Adversarial Learning Under Hybrid Perturbations for Robust Acute Lymphoblastic Leukemia Classification

Jie Chen[1], **Xinyuan Liu**[1], **Xintong Liu**[1], **Jianqiang Li**[1,2*]

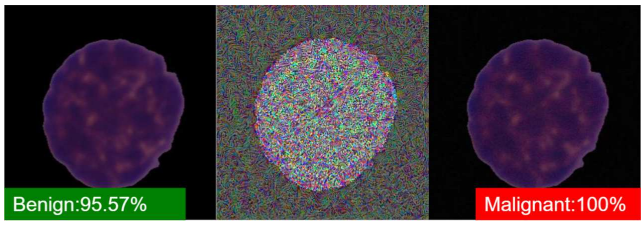
[1]College of Computer Science and Software Engineering, Shenzhen University, China
[2]National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, China
chenjie@szu.edu.cn, 1227653652@qq.com, 2210273132@email.szu.edu.cn, lijq@szu.edu.cn

Abstract

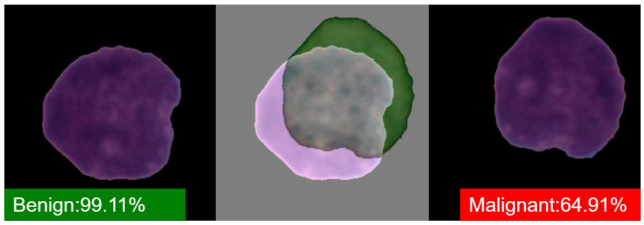
Acute lymphoblastic leukemia is a childhood cancer prevalent worldwide, which can prove fatal within weeks or months. However, current diagnosis models based on machine learning and deep learning methods fail to consider device noise (pixel-level perturbations) and rotation/translation (spatial-transformed perturbations), which can undermine the model's robustness. Adversarial training is a potential solution to this issue. This paper presents a hybrid perturbation adversarial training (HPAT) strategy that leverages two types of adversarial samples: pixel-level adversarial samples and spatial adversarial samples. This work generates these hybrid adversarial samples through Projected Gradient Descent (PGD) in couple with spatial transformation based on the Bayesian optimization (STBO) algorithm, respectively. This work introduced the Mixed Batch Normalization (MixBN) module to handle both adversarial samples and clean samples, alleviating the problem of clean accuracy degradation due to adversarial training. The proposed hybrid adversarial training strategy is tested on the public acute lymphoblastic leukemia dataset and found that it outperformed existing acute lymphoblastic cell classification models.

Introduction

Acute Lymphoblastic Leukemia (ALL) is one of the most common cancer among children and the most frequent cause of death from cancer before 20 years old (Smith et al. 2010). Recently, with the development of medical treatment, the survival rate of ALL has increased to 90% (Hunger and Mullighan 2015). However, Low and Middle-Income Countries (LMICs) still suffer from this disease due to the lack of costly infrastructure, trained human resources, and diagnostic tools at the required scale. Hence, these countries experience higher fatality rates. For example, 84% of ALL cases are reported from LMICs(Gehlot, Gupta, and Gupta 2020). Therefore, it is very important to develop a model for the automatic diagnosis of acute lymphoid leukocytes. Deep Learning (DL)-based outperforms Traditional Machine Learning in ALL classification due to the capability of automatic features extraction (Lamberti 2022; Prellberg and Kramer 2019; Hu, Shen, and Sun 2018; Tan and Le 2019). Additionally, due to variations in doctors' clinical practice, spatial-transformed perturbations also exist when they take images. These perturbations occur simultaneously in the scenario of ALL classification, which could cause misclassifications by the model. As illustrated in Fig. 1, after adding random but slight noise and spatial transformations to the original examples, the model made opposite predictions. Consequently, if an ALL cell classification model failed to consider the robustness, it will lead to diagnostic errors, inappropriate treatment decisions, and further health risks, which could cause a significant impact on the patient's life. Thus, this work dedicates to the ALL cell classification models' robustness.



(a) Pixel-level perturbation



(b) Rotation and transition

Figure 1: Robustness issue in Acute Lymphoblastic Leukemia Cell Classification (ALLCC) due to pixel-level perturbation and spatial-transformed perturbation.

To incorporate model robustness in the acute lymphoblastic leukocyte classification, this paper presents a novel framework, called Hybrid-Perturbation Adversarial Training (HPAT) which considers the existence of pixel-level perturbations and spatial-transformed perturbations in the scenario of acute lymphoblastic leukemia cell classification scenario.

*Corresponding author.

The HPAT framework is designed in the adversarial training framework which consists of two stages: generating adversarial examples and training. To alleviate the issue due to existence of pixel-level and spatial-transformed perturbations in the ALL cell classification, this work first generats two types of adversarial examples using Projected Gradient Descent (PGD) coupled with STBO, and then add them to the original examples to obtain mixed adversarial examples. Then, the mixed adversarial examples are used to train the model to be able to genenalize in such environment with both types of perturbations. As a result, the model can effectively reduce the impact of pixel-level perturbation and spatial-transformed perturbation, thereby improving the robustness of the deep learning model.

Our contributions are summarized as follows:

• To improve the efficiency of spatial adversarial examples generation, this work proposed a novel method based on Bayesian optimization for search the spatial-transformed adversarial examples with only a few iterations (Section 3.2).

• To alleviate the accuracy degradation in clean examples, this work present a mixed batch normalization component to capture different patterns (Section 3.3).

• To improve the model's robustness allowing to work properly in the environment with mixed types of perturbations, this work propose a Hybrid-Perturbation Adversarial Training (HPAT) framework which can generate mixed adversarial examples by combining pixel-level perturbation and spatial-transformed perturbation, improving the model's robustness by allowing it to adapt to both types of perturbation (Section 3.4).

• Empirically, this work is the first to identify the poor robustness issue in existing acute lymphocyte classifications models (Section 4.1). It is also verified in the real world dataset to show the effectiveness of proposed framework(Section 4.2).

Related Work

Existing blood cell classification methods can be divided into traditional machine-learning and deep learning-based methods. Traditional machine-learning methods are based on the extraction of image features and classification with machine-learning models. For example, (Tai et al. 2011) proposed a blood cell classification method based on feature selection and hierarchical SVM to classify white blood cells, red blood cells, and platelets. Recently, deep learning methods are widely used in cell classification, which uses deep learning methods to extract image features and classify them.(Prellberg and Kramer 2019) using a ResNet50 with Squeeze-and-Excitation modules to automatically classify white blood cell images into benign B-lymphoid precursors and malignant B-lymphoblasts, this work achieves a weighted F1-score of 88.91% on the test set and ranked 10th in the IEEE ISBI-2019 B-ALL Challenge(Gupta et al. 2020). Deep learning models are susceptible to adversarial attacks. In the context of autonomous driving, this vulnerability may result in traffic accidents(Ma, Driggs-Campbell, and Kochenderfer 2019). Furthermore, the Segment Anything Model (SAM) exhibits limitations in identifying instruments within certain complex surgical scenarios(Wang et al. 2023). This deficiency can potentially contribute to medical fraud and compromise the confidentiality of patient information in medical settings (Finlayson et al. 2019). Adversarial attacks on medical images can be carried out through various methods, such as adding pixel-level perturbations or rotations. Constrained by the fact that traditional data augmentation methods typically do not consider adversarial attack scenarios, more advanced techniques such as adversarial training may prove to be more effective in addressing adversarial challenges across all scenarios(Rebuffi et al. 2021). Researchers have conducted antagonistic attacks on different clinical fields, including Fundoscopy, Chest X-Ray, and Dermoscopy, proving the vulnerability of medical deep learning models (Finlayson et al. 2018; Guo et al. 2019; Ma et al. 2021). Adversarial training strategies have been proposed to defend against attacks and improve the model's robustness and reliability (Koga and Takemoto 2022; Bortsova et al. 2021; Paschali et al. 2018; Han et al. 2021). For example, (Uwimana and Senanayake 2021) propose methods for the detection of malaria, this method is based on Mahalanobis distance to detect antagonistic samples and Out-of-Distribution samples, the test samples that are not well covered by training data. To improve the robustness of the model, comparing the Local Intrinsic Dimensionality adversarial detection method, better performance was achieved on the four adversarial examples: FGSM(Goodfellow, Shlens, and Szegedy 2015), BIM(Kurakin, Goodfellow, and Bengio 2016), CW(Carlini and Wagner 2017), and Deep Fool(Moosavi-Dezfooli, Fawzi, and Frossard 2016). Furthermore, the use of hybrid-perturbation adversarial training and misclassification-aware adversarial training has shown promising results in improving the robustness of deep diagnostic models. For example, (Mao et al. 2021) proposes a composite attack method Composite Adversarial Attacks, combining Spatial attack and FGSM attack, two attackers can generate four possible permutations, and then find the best permutation by using a search algorithm, this method achieves the most Advanced white-box scenarios, significantly reducing attack time cost. It is crucial to consider the robustness and reliability of medical deep learning models to ensure safe and effective patient care(Khakzar, Albarqouni, and Navab 2019).

There are a series of adversarial training strategies that mix them with multiple perturbations. Based on whether the types of interference are the same, existing mixed adversarial training methods can be divided into isomorphic methods(Salman et al. 2019; Schott et al. 2018; Tramèr et al. 2018; Araujo et al. 2019; Zhang and Gao 2021; Liu et al. 2020) and heterogeneous methods(Kamath et al. 2021). Isomorphic interference refers to the robust performance of a model against the same type of interference. For example, (Mao et al. 2021) proposes a composite attack method Composite Adversarial Attacks (CCA), combining Spatial attack and FGSM attack, this method achieves the most Advanced white-box scenarios, significantly reducing attack time cost. (Tramèr and Boneh 2019) proposes an adversarial training strategy combining random rotation and translation perturbation to defend against multiple types of adversarial attack

accuracy. In the ALL scenarios, there could be two types of perturbation, one is the pixel-level perturbation due to device reasons, and the other is the perturbation due to the different rotation and translation of the doctor's shooting angles. In this case of heterogeneous interference, due to the different types of interference requiring different robustness methods, models with homogeneous interference are no longer used due to their lack of ability to handle heterogeneous interference(Song et al. 2022). Heterogeneous methods (Kamath et al. 2021) proposed a strategy of adversarial training by mixing PGD adversarial examples and spatially transformed adversarial examples, balancing adversarial accuracy and spatial accuracy. However, existing studies lack consideration for spatial-transformed perturbations and cannot fully reflect the diversity and complexity of ALL scenarios.

Methodology

Problem Formulation

The Acute Lymphoblastic Leukemia Cell Classification (ALLCC) task aims to learn a mapping $f : \mathbb{R}^{h\times w\times 3} \mapsto \{0, 1\}$ from a given dataset $\mathcal{D}$. Each point in $\mathcal{D}$ is a tuple $(\mathbf{X}, y)$ where $\mathbf{X} \in \mathbb{R}^{h\times w\times 3}$ is an RGB cell image of h rows and w columns; $y \in \{0, 1\}$ is a ground truth label with 0 and 1 indicating a benign and malignant cell, respectively. Commonly, the models of ALLCC can be chosen as deep neural networks (e.g., ResNet, EfficientNet, and DesNet) denoted as $f_\theta(\cdot)$ with θ being the learnable weights. Then, given any unlabeled cell image $\hat{\mathbf{X}}$, an accurate predictive label $\hat{y} = f_\theta(\hat{\mathbf{X}})$ can be used to assist the doctors to alleviate the burden of spotting malignant cells from the enormous number of benign cells. Many research works(Rehman et al. 2018; Prellberg and Kramer 2019) have addressed the aforementioned ALLCC tasks in a deep learning framework. In practice, the robustness of deep ALLCC models could be compromised due to two types of perturbation on the input image $\mathbf{X}$: (1) The pixel-level perturbation originated from the instruments for sensing the cells. The pixel-level perturbation can be denoted by a function $A_\epsilon : \mathbb{R}^{h\times w\times 3} \mapsto \mathbb{R}^{h\times w\times 3}$ which captures the pixel-level variations due to instrumental measure noise. Sometimes, with a very small variation ϵ (i.e., $\|A_\epsilon(\mathbf{X}) - \mathbf{X}\| < \epsilon$), the predictive label could unexpectedly flipped: $f_\theta(\mathbf{X}) \neq f_\theta(A_\epsilon(\mathbf{X}))$. (2) The spatial-transformed perturbation resulted from the doctors' manual operations when collecting the cell images. Similarly, let the function $T_\mathbf{s}(\mathbf{X})$ capture the spatial-transformed perturbation on any cell image $\mathbf{X}$ where $\mathbf{s} \triangleq (\xi, \zeta, \alpha)$ are the parameters for modeling the patterns with ξ and ζ control the amount of horizontal and vertical translation, and α is the degree of rotation of the image. At certain circumstance, the prediction could also flip $f_\theta(\mathbf{X}) \neq f_\theta(T_\mathbf{s}(\mathbf{X}))$ with a small change in $\mathbf{s}$. The effects of the above two kinds of perturbations are not well-studied in existing ALLCC models but they are extremely important to consider in life-critical applications. This leads to the research question : *How to improve the robustness of ALLCC models under hybrid perturbations?* The adversarial learning framework can be exploited to improve the robustness of the deep learning models:

$$\min_\theta \mathbb{E}_{(\mathbf{X},y)\sim\mathcal{D}} \left[\max_{\epsilon,\mathbf{s}} \ell\left(f_\theta(T_\mathbf{s}(A_\epsilon(\mathbf{X}))), y\right)\right] \quad (1)$$

where θ captures the parameters of ALLCC model $f_\theta(\mathbf{X})$, $\ell(f_\theta(\mathbf{X}), y)$ is the empirical loss for fitting the training data, and the inner part of the formula is a process of maximizing the adversarial loss for fixing the weakest parts of the ALLCC model under hybrid perturbations. However, in order to apply the adversarial training framework to an ALLCC model, both pixel-level perturbations and spatial-transformed perturbations need to be explicitly specified. In the following section, this work proposes an adversarial training framework called Hybrid Perturbation Adversarial Training for Acute Lymphoblastic Leukemia Cell Classification (HPAT-ALLCC). The workflow of the proposed framework is shown in Fig. 2.

Generating Adversarial Samples under Hybrid Perturbations

Due to the instrumental noise when microscopy equipment senses the cells, the images inevitably have pixel-level perturbations. Classification models could output incorrect prediction results when the images have pixel-level perturbations. In order to improve the robustness of the model, it is necessary to train the model with pixel-level adversarial examples. In this work, we uses the gradient-based multiple iteration method PGD(Madry et al. 2018) to generate pixel-level adversarial examples $\mathbf{X}_\epsilon = A_\epsilon(\mathbf{X})$ where $A_\epsilon(\cdot)$ can be specified as

$$\mathbf{X}_\epsilon^{t+1} = \mathrm{clip}_{[0,1]}\left(\mathbf{X}_\epsilon^t + \epsilon \cdot \mathrm{sign}\left(\frac{\mathbf{G}^t}{\|\mathbf{G}^t\|}\right)\right), \quad (2)$$

$$\mathbf{G}^t = \nabla_\mathbf{X} \ell\left(f_\theta\left(\mathbf{X}^t\right), y\right) \quad (3)$$

where $\mathbf{X}_\epsilon^t$ represents the adversarial examples generated in the t-th iteration, and $\mathbf{G}^t$ represents the gradient with respect to $\mathbf{X}_\epsilon^t$. In each iteration of the PGD algorithm, $\mathbf{G}^t$ needs to be updated. ϵ is the parameter that controls the level of the adversarial perturbation. $\mathrm{sign}(\cdot)$ extracts the sign of the gradient and $\mathrm{clip}_{[0,1]}$ restricts the pixel values to the range $[0, 1]$. In each iteration, the direction of generating adversarial examples is updated by computing the gradient of the model at the $\mathbf{X}_\epsilon^t$ with a limited perturbation range. As the cell images are normally collected by doctors, various spatial-transformed perturbations (e.g., rotations and translations) could be introduced into cell images for acute lymphocyte classification. Such spatial-transformed perturbations could cause the model to misclassify. Thus, additional adversarial examples are required in the training process to make the model resist the spatial-transformed perturbation. Similarly, let $\mathbf{X}_\mathbf{s} = T_\mathbf{s}(\mathbf{X})$ be transformed image of $\mathbf{X}$ under small rotation and translation perturbation specified by $\mathbf{s}$. Then, in order to implement the maximization operation in Eq. (1), this work presents a Spatial Transform based on Bayesian Optimization (STBO) technique to generate spatial-transformed adversarial examples.

Spatial Transform based on Bayesian Optimization: Inspired by (Engstrom et al. 2019a), this work specify the

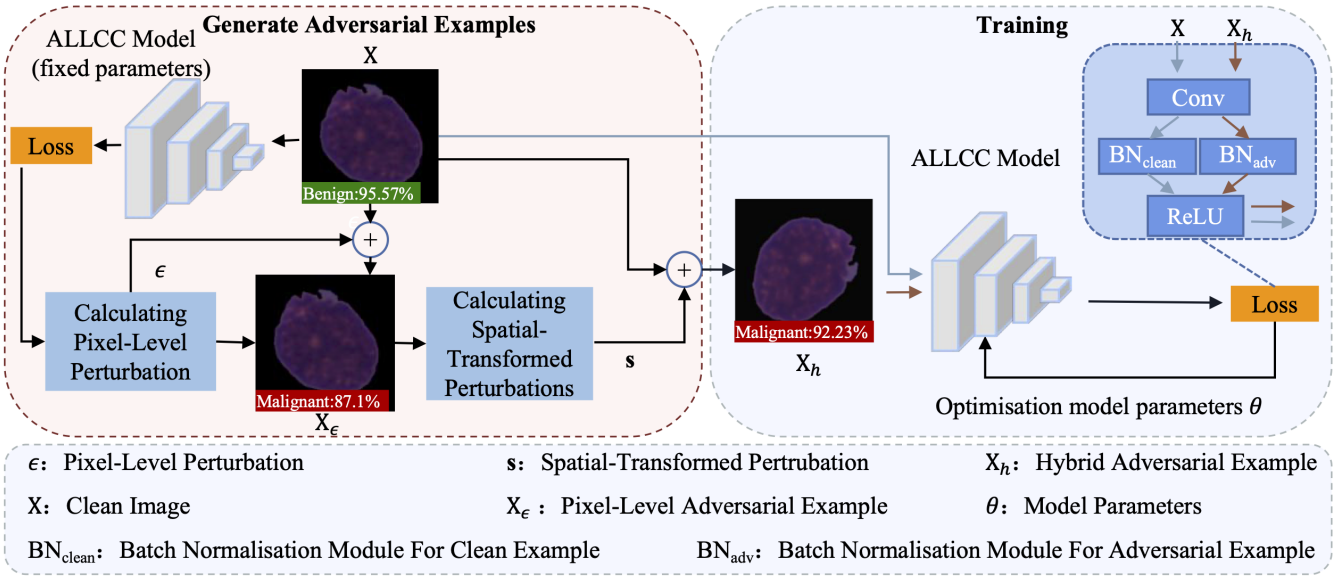


Figure 2: Overview of our proposed HPAT-ALLCC framework for Acute Lymphoblastic data.

function $T_{\mathbf{s}}(\cdot)$ in Eq. (1) as follows. Let $\mathbf{X}_s = T_{\mathbf{s}}(\mathbf{X})$, then the spatial-transformed image can be computed as $[\mathbf{X}_s]_{(u',v')} = [\mathbf{X}]_{(u,v)}$ where

$$\begin{pmatrix} u' \\ v' \end{pmatrix} = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} \cdot \begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} \xi \\ \zeta \end{pmatrix}. \quad (4)$$

Here, (u, v) is the position of a specific pixel in cell image $\mathbf{X}$ and (u', v') is a position after spatial transform specified by $\mathbf{s} = (\xi, \zeta, \alpha)$ where ξ, ζ control translation and α controls rotation. Unlike using the grid search(Engstrom et al. 2019a), this work uses Bayesian optimization(Kandasamy et al. 2017) to find the optimal hyperparameter $\mathbf{s}$ in a more efficient manner. Then, the spatial-transformed adversarial samples can be obtained given any clean images. In the following, this work will describe how the Bayesian Optimization technique is used to find the spatial-transformed adversarial examples.

The goal of STBO is to find the optimal spatial-transformed perturbation parameters $\mathbf{s}$:

$$\mathbf{s}^* = \operatorname{argmax} \mathbb{E}_{(\mathbf{X},y)\sim\mathcal{D}} \left[\ell\left(f_\theta\left(T_{\mathbf{s}}(A_\epsilon(\mathbf{X}))\right), y\right)\right] \quad (5)$$

where θ and ϵ are fixed. The above maximization can be achieved using Bayesian Optimization (BO) which relies on two components: the prior and the acquisition function. Gaussian process (GP)(Williams and Rasmussen 2006) model can serve as a prior in BO, allowing for the integration of domain knowledge about the problem through a kernel over the input space. A GP is fully specified by a mean function $m(\mathbf{s})$ and the covariance function $k(\mathbf{s}, \mathbf{s})$, i.e., $g(\cdot) \sim \mathrm{GP}(m(\cdot), k(\cdot,\cdot))$. At t-th iteration, a set $\mathcal{S} = \{(\mathbf{s}_i, g(\mathbf{s}_i))\}_{i=1}^t$ is collected by measuring the loss $g(\mathbf{s})$ with the best candidate selected using Eq. (5). Then, the posterior process $g(\hat{\mathbf{s}} \mid \mathcal{S})$ is also a GP with mean $\widehat{\mu}^t(\hat{\mathbf{s}})$ and variance $\widehat{\sigma}^t(\hat{\mathbf{s}})$ given by

$$\widehat{\mu}^t(\hat{\mathbf{s}}) = \kappa^T \mathbf{K}^{-1} \mathbf{g}, \quad (6)$$

$$\widehat{\sigma}^t(\hat{\mathbf{s}}) = k(\hat{\mathbf{s}}, \hat{\mathbf{s}}) - \kappa^T \mathbf{K}^{-1} \kappa \quad (7)$$

where $\mathbf{g}$ is a vector with $[\mathbf{g}]_i = g(\mathbf{s}_i)$, the matrix $\mathbf{K}$ is computed by $[\mathbf{K}]_{i,j} = k(\mathbf{s}_i, \mathbf{s}_j)$, κ is a vector with $[\kappa]_i = k(\hat{\mathbf{s}}, \mathbf{s}_i)$, and κ^T is the transpose of κ.

Particular acquisition functions based on $\widehat{\mu}^t$ and $\widehat{\sigma}^t$ can be used for selecting near-optimal points to be evaluated in the next iteration t. This work uses Expected Improvement (EI)(Jones, Schonlau, and Welch 1998) as the acquisition function denoted as $\alpha_{\mathrm{EI}}(\cdot)$. However, many other options cite can be used too. Intuitively, if sampling a point can result in a better value than the current optimal solution, the EI value will be positive. In this case, the maximum value of the EI function needs to be determined in order to obtain the next point. Specifically,

$$\begin{aligned} \alpha_{\mathrm{EI}}(\hat{\mathbf{s}}) = \left(\widehat{\mu}^t(\hat{\mathbf{s}}) - g(\mathbf{s}^*)\right) \phi\left(\frac{\widehat{\mu}^t(\hat{\mathbf{s}}) - g(\mathbf{s}^*)}{\widehat{\sigma}^t(\hat{\mathbf{s}})}\right) \\ + \widehat{\sigma}^t(\hat{\mathbf{s}}) \phi\left(\frac{\widehat{\mu}^t(\hat{\mathbf{s}}) - g(\mathbf{s}^*)}{\widehat{\sigma}^t(\hat{\mathbf{s}})}\right) \end{aligned} \quad (8)$$

where $g(\mathbf{s}^*) = \max_{i=1}^t g(\mathbf{s}_i)$ is the optimal value among the currently available measured points in $\mathcal{S}$, and $\widehat{\mu}^t(\hat{\mathbf{s}})$& $\widehat{\sigma}^t(\hat{\mathbf{s}})$ can be evaluated by Eqs. (6)&(7). Then, the next point $\mathbf{s}^{t+1}$ to be measured can be by determined by the following equation:

$$\mathbf{s}^{t+1} = \operatorname*{argmax}_{\hat{\mathbf{s}}} \alpha_{\mathrm{EI}}(\hat{\mathbf{s}}). \quad (9)$$

The evaluated points will be added into the set $\mathcal{S}$ during the iteration. The above procedure can be repeated until the maximum number of iterations is reached. The parameter corresponding to the current $g(\mathbf{s}^*)$ is the parameter for spatial-transformed perturbation model $T_{\mathbf{s}}(\cdot)$. As the acquisition function can trade off exploitation and exploration, it tends to converge to the optimal point with much less evaluation than grid search.

Alleviating Degradation of Predictive Accuracy in Clean Examples

In practice, the predictive accuracy of the clean samples could degenerate as a side-effect of adversarial samples. Xie et. al. give an explanation of such effect that the adversarial and clean examples reside in different distributions(Xie et al. 2020). Then, a uniform normalization could lead to poorly-posed statistical estimation which subsequently hurts the model's performance. Unlike existing models, this work introduces the Mixed Batch Normalization (MixBN) module in the HPAT framework to handle hybrid perturbations. Specifically, the MixBN module is inserted before the activation function, each Batch Normalization (BN) branch is responsible for processing one type of distribution. Then ReLU function follows after the mixed BN module which is computed as follows:

$$\begin{aligned}\eta\left(\mathbf{X}, \mathbf{X}_h\right) = \operatorname{ReLU}\left(\operatorname{BN}_{\text{clean}}(\operatorname{Conv}(\mathbf{X}))\right. \\ \left. + \operatorname{BN}_{\text{adv}}(\operatorname{Conv}(\mathbf{X}_h))\right)\end{aligned} \tag{10}$$

where $\operatorname{Conv}(\cdot)$ is a stack of convolutional layers, $\operatorname{BN}_{\text{clean}}(\cdot)$ and $\operatorname{BN}_{\text{adv}}(\cdot)$ are computed similarly with different parameters. Each BN branch $\operatorname{BN}(\cdot;\beta,\gamma)$ is parameterized by β and γ. Given any vectorized input $\mathbf{x}$, the normalized vector $\widetilde{\mathbf{x}}$ is calculated as follows:

$$[\widetilde{\mathbf{x}}]_i = \gamma \cdot \left(\frac{[\mathbf{x}]_i - \mu_{\text{BN}}}{\sqrt{\sigma^2_{\text{BN}} + \omega}}\right) + \beta \tag{11}$$

where ω is a small value to avoid division by zero and μ_{BN} and σ^2_{BN} are computed by

$$\mu_{\text{BN}} = \frac{1}{|\mathbf{x}|}\sum_{i=1}^{|\mathbf{x}|}[\mathbf{x}]_i, \tag{12}$$

$$\sigma^2_{\text{BN}} = \frac{1}{|\mathbf{x}|}\sum_{i=1}^{|\mathbf{x}|}([\mathbf{x}]_i - \mu_{\text{BN}})^2. \tag{13}$$

Then, the $\operatorname{BN}_{\text{clean}}(\cdot)$ and $\operatorname{BN}_{\text{adv}}(\cdot)$ in Eq. (10) are required to specified with different parameters (i.e., β and γ) for clean samples and hybrid-perturbed adversarial samples.

Adversarial Learning under Hybrid Perturbations

Finally, the Hybrid-Perturbations Adversarial Training (HPAT) framework combines both pixel-level perturbation and spatial-transformed perturbation to generate hybrid adversarial examples to improve the robustness of any deep ALLCC models. The final min-max training framework is as follows:

$$\min_{\theta}\mathbb{E}\left[\ell\left(f_\theta(\mathbf{X}), y\right) + \max_{\epsilon,\mathbf{s}}\ell\left(f_\theta\left(\mathbf{X}_h\right), y\right) + \eta\left(\mathbf{X}, \mathbf{X}_h\right)\right] \tag{14}$$

where $\mathbf{X}_h$ denotes $T_{\mathbf{s}}(A_\epsilon(\mathbf{X}))$ and additional $\ell(f_\theta(\mathbf{X}), y)$ is used to treat clean samples separately. The maximization aims to maximize the inner adversarial losses in order to obtain the optimal parameters of pixel-level perturbation and the spatial-transformed perturbation (i.e., ϵ and $\mathbf{s}$). The minimization process aims to minimize both clean loss and adversarial loss by optimizing model parameters θ. The MixBN module $\eta(\mathbf{X}, \mathbf{X}_h)$ can well handle the discrepancy among distributions of clean samples and adversarial samples. The HPAT training framework is shown in Algorithm 1. It contains two phases: (1) generating the hybrid adversarial samples and (2) minimizing the losses. Specifically, the model f_θ was pre-trained using ImageNet data. Then, a batch size of m is sampled from the set $\mathcal{B} = \{\mathbf{X}_i, y_i\}_{i=1}^m$ in each training round. Next, pixel-level adversarial examples are generated in an iterative process, and then the pixel-level adversarial examples are fed into the $T_{\mathbf{s}}(\cdot)$ function to generate a mixed adversarial sample $\mathbf{X}_i^{'}$. Then, the parameter θ is updated to optimize the loss function, which includes the loss of the clean example and the loss of the mixed adversarial example. Finally, repeat steps until the number of iterations reaches L_O and the final model f_θ is output.

Algorithm 1: Hybrid-Perturbation Adversarial Training (HPAT)

Require: Training data $\mathcal{D}$, m is batch size, $\mathbf{X}_i'$ is mixed adversarial example, L_I is inner iteration step and L_O is outer iteration epoch, ϵ is maximum perturbation, step size for inner optimization Δ_I and step size for outer optimization Δ_O.

initialize network f_θ with pre-trained configuration
for $k = 1, ..., L_O$ **do**
  Read mini-batch $\mathcal{B}$=$\{(\mathbf{X}_i, y_i)\}_{i=1}^m$ from training set $\mathcal{D}$
  for $\mathbf{X}_i \in \mathcal{B}$ **do**
    $\mathbf{X}_i' \leftarrow \mathbf{X}_i + 0.001 \cdot \mathcal{N}(\mathbf{0}, \mathbf{1})$
    for $t = 1, ..., L_I$ **do**
      Compute pixel-level adversarial examples $\mathbf{X}_i'$ using Eqs. (2)&(3).
      Compute $\mathbf{X}_i' \leftarrow T_{\mathbf{s}}(\mathbf{X}_i')$ based on STBO
    end for
  end for
  Compute mixed batch norm $\eta(\mathbf{X}, \mathbf{X}')$ using Eq. (10)
  $\theta \leftarrow \theta - (\Delta_O/m)\sum_{i=1}^m \nabla_\theta[\ell(f_\theta(\mathbf{X}_i), y_i) + \ell(f_\theta(\mathbf{X}_i'), y_i) + \eta(\mathbf{X}, \mathbf{X}')]$
end for

Ensure: Return ALLCC model parameter θ

Experiments and Results

Dataset description and pre-processing: The acute lymphoblastic leukemia dataset C_NMC from the IEEE ISBI-2019 B-ALL benign and malignant cell classification challenge(Gupta et al. 2020), which contains a large number of labeled images of begin and malignant cells. C_NMC has a total of 12527 images. Among them, 8258 images are used for the training, 2132 for validation, and 1867 for the final test. ALL the cells have been preprocessed via stain normalization and cell segmentation(Gupta et al. 2017) and 450x450 pixel images were center cropped to 300x300 pixels in order to reduce the interference of the black background better.

Benchmark models: This work chooses some model back-

Model	ACC	PRE	F1	AUC
DesNet-201	78.865	81.213	73.745	86.112
ResNet50	79.535	81.846	74.224	89.421
VGG16	76.889	81.662	68.814	82.043
SE-ResNet50Xt	76.012	76.264	68.591	77.391
EfficientNet-B3	**84.524**	**83.989**	**82.373**	**91.681**

Table 1: Performance evaluation of existing ALLCC benchmark models.

bones as the baseline, including VGG16, DesNet-201, ResNet50, SE-ResNet50Xt, and EfficientNet-B3. All models were trained on the C_NMC dataset for 20 epochs, using Adam optimizer with a batch size of 16 to update network parameters. The initial learning rate is 0.001 and decay by 0.1 every ℓ_2 steps.

Performance metrics: Accuracy (ACC), Precision (PRE), Recall (REC), F1-score (F1), and area under the curve (AUC) are employed as performance metrics in this work. Additionally, the rate of invariance (ROI)(Kamath et al. 2021) is also used to evaluate models' adversarial and spatial robustness. After adding perturbations or spatial transformations to the origin examples, the predicted labels will change, resulting in a lower ROI. The formula for the ROI is defined as follows:

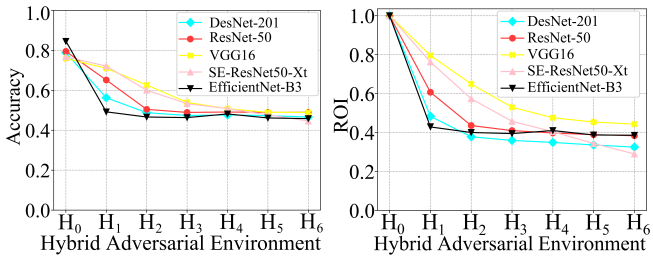
$$ROI = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left(f_\theta\left(\mathbf{X}_i\right) = f_\theta\left(\widetilde{\mathbf{X}}_i\right)\right) \tag{15}$$

where $\widetilde{\mathbf{X}}_i$ is a cell image $\mathbf{X}_i$ under hybrid perturbation and $\mathbb{I}(\cdot)$ is the indicator function which returns 1 if the input expression is true and 0 otherwise.

Results and Analysis

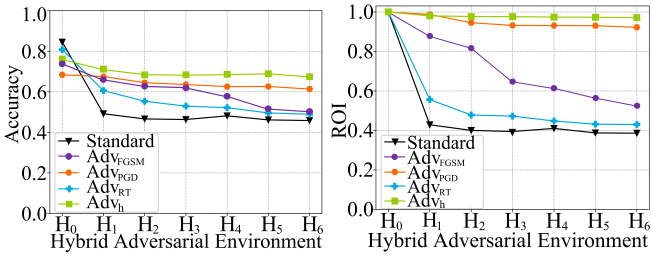
Evaluating ALLCC model This work first evaluates several Acute Lymphoblastic Leukemia Cell Classification (ALLCC) models. These ALLCC models were pre-trained on ImageNet(Deng et al. 2009). Then, the existing ALLCC models are tested on the C_NMC dataset under various performance metrics (see section 4.1). As shown in Table. 1, the EfficientNet-B3 model outperforms other models in all four metrics. Based on the experimental results the EfficientNet-B3 model will be a good candidate as a backbone model for evaluating the HPAT framework.

Evaluating model robustness in hybrid adversarial environment As there is no prior work attempted to evaluate the robustness of existing ALLCC models, this work first build a mixed adversarial environment using hybrid adversarial examples. Such a mixed adversarial environment can serve as a testbed for evaluating the proposed training scheme. As shown in Fig. 3, the ROI and accuracy of the model are evaluated in a mixed adversarial environment. The horizontal coordinates in Fig. 3 represent in a hybrid adversarial environment, denoted as $H_{0\sim6}$. Here, H_0 represents the clean examples. For $H_{1\sim6}$, their parameters $(\epsilon, \alpha, \xi, \zeta)$ are set by element-wise combination of (1/255, 2/255, 3/255, 4/255, 5/255, 6/255), $(30°, 60°, 90°, 120°, 150°, 180°)$, and



(a) ACC (b) ROI

Figure 3: Evaluating robustness of ALLCC models in hybrid adversarial environment.

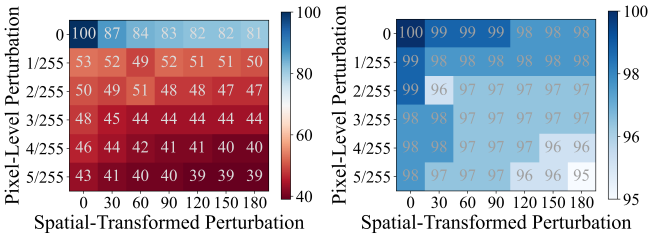


(a) ACC (b) ROI

Figure 4: Evaluating robustness of the proposed method in hybrid adversarial environment.

(2px, 4px, 6px, 8px, 10px, 12px). As shown in Fig. 3, the accuracy of several ALLCC models decreased significantly. It is evident that existing ALLCC models are vulnerable to pixel-level perturbations or spatial-transformed perturbations and the model is easily misclassified due to both types of perturbations.

Comparison of standard training and hybrid adversarial training To demonstrate the effectiveness of the proposed HPAT framework, (see Fig. 4), this work compares the performance of the hybrid perturbation adversarial training model ($\mathrm{Adv_h}$) with the pixel-level adversarial PGD method ($\mathrm{Adv_{PGD}}$), the pixel-level FGSM method ($\mathrm{Adv_{FGSM}}$), and the spatially transformed adversarial training model ($\mathrm{Adv_{RT}}$). The experiments also compared the standard trained model (trained with clean examples only), denoted as $\mathrm{Standard}$. $\mathrm{Adv_h}$ is trained using EfficientNet-B3 as the backbone model, introducing mixed adversarial examples and clean examples for adversarial training. The mixed adversarial example is generated by two types of perturbations (pixel-level perturbations and spatial-transformed perturbations). The pixel-level perturbations were obtained using an iterative PGD algorithm with a perturbation strength set to 6/225, an iteration number of 6, and a step size of 1/255. The spatial-transformed perturbations were implemented using the STBO algorithm, with the rotational and translational perturbations are $[-180°, 180°]$ and $[-12\mathrm{px}, +12\mathrm{px}]$. The number of hyperparameter searches is 100. $\mathrm{Adv_p}$ is trained by introducing pixel-level adversarial examples and clean examples, and the $\mathrm{Adv_{RT}}$ model is trained by introducing spatial adversarial examples



(a) Benchmark (b) HPAT

Figure 5: ROI of benchmark and HPAT model under hybrid perturbation (pixel-level perturbation + spatial-transformed (rotation) perturbation).

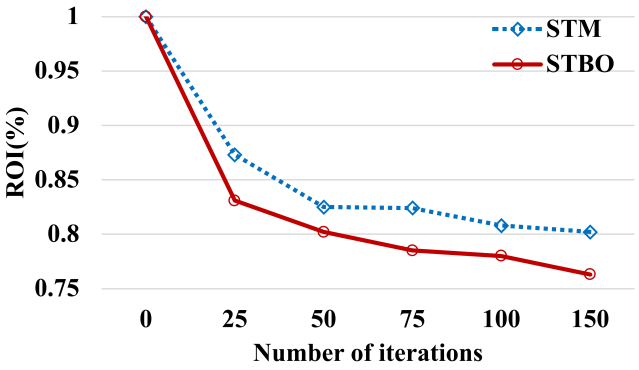


Figure 6: Comparing ROI of STM and STBO.

and clean examples. In comparison with several different training strategies, the accuracy and ROI for the Standard model have dropped sharply, from 84.5% to 48.3%. The $\mathrm{Adv_{PGD}}$ model (orange), the $\mathrm{Adv_{FGSM}}$ model (purple) and the $\mathrm{Adv_{RT}}$ model (blue) show a slower dropping trend in accuracy, but a fast decrease in ROI. The $\mathrm{Adv_h}$ model (green) has a more stable trend in both accuracy and ROI and is not affected by adversarial perturbations. It is implies that the HPAT algorithm improves the robustness of the model and is much better than the standard training and single-perturbation adversarial training.

We also test the ALLCC models in more combinations between two types of perturbations. As shown in Fig. 5(a), with both pixel-level and spatial-transformed perturbations available, the ROI of benchmark model decreases significantly when level of hybrid perturbations increases. In contrast, the models trained using HPAT framework can effectively mitigate the rate of ROI decline (see Fig. 5(b)).

Performance of spatial transform based on Bayesian optimization In this work, the proposed STBO algorithm is used to generate spatial-transformed adversarial samples, which relies on Bayesian Optimization to improve the efficiency of searching for spatially perturbed parameters. Compared to the grid search-based method STM(Engstrom et al. 2019b), STBO can better trade off the exploration and exploitation in searching the most vulnerable examples. As shown in Fig. 6, we compare the trends of ROI changes for STM and STBO algorithms. If the ROI is lower, then the attack success rate is higher which means generating more samples that are difficult to classify, and these samples are extremely useful for adversarial training. It can be found that the ROI of the STBO algorithm is lower. The results show that the curve of STBO is consistently under the the curve of STM. It implies adversarial examples of better quality can be obtained, which are more useful for adversarial training.

Model	ACC	PRE	F1	AUC
Adv_p	68.398	70.329	78.669	60.457
Adv_p+MixBN	**70.166**	**73.778**	**78.667**	**72.919**
Adv_{RT}	80.824	80.125	86.480	87.306
Adv_{RT}+MixBN	**82.110**	**84.619**	**86.714**	**87.047**
Adv_h	75.416	72.124	79.816	75.148
Adv_h+MixBN	**77.647**	**74.818**	**82.503**	**79.809**

Table 2: Experimental results comparing the model with and without the addition of MixBN.

Ablation experiments of mixed batch normalization In order to demonstrate the effectiveness of introducing the MixBN module. This conducted ablation experiments on the $\mathrm{Adv_p}$, $\mathrm{Adv_{RT}}$ and $\mathrm{Adv_h}$ models respectively. The results of the ablation experiments are shown in Table. 2. It can be found that the models improved in accuracy, precision, F1 score, and AUC metrics after the introduction of the MixBN module, with the overall results being higher than the baseline model. It implies the effectiveness of the MixBN module, which improved the model performance by normalizing the mixed adversarial samples and clean samples respectively and improved the problem of clean accuracy degradation caused by adversarial training.

Conclusion

This work proposes an HPAT algorithm for the acute lymphocyte classification task to improve the robustness of the ALL model. Different from existing ALL models, this work considers the doctor's different camera angles and different devices, which will lead to images with pixel-level perturbations and spatial-transformed perturbations, thus this work considers the robustness of the model. This work proposes the Bayesian optimization-based generative space adversarial sample algorithm STBO, which experimentally proves to be more efficiently queried than the STM algorithm. This work demonstrates the vulnerability of existing ALL models to adversarial samples and proposes a deep diagnostic framework based on HPAT adversarial training. Experiments show that it greatly improves the robustness of the model.

Acknowledgments

This work is supported in part by the National Natural Science Funds for Distinguished Young Scholar under Grant 62325307, in part by the National Natural

Science Foundation of China under Grants 62473264, 62073225, 62203134, 62072315 in part by the Natural Science Foundation of Guangdong Province under Grants 2023B1515120038, in part by Shenzhen Science and Technology Innovation Commission (20220809141216003, KCXFZ20230731094001003, JCYJ20220531102817040, KJZD20230923113801004), in part by the Scientific Instrument Developing Project of Shenzhen University under Grant 2023YQ019.

References

Araujo, A.; Meunier, L.; Pinot, R.; and Negrevergne, B. 2019. Robust neural networks using randomized adversarial training. *arXiv preprint arXiv:1903.10219*.

Bortsova, G.; González-Gonzalo, C.; Wetstein, S. C.; Dubost, F.; Katramados, I.; Hogeweg, L.; Liefers, B.; van Ginneken, B.; Pluim, J. P.; Veta, M.; et al. 2021. Adversarial attack vulnerability of medical image analysis systems: Unexplored factors. *Medical Image Analysis*, 73: 102141.

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *Proc. SP*, 39–57. IEEE.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proc. ICCV*, 248–255. IEEE.

Engstrom, L.; Tran, B.; Tsipras, D.; Schmidt, L.; and Madry, A. 2019a. Exploring the landscape of spatial robustness. In *Proc. ICML*, 1802–1811.

Engstrom, L.; Tran, B.; Tsipras, D.; Schmidt, L.; and Madry, A. 2019b. Exploring the landscape of spatial robustness. In *International conference on machine learning*, 1802–1811. PMLR.

Finlayson, S. G.; Bowers, J. D.; Ito, J.; Zittrain, J. L.; Beam, A. L.; and Kohane, I. S. 2019. Adversarial attacks on medical machine learning. *Science*, 363(6433): 1287–1289.

Finlayson, S. G.; Chung, H. W.; Kohane, I. S.; and Beam, A. L. 2018. Adversarial attacks against medical deep learning systems. *arXiv preprint arXiv:1804.05296*.

Gehlot, S.; Gupta, A.; and Gupta, R. 2020. SDCT-AuxNetθ: DCT augmented stain deconvolutional CNN with auxiliary classifier for cancer diagnosis. *Medical image analysis*, 61: 101661.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In Bengio, Y.; and LeCun, Y., eds., *Proc. ICLR*.

Guo, C.; Gardner, J.; You, Y.; Wilson, A. G.; and Weinberger, K. 2019. Simple black-box adversarial attacks. In *Proc. ICML*, 2484–2493. PMLR.

Gupta, A.; Duggal, R.; Gehlot, S.; Gupta, R.; Mangal, A.; Kumar, L.; Thakkar, N.; and Satpathy, D. 2020. GCTI-SN: Geometry-inspired chemical and tissue invariant stain normalization of microscopic medical images. *Medical Image Analysis*, 65: 101788.

Gupta, R.; Mallick, P.; Duggal, R.; Gupta, A.; and Sharma, O. 2017. Stain color normalization and segmentation of plasma cells in microscopic images as a prelude to development of computer assisted automated disease diagnostic tool in multiple myeloma. *Clinical Lymphoma, Myeloma and Leukemia*, 17(1): e99.

Han, T.; Nebelung, S.; Pedersoli, F.; Zimmermann, M.; Schulze-Hagen, M.; Ho, M.; Haarburger, C.; Kiessling, F.; Kuhl, C.; Schulz, V.; et al. 2021. Advancing diagnostic performance and clinical usability of neural networks via adversarial training and dual batch normalization. *Nature communications*, 12(1): 1–11.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proc. ICCV*, 7132–7141.

Hunger, S. P.; and Mullighan, C. G. 2015. Acute lymphoblastic leukemia in children. *New England Journal of Medicine*, 373(16): 1541–1552.

Jones, D. R.; Schonlau, M.; and Welch, W. J. 1998. Efficient global optimization of expensive black-box functions. *Proc. JGO*, 13(4): 455.

Kamath, S.; Deshpande, A.; Kambhampati Venkata, S.; and N Balasubramanian, V. 2021. Can we have it all? On the Trade-off between Spatial and Adversarial Robustness of Neural Networks. In *Proc. NeurIPS*, volume 34.

Kandasamy, K.; Dasarathy, G.; Schneider, J.; and Póczos, B. 2017. Multi-fidelity bayesian optimisation with continuous approximations. In *Proc. ICML*, 1799–1808. PMLR.

Khakzar, A.; Albarqouni, S.; and Navab, N. 2019. Learning interpretable features via adversarially robust optimization. In *Proc. MICCAI*, 793–800. Springer.

Koga, K.; and Takemoto, K. 2022. Simple Black-Box Universal Adversarial Attacks on Deep Neural Networks for Medical Image Classification. *Algorithms*, 15(5): 144.

Kurakin, A.; Goodfellow, I.; and Bengio, S. 2016. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*.

Lamberti, W. F. 2022. Classification of White Blood Cell Leukemia with Low Number of Interpretable and Explainable Features. *Arxiv*.

Liu, A.; Tang, S.; Liu, X.; Chen, X.; Huang, L.; Tu, Z.; Song, D.; and Tao, D. 2020. Towards defending multiple adversarial perturbations via gated batch normalization. *arXiv preprint arXiv:2012.01654*.

Ma, X.; Driggs-Campbell, K. R.; and Kochenderfer, M. J. 2019. Improved Robustness and Safety for Autonomous Vehicle Control with Adversarial Reinforcement Learning. *CoRR*, abs/1903.03642.

Ma, X.; Niu, Y.; Gu, L.; Wang, Y.; Zhao, Y.; Bailey, J.; and Lu, F. 2021. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition*, 110: 107332.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *Proc. ICLR*.

Mao, X.; Chen, Y.; Wang, S.; Su, H.; He, Y.; and Xue, H. 2021. Composite adversarial attacks. In *Proc. AAAI*, volume 35, 8884–8892.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; and Frossard, P. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proc. ICCV*, 2574–2582.

Paschali, M.; Conjeti, S.; Navarro, F.; and Navab, N. 2018. Generalizability vs. robustness: investigating medical imaging networks using adversarial examples. In *Proc. MICCAI*, 493–501. Springer.

Prellberg, J.; and Kramer, O. 2019. Acute lymphoblastic leukemia classification from microscopic images using convolutional neural networks. In *Proc. ISBI*, 53–61. Springer.

Rebuffi, S.-A.; Gowal, S.; Calian, D. A.; Stimberg, F.; Wiles, O.; and Mann, T. 2021. Fixing data augmentation to improve adversarial robustness. *arXiv preprint arXiv:2103.01946*.

Rehman, A.; Abbas, N.; Saba, T.; Rahman, S. I. u.; Mehmood, Z.; and Kolivand, H. 2018. Classification of acute lymphoblastic leukemia using deep learning. *Microscopy Research and Technique*, 81(11): 1310–1317.

Salman, H.; Li, J.; Razenshteyn, I.; Zhang, P.; Zhang, H.; Bubeck, S.; and Yang, G. 2019. Provably robust deep learning via adversarially trained smoothed classifiers. In *Proc. NeurIPS*, volume 32.

Schott, L.; Rauber, J.; Bethge, M.; and Brendel, W. 2018. Towards the first adversarially robust neural network model on MNIST. In *Proc. ICLR*.

Smith, M. A.; Seibel, N. L.; Altekruse, S. F.; Ries, L. A.; Melbert, D. L.; O'Leary, M.; Smith, F. O.; and Reaman, G. H. 2010. Outcomes for children and adolescents with cancer: challenges for the twenty-first century. *Journal of clinical oncology*, 28(15): 2625.

Song, H.; Kim, M.; Park, D.; Shin, Y.; and Lee, J.-G. 2022. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*.

Tai, W.-L.; Hu, R.-M.; Hsiao, H. C.; Chen, R.-M.; and Tsai, J. J. 2011. Blood Cell Image Classification Based on Hierarchical SVM. In *Proc. ISM*, 129–136.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proc. ICML*, 6105–6114. PMLR.

Tramèr, F.; and Boneh, D. 2019. Adversarial training and robustness for multiple perturbations. In *Proc. NeurIPS*, 5866–5876.

Tramèr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I.; Boneh, D.; and McDaniel, P. 2018. Ensemble Adversarial Training: Attacks and Defenses. In *Proc. ICLR*.

Uwimana, A.; and Senanayake, R. 2021. Out of distribution detection and adversarial attacks on deep neural networks for robust medical image analysis. *arXiv preprint arXiv:2107.04882*.

Wang, A.; Islam, M.; Xu, M.; Zhang, Y.; and Ren, H. 2023. SAM Meets Robotic Surgery: An Empirical Study in Robustness Perspective. arXiv:2304.14674.

Williams, C. K.; and Rasmussen, C. E. 2006. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA.

Xie, C.; Tan, M.; Gong, B.; Wang, J.; Yuille, A. L.; and Le, Q. V. 2020. Adversarial examples improve image recognition. In *Proc. CVPR*, 819–828.

Zhang, C.; and Gao, P. 2021. Countering Adversarial Examples: Combining Input Transformation and Noisy Training. In *Proc. ICCV*, 102–111.