| | *Det* | $1^{-10}$ | $1^{-9}$ | $1^{-8}$ | $1^{-7}$ | $1^{-6}$ | $1^{-5}$ | $1^{-4}$ | $1^{-3}$ | $1^{-2}$ | $1^{-1}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| No attack | 96 | 95 | 95 | 95 | 95 | 95 | 95 | 95 | 95 | 96 | 96 |
| FGSM | 10 | 17 | 17 | 16 | 17 | 42 | 15 | 12 | 13 | 14 | 12 |
| CW | 4 | 5 | 5 | 6 | 5 | 75 | 5 | 6 | 8 | 6 | 9 |
| PGD1 | 8 | 13 | 14 | 12 | 13 | 42 | 12 | 9 | 9 | 11 | 9 |
| PGD2 | 4 | 5 | 5 | 5 | 5 | 41 | 5 | 5 | 5 | 4 | 4 |
| PGD5 | 4 | 5 | 5 | 5 | 5 | 40 | 5 | 5 | 5 | 4 | 4 |
| SPSA5, $\epsilon = 0.02$ | 96 | 95 | 96 | 95 | 95 | 95 | 96 | 95 | 96 | 95 | 96 |
| SPSA5, $\epsilon = 0.05$ | 95 | 95 | 95 | 95 | 95 | 95 | 95 | 96 | 96 | 96 | 96 |
| SPSA5, $\epsilon = 0.1$ | 95 | 95 | 95 | 95 | 95 | 95 | 95 | 96 | 95 | 95 | 95 |
| SPSA5, $\epsilon = 0.5$ | 80 | 77 | 77 | 83 | 84 | 82 | 81 | 76 | 80 | 83 | 83 |
| SPSA5, $\epsilon = 1.0$ | 58 | 57 | 57 | 65 | 63 | 57 | 61 | 57 | 62 | 67 | 62 |

Table 6: Classification accuracies (%): CXR Pneumonia dataset, ResNet50 + VIB trained with different $\beta$ values evaluated against adversarial attacks. *Det* stands for 'deterministic'. Significant increase in robustness against all adversarial attacks is achieved when VIB is trained with $\beta = 1^{-6}$. Most notably, classification performance against CW attacks reaches a respectable accuracy. Still, against the other attacks, the predictive power is low, and these results are mostly just encouraging.

| | *Det* | $1^{-10}$ | $1^{-9}$ | $1^{-8}$ | $1^{-7}$ | $1^{-6}$ | $1^{-5}$ | $1^{-4}$ | $1^{-3}$ | $1^{-2}$ | $1^{-1}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| No attack | 94 | 94 | 95 | 94 | 94 | 94 | 94 | 94 | 94 | 94 | 94 |
| FGSM | 62 | 63 | 62 | 61 | 61 | 60 | 60 | 63 | 63 | 62 | 63 |
| CW | 6 | 7 | 9 | 6 | 7 | 7 | 16 | 21 | 30 | 37 | 42 |
| PGD1 | 70 | 71 | 74 | 72 | 71 | 69 | 70 | 73 | 71 | 76 | 77 |
| PGD2 | 55 | 57 | 57 | 57 | 56 | 55 | 55 | 57 | 56 | 57 | 61 |
| PGD5 | 39 | 44 | 42 | 41 | 42 | 41 | 41 | 43 | 43 | 44 | 43 |
| SPSA5, $\epsilon = 0.02$ | 92 | 92 | 92 | 92 | 92 | 91 | 92 | 92 | 91 | 91 | 92 |
| SPSA5, $\epsilon = 0.05$ | 74 | 73 | 73 | 74 | 76 | 75 | 75 | 74 | 73 | 71 | 74 |
| SPSA5, $\epsilon = 0.1$ | 55 | 54 | 55 | 55 | 56 | 55 | 56 | 55 | 55 | 55 | 55 |
| SPSA5, $\epsilon = 0.5$ | 49 | 49 | 49 | 49 | 49 | 49 | 49 | 49 | 49 | 49 | 49 |
| SPSA5, $\epsilon = 1.0$ | 49 | 49 | 49 | 49 | 49 | 49 | 49 | 49 | 49 | 49 | 49 |

Table 7: Classification accuracies (%): DR dataset, InceptionV3 + VIB trained with different $\beta$ values evaluated against adversarial attacks. *Det* stands for 'deterministic'. Only slight increases are visible when it comes to predictive accuracy. An exception is that against CW attack, we can see significant increase in predictive performance as $\beta$ increases towards $1^{-1}$.

| | *Det* | $1^{-10}$ | $1^{-9}$ | $1^{-8}$ | $1^{-7}$ | $1^{-6}$ | $1^{-5}$ | $1^{-4}$ | $1^{-3}$ | $1^{-2}$ | $1^{-1}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| No attack | 93 | 93 | 94 | 94 | 93 | 93 | 93 | 93 | 93 | 93 | 84 |
| FGSM | 45 | 47 | 47 | 46 | 47 | 47 | 47 | 46 | 46 | 46 | 58 |
| CW | 7 | 8 | 7 | 8 | 8 | 8 | 7 | 9 | 49 | 21 | 52 |
| PGD1 | 35 | 36 | 34 | 37 | 35 | 34 | 35 | 38 | 39 | 36 | 53 |
| PGD2 | 20 | 19 | 18 | 22 | 19 | 14 | 19 | 24 | 26 | 22 | 40 |
| PGD5 | 9 | 9 | 8 | 8 | 8 | 8 | 8 | 9 | 10 | 9 | 24 |
| SPSA5, $\epsilon = 0.02$ | 93 | 94 | 93 | 93 | 94 | 93 | 94 | 94 | 93 | 94 | 84 |
| SPSA5, $\epsilon = 0.05$ | 94 | 94 | 94 | 94 | 93 | 93 | 94 | 93 | 92 | 93 | 86 |
| SPSA5, $\epsilon = 0.1$ | 94 | 93 | 94 | 92 | 93 | 93 | 93 | 93 | 91 | 93 | 85 |
| SPSA5, $\epsilon = 0.5$ | 79 | 82 | 83 | 80 | 82 | 84 | 84 | 79 | 75 | 78 | 58 |
| SPSA5, $\epsilon = 1.0$ | 51 | 51 | 51 | 51 | 51 | 51 | 51 | 51 | 51 | 51 | 51 |

Table 8: Classification accuracies (%): DR dataset, ResNet50 + VIB trained with different $\beta$ values evaluated against adversarial attacks. *Det* stands for 'deterministic'. Clearly, the best adversarial robustness is achieved with $\beta = 1^{-1}$. Interestingly, the predictive accuracy against non-perturbed images is worse with the $\beta$ value that provides the best adversarial robustness. In comparison to experiments on the DR dataset with InceptionV3 + VIB model, now we also see an increase in robustness against not only CW, but all adversarial attacks.