Erciyes University Journal of Institute Of Science and Technology

Research Article
10.65520/erciyesfen.1997805

Imprint:
Volume: 42(3)
Year: 2026
Page: 445-466

Tuğba KARATAŞ[a]
Mustafa ETCİL[b]
Burak KOLUKISA[c*]

[a] Undergraduate Student, Kayseri University, karatastugba132@gmail.com
[b] Res. Asst., Abdullah Gul University, mustafa.etcil@agu.edu.tr.
[c] Asst. Prof., Kayseri University, burakkolukisa@kayseri.edu.tr

* Corresponding Author

Received: 7/18/2026
Accepted: 10/19/2026

Citation:
Tuğba KARATAŞ, Mustafa ETCİL, Burak KOLUKISA (2026). Evaluation of Adversarial Robustness in Deep Learning-Based Colorectal Histopathology Classification. *Erciyes University Journal of Institute Of Science and Technology*, 42(3), 445-466. https://doi.org/10.65520/erciyesfen.1997805

Screened by iThenticate for Authors & Researchers

Except where otherwise noted, content in this article is licensed under a Creative Commons 4.0 International license. Icons by Font Awesome.

Evaluation of Adversarial Robustness in Deep Learning-Based Colorectal Histopathology Classification

Abstract

Background/Aim: Deep learning models can achieve high accuracy in colorectal histopathology classification, but their performance may decrease under adversarial perturbations. This study aimed to evaluate the adversarial robustness of a 2D-CNN across two colorectal histopathology datasets under FGSM, PGD, and BIM attacks and to compare adversarial training with convolutional autoencoder (CAE)-based input purification as defense strategies.

Methods: Experiments were conducted on the Kather Texture 2016 and CRC-VAL-HE-7K datasets, containing 5,000 and 7,180 image patches from eight and nine colorectal tissue classes, respectively. A 2D-CNN was trained using five-fold stratified cross-validation and evaluated on held-out test sets. White-box FGSM, PGD, and BIM attacks were applied across multiple perturbation levels. Adversarial training was examined using three clean-to-adversarial training compositions. For input purification, shallow, deep, and mixed-ε deep CAE models were evaluated. Performance was assessed using accuracy, macro precision, macro recall, and macro F1-score.

Results: The 2D-CNN achieved 87.70% and 89.35% clean test accuracy on Kather Texture 2016 and CRC-VAL-HE-7K, respectively, but performance declined markedly under adversarial attacks, particularly on CRC-VAL-HE-7K. Adversarial training substantially improved robustness at ε = 0.01, although the best training composition varied across datasets and attacks. Mixed-ε deep CAE generally provided the strongest purification performance at higher perturbation levels.

Conclusion: Despite its clean-image classification performance, the 2D-CNN remained vulnerable to adversarial perturbations. Both defense strategies generally improved performance under the evaluated adversarial conditions: adversarial training substantially improved performance at ε = 0.01, whereas mixed-ε CAE purification generally showed stronger relative performance at higher perturbation levels.

Keywords: Colorectal histopathology, Deep learning, Adversarial attacks, Adversarial training, Convolutional autoencoder

Derin Öğrenme Tabanlı Kolorektal Histopatoloji Sınıflandırmasında Saldırılara Karşı Dayanıklılığın Değerlendirilmesi

Öz

Arka Plan/Amaç: Derin öğrenme modelleri kolorektal histopatoloji sınıflandırmasında yüksek doğruluk sağlayabilse de adversarial pertürbasyonlar altında performansları azalabilmektedir. Bu çalışmada, bir 2D-CNN modelinin iki kolorektal histopatoloji veri setinde FGSM, PGD ve BIM saldırıları altındaki adversarial dayanıklılığının değerlendirilmesi ve adversarial eğitim ile evrişimsel otoenkoder (CAE) tabanlı girdi arındırma

yöntemlerinin savunma stratejileri olarak karşılaştırılması amaçlanmıştır.

Yöntemler: Deneyler, sırasıyla sekiz ve dokuz kolorektal doku sınıfına ait 5.000 ve 7.180 RGB görüntü yaması içeren Kather Texture 2016 ve CRC-VAL-HE-7K veri setleri üzerinde gerçekleştirilmiştir. Bir 2D-CNN modeli beş katlı tabakalı çapraz doğrulama kullanılarak eğitilmiş ve eğitimden ayrı tutulan test setlerinde değerlendirilmiştir. Beyaz kutu FGSM, PGD ve BIM saldırıları birden fazla pertürbasyon düzeyinde uygulanmıştır. Adversarial eğitim, üç farklı temiz-adversarial eğitim verisi bileşimi kullanılarak incelenmiştir. Girdi arındırma için sığ, derin ve karma-ε derin CAE modelleri değerlendirilmiştir. Performans; doğruluk, makro kesinlik, makro duyarlılık ve makro F1-skoru kullanılarak değerlendirilmiştir.

Bulgular: 2D-CNN modeli, Kather Texture 2016 ve CRC-VAL-HE-7K veri setlerinde temiz test görüntüleri üzerinde sırasıyla %87,70 ve %89,35 doğruluk elde etmiş; ancak adversarial saldırılar altında, özellikle CRC-VAL-HE-7K veri setinde, performans belirgin şekilde azalmıştır. Adversarial eğitim, ε=0,01 düzeyinde dayanıklılığı önemli ölçüde artırmış olmakla birlikte, en iyi eğitim bileşimi veri setine ve saldırı türüne göre değişiklik göstermiştir. Karma-ε derin CAE modeli ise genel olarak daha yüksek pertürbasyon düzeylerinde en güçlü arındırma performansını sağlamıştır.

Sonuç: 2D-CNN, temiz görüntülerdeki sınıflandırma performansına rağmen adversarial pertürbasyonlara karşı savunmasız kalmıştır. Her iki savunma stratejisi de değerlendirilen adversarial koşullar altında genel olarak performansı artırmıştır: adversarial eğitim ε = 0.01 düzeyinde performansı önemli ölçüde iyileştirirken, karma-ε CAE tabanlı girdi arındırma genel olarak daha yüksek pertürbasyon düzeylerinde daha güçlü göreli performans göstermiştir.

Anahtar kelimeler: Kolorektal histopatoloji, Derin öğrenme, Adversarial saldırılar, Adversarial eğitim, Evrişimsel otoenkoder

1. INTRODUCTION

Colorectal cancer (CRC) is a malignant neoplasm arising from the glandular epithelial cells of the colonic or rectal mucosa. It is the third most commonly diagnosed cancer and the second leading cause of cancer-related mortality worldwide [1]. In 2022, approximately 1.9 million new cases and 904,000 deaths were reported, accounting for nearly 10% of all cancer diagnoses and cancer-related deaths globally [2]. CRC occurs most frequently among individuals aged 65–74 years [3]. A substantial proportion of patients are diagnosed at advanced stages, contributing to increased mortality and disease-related morbidity [4]. Although the five-year survival rate is approximately 90% among patients diagnosed at an early stage, it declines to nearly 13% in those with advanced-stage disease [5]. Therefore, early detection and accurate diagnosis of CRC are critically important for the timely initiation of appropriate treatment and improvement of patient prognosis [6].

The primary methods used for CRC diagnosis are colonoscopy and histopathological examination [6]. Although colonoscopy enables the detection and sampling of suspicious lesions, small, flat, or sessile lesions may be overlooked because of their subtle appearance [7]. Histopathological examination of biopsy or resection specimens remains the reference standard for definitive diagnosis [8]. During histopathological assessment, pathologists microscopically examine tissue specimens to identify neoplastic alterations. However, this process is time-consuming and susceptible to both intraobserver and interobserver variability [9]. In recent years, artificial intelligence (AI)-based diagnostic approaches have emerged as valuable tools for supporting clinical decision-making because of their potential to improve diagnostic accuracy and automate the evaluation process [10]. These technologies may assist pathologists and contribute to the standardization of diagnostic workflows [11, 12]. Accordingly, the development of computer-aided diagnosis (CAD) systems for CRC has attracted increasing research interest.

With the growing interest in CAD systems, the use of deep learning (DL)-based methods in medical image analysis has become increasingly widespread [13, 14, 15]. Owing to their ability to learn hierarchical feature representations directly from raw images in an end-to-end manner, DL algorithms offer considerable flexibility [16] and facilitate faster and more efficient diagnostic workflows [17]. Consequently, several DL approaches have been proposed for colorectal histopathology classification, including lightweight convolutional neural network (CNN) architectures [18], ensembles of adapted pretrained CNNs [19], and attention-based CNN models [20], all of which have demonstrated

CC: BY-NC-ND

promising performance on clean histopathological images. More recently, transformer-based representation learning and large-scale pathology foundation models have extended this landscape, providing transferable feature representations to a variety of histopathological tasks [21, 22].

However, high classification performance under conventional evaluation settings does not necessarily imply robustness. DL models remain vulnerable to carefully crafted adversarial examples that can induce incorrect predictions through visually imperceptible perturbations [23, 24, 25]. Consequently, robustness evaluation has moved from single-step attacks to stronger iterative attacks such as PGD and BIM, as well as standardized attack ensembles such as AutoAttack, which were introduced to provide more reliable estimates of adversarial robustness [26]. Such vulnerability is of particular concern in medical image analysis, where diagnostic accuracy is critical and prediction errors may affect clinical decision-making [27, 28, 29]. Therefore, evaluating the response of AI systems to adversarial perturbations is essential for assessing their reliability in clinical applications [30]. In this context, adversarial vulnerability has been studied for different medical imaging tasks [30]. CNN-based classifiers have been shown to be vulnerable to white-box and black-box attacks in computational pathology [31], and adversarial perturbations have been used to manipulate pathology vision-language models on colorectal histopathology images [32]. Besides adversarial perturbations, the robustness of computational pathology is also influenced by domain shifts arising from changes in image acquisition and data distributions. Thus, uncertainty estimation is also a way to study unreliable predictions in such scenarios [33]. Previous defense-oriented studies have examined adversarial training [34, 35] and reconstruction-based input purification [36, 37]. Together, these studies motivate a controlled evaluation of classifier-level and input-level defense strategies within a common colorectal histopathology classification framework.

Accordingly, this study evaluates the adversarial robustness of a two-dimensional convolutional neural network (2D-CNN) for colorectal histopathological image classification using two benchmark datasets, Kather Texture 2016 and CRC-VAL-HE-7K. For each dataset, the classifier is assessed against three white-box attacks—the Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and the Basic Iterative Method (BIM)—across multiple perturbation magnitudes. Two complementary defense strategies are investigated within a common experimental framework: adversarial training as a classifier-level defense and convolutional autoencoder (CAE)-based input purification as an input-level defense. Adversarially trained 2D-CNN models with different clean-to-adversarial training ratios and CAE-based purification models trained using single or multiple perturbation levels are evaluated under both clean and adversarial test conditions. These experiments are designed to characterize the vulnerability of colorectal histopathology classifiers to adversarial perturbations and to assess how effectively classifier-level and input-level defenses preserve tissue-classification performance under attack.

The remainder of this paper is organized as follows. Section 2 describes the dataset, 2D-CNN architecture, threat model, and defense strategies. Section 3 presents the experimental protocol. Section 4 reports the classification and adversarial robustness results. Finally, Section 5 discusses the findings, addresses the limitations of the study, and presents the conclusions.

2. MATERIALS AND METHODS

This section presents the methodological framework used to evaluate the adversarial robustness of a two-dimensional convolutional neural network for colorectal histopathological tissue classification. The following subsections describe the datasets and preprocessing procedures, the 2D-CNN classifier, the white-box adversarial threat model and the generation of FGSM, PGD, and BIM attacks, the adversarial training and convolutional autoencoder-based input purification strategies, and the statistical and experimental evaluation protocols. An overview of the complete workflow is provided in Fig. 1.

CC: BY-NC-ND

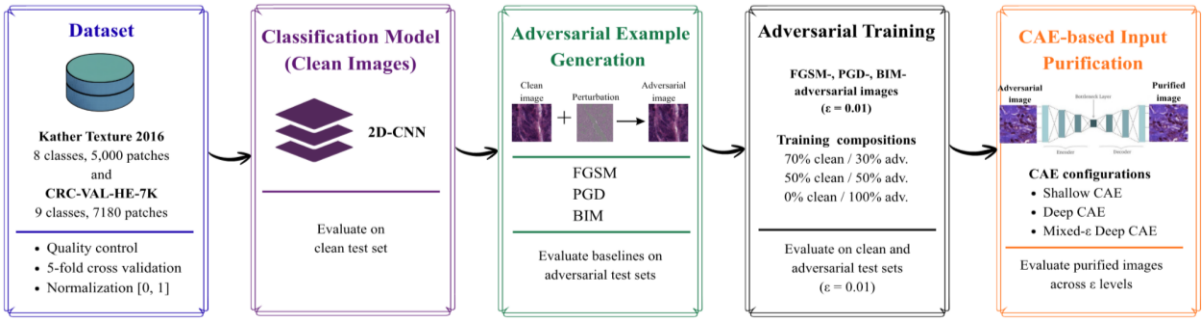


Figure 1. Overall workflow of the proposed framework for evaluating adversarial robustness in deep learning-based colorectal histopathology classification.

2.1. Dataset Description

The Kather Texture 2016 dataset [9] was used for model development and robustness assessment. It contains 5,000 image patches of size 150 × 150 pixels acquired using an Aperio ScanScope digital slide scanner (Leica Biosystems) at a spatial resolution of 0.495 µm/pixel. The dataset is evenly distributed across eight colorectal tissue classes, with 625 images per class: tumor epithelium, simple stroma, complex stroma, immune cell conglomerates, debris, mucosal glands, adipose tissue, and background. Representative examples are shown in Fig. 2.

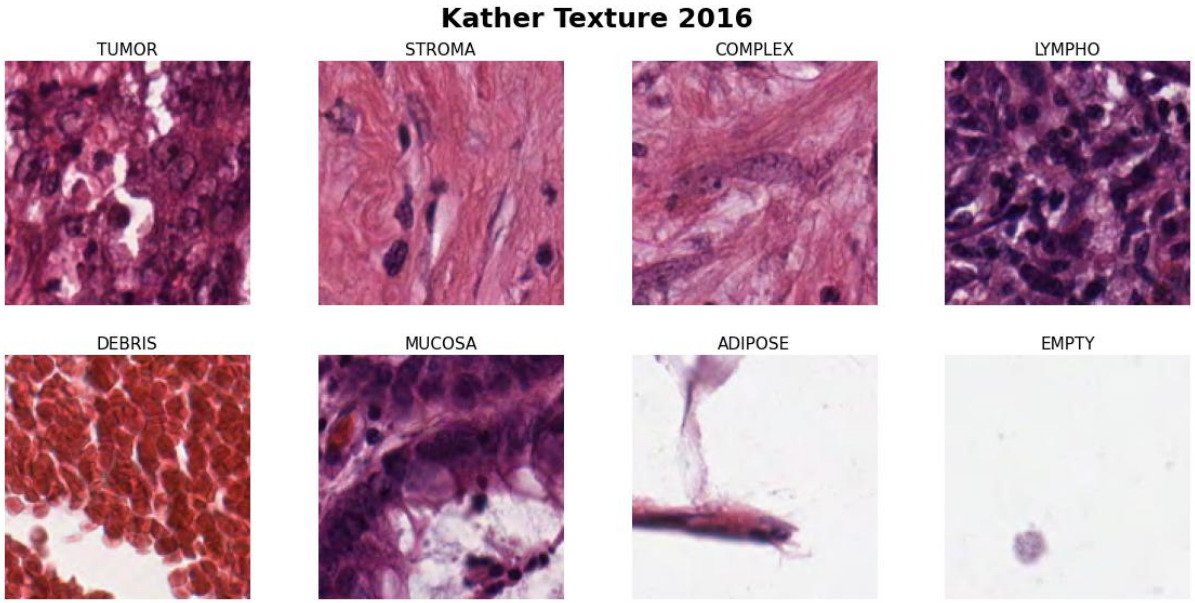


Figure 2. Representative histopathological image patches from the eight tissue classes in the Kather Texture 2016 dataset [9].

CRC-VAL-HE-7K [8] was used as a second benchmark dataset for additional evaluation. It contains 7,180 H&E-stained image patches of size 224 × 224 pixels at 0.5 µm/pixel across nine tissue classes: adipose tissue (ADI, n = 1,338), background (BACK, n = 847), debris (DEB, n = 339), lymphocytes (LYM, n = 634), mucus (MUC, n = 1,035), smooth muscle (MUS, n = 592), normal colon mucosa (NORM, n = 741), cancer-associated stroma (STR, n = 421), and colorectal adenocarcinoma epithelium (TUM, n = 1,233). Examples from the CRC-VAL-HE-7K tissue classes are shown in Fig. 3.

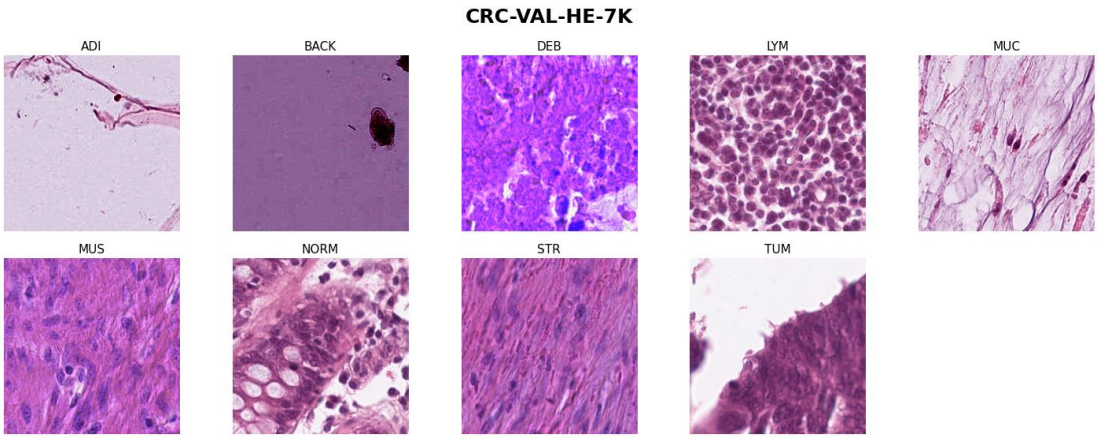


Figure 3. Representative histopathological image patches from the nine tissue classes in the CRC-VAL-HE-7K dataset [8].

2.2. Two-Dimensional Convolutional Neural Network Classifier

A two-dimensional convolutional neural network (2D-CNN) was designed for the classification of colorectal histopathological tissue images. The model received normalized RGB image patches of size 150×150×3 for the Kather Texture 2016 dataset and 224×224×3 for the CRC-VAL-HE-7K dataset, while retaining their two-dimensional spatial organization and relevant local morphological information. CNNs employ local receptive fields and shared weights to efficiently learn hierarchical spatial features from image data [38].

The feature extraction stage comprised four consecutive convolutional blocks. Each block consisted of a Conv2D layer followed by batch normalization, ReLU activation, and max-pooling. The number of filters progressively increased from 32 to 64, 128, and 256. This hierarchical design was intended to capture basic visual patterns in the early layers and increasingly complex tissue-related features in the deeper layers [39]. The output feature maps were then flattened and passed to a dense layer with 256 neurons. Batch normalization and ReLU activation were applied to this layer, followed by dropout at a rate of 0.50 to reduce overfitting. The final classification layer used softmax activation, with eight output neurons for the Kather Texture 2016 dataset and nine output neurons for the CRC-VAL-HE-7K dataset, corresponding to the number of tissue classes in each dataset.

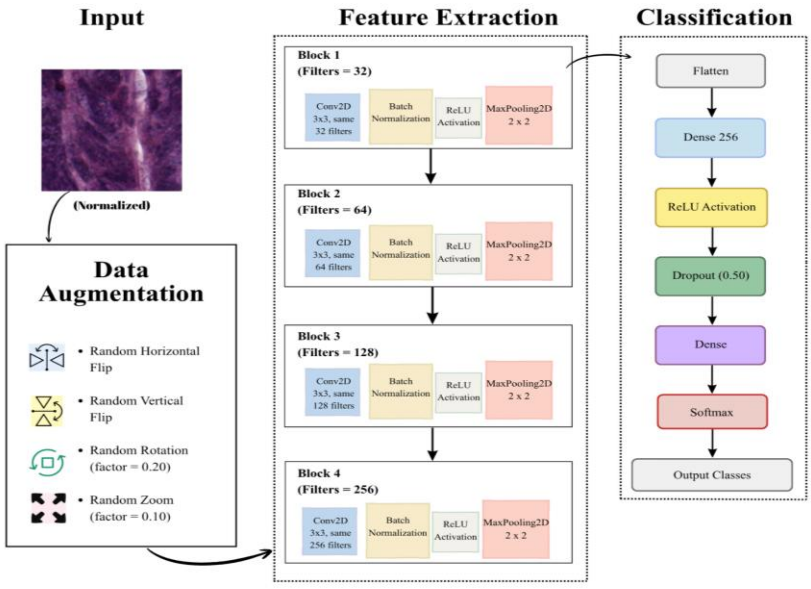


Figure 4. Architecture of the developed 2D-CNN model.

Data augmentation was incorporated into the model as an in-model preprocessing pipeline to improve generalization. The augmentation sequence included random horizontal and vertical flipping, random rotation with a factor of 0.20, and random zoom with a factor of 0.10. These transformations increased the variability of the training data and were intended to reduce the model's sensitivity to changes in image orientation and scale [40]. The augmentation operations were applied only during training and were disabled during validation and testing.

The model was implemented using TensorFlow/Keras and trained for 100 epochs with a batch size of 32. Early stopping was applied by monitoring the validation loss with a patience of 20 epochs, and the weights corresponding to the lowest validation loss were restored. The Adam optimizer was used with a learning rate of 1×10^{-4} , and sparse categorical cross-entropy was employed as the loss function. The main architectural configuration and training settings are summarized in Fig. 4.

2.3. Adversarial Threat Model and Attack Generation

The adversarial robustness of the proposed 2D-CNN classifier was evaluated under a white-box threat model. In this setting, the attacker was assumed to have full access to the target model, including its architecture, trained parameters, loss function, and input gradients [30]. Unlike black-box attacks, which rely on model outputs, predicted labels, or surrogate models, white-box attacks directly exploit the gradient information of the target model to generate adversarial perturbations [28]. Three gradient-based attacks were considered: FGSM, PGD, and BIM. This combination enabled evaluation under both single-step and iterative attack settings. In this study, the white-box threat model refers specifically to adversarial example generation against the target 2D-CNN classifier; the CAE was not included in the attack-generation gradient path.

FGSM generates an adversarial example in a single step by perturbing the input in the direction of the sign of the loss gradient [25]. Given an input image (x) with ground-truth label (y), the adversarial example x_{adv} is generated by perturbing the input in the direction that maximally increases the classification loss:

$$x_{adv} = clip_{[0,1]}(x + \varepsilon.sign(\nabla_x J(\theta, x, y))), \quad (1)$$

where x_{adv} denotes the adversarial example, θ represents the model parameters, $J(\theta, x, y)$ is the classification loss, $\nabla_x J(\theta, x, y)$ denotes the gradient of the loss with respect to the input image, and ε controls the perturbation magnitude. The clipping operation constrains pixel values to the normalized input range of [0,1]. Adversarial examples were generated using multiple perturbation magnitudes (ε) to evaluate model robustness under increasing attack strengths.

BIM [41] generalizes the FGSM formulation by taking multiple smaller perturbation steps instead of a single update. The adversarial example is iteratively updated from the original input image in the direction of the sign of the loss gradient, while bounding the total perturbation within the specified ε-bounded neighborhood of the original image. The iterative update is given by:

$$x_{adv}^{(t+1)} = clip_{x,\varepsilon}(x_{adv}^{(t)} + \alpha.sign(\nabla_x J(\theta, x_{adv}^{(t)}, y))), \quad (2)$$

where $x_{adv}^{(0)} = x$, t denotes the iteration index, and α represents the step size. The clipping operation constrains the adversarial image to remain within the ε-bounded neighborhood of the original image and within the valid normalized pixel range of [0,1]. In this study, BIM was applied for 10 iterations, with the step size defined as $\alpha = \varepsilon/10$.

Similarly, PGD [34] performs iterative gradient-based updates but, unlike BIM, starts from a randomly perturbed point within the ε-bounded region. The first adversarial image is defined as:

$$\delta \sim u(-\varepsilon, \varepsilon), \quad (3)$$

$$x_{adv}^{(0)} = clip_{[0,1]}(x + \delta), \quad (4)$$

where δ denotes a random perturbation sampled uniformly from the interval $[-\varepsilon, \varepsilon]$. Following this random initialization, PGD applies the same iterative update rule given in Eq. (2). As with BIM,

CC: BY-NC-ND

PGD was performed for 10 iterations with a step size of $\alpha = \varepsilon/10$, while constraining the adversarial examples to remain within the ε-bounded neighborhood of the original image and the valid normalized pixel range of [0,1].

2.4. Adversarial Defense Mechanisms

Two complementary defense strategies were investigated within the same 2D-CNN classification framework: adversarial training and CAE-based input purification. Adversarial training modifies the classifier by incorporating adversarially perturbed images into the training data. In contrast, CAE-based purification processes perturbed images before classification with the aim of reducing the effects of adversarial perturbations.

2.4.1 Adversarial Training

Adversarial training is a defense strategy in which adversarially perturbed samples are incorporated into the training process to improve model robustness against adversarial attacks [34]. In this study, adversarial training was performed separately using FGSM-, PGD-, and BIM-generated training images, while preserving their ground-truth tissue-class labels.

First, the 2D-CNN trained using only clean training images was evaluated on the clean test set and on adversarial test images generated using FGSM, PGD, and BIM at perturbation magnitudes of ε = 0.01, 0.03, 0.05, 0.08, 0.10, 0.15, 0.20, and 0.30. This evaluation was performed to examine the change in classification performance as the perturbation magnitude increased.

For the adversarial training experiments, perturbed versions of the training images were generated using FGSM, PGD, and BIM with ε = 0.01. The level of perturbation was chosen to result in a small input distortion, but one which led to a measurable performance degradation of the undefended classifier, so as to provide a practical setting for evaluation of the impact of adversarial training. For each attack type, three separate training-data compositions were constructed: 70% clean and 30% adversarial images, 50% clean and 50% adversarial images, and 100% adversarial images. The total number of training samples was kept constant across all three compositions. Adversarial samples replaced the corresponding proportion of clean samples rather than being added to the original training set.

A separate 2D-CNN was trained for each attack type and training-data composition using the same architecture and optimization settings described in Section 2.2. The resulting models were evaluated on the clean test set and on adversarial test images generated with the corresponding attack at ε=0.01. For robustness evaluation, adversarial test examples were generated separately against each corresponding adversarially trained 2D-CNN model. Clean and adversarial classification performance was then compared across the different training compositions.

2.4.2 Convolutional Autoencoder-Based Input Purification

A CAE was used as an input-level defense to reduce the effect of adversarial perturbations before classification [36, 37]. The CAE consisted of an encoder and a decoder. The encoder transformed each input image into a lower-dimensional latent representation, whereas the decoder reconstructed an output image from this representation. During training, histopathological image patches perturbed using FGSM, PGD, or BIM were used as inputs, and the corresponding clean image patches were used as reconstruction targets. Thus, the CAE was optimized to reconstruct clean image content from adversarially perturbed inputs. After reconstruction, the output images were passed to the pretrained 2D-CNN for tissue classification. To assess the effect of the purification stage on unperturbed inputs, the shallow, deep, and mixed-ε CAE pipelines were also evaluated on the clean held-out test sets. In these experiments, adversarial examples were generated against the target classifier and then passed through the CAE for purification. Accordingly, the CAE evaluation reflects purification performance for classifier-targeted adversarial examples rather than robustness against adaptive white-box attacks jointly targeting the complete CAE–classifier pipeline.

CC: BY-NC-ND

Table 1. Architectural configurations of the CAE purification models.

Component	Shallow CAE	Deep CAE
Encoder	2 × (Conv2D (32, 64) → MaxPooling2D)	2 × (Conv2D (32, 64) → Batch Normalization → LeakyReLU → MaxPooling2D)
Bottleneck	Conv2D (128) → Conv2D (64)	Conv2D (128) → Batch Normalization → LeakyReLU → Conv2D (64) → Batch Normalization → LeakyReLU
Decoder	2 × (UpSampling2D → Conv2D (64, 32) → Cropping2D)	2 × (UpSampling2D → Conv2D (64, 32) → Batch Normalization → LeakyReLU → Cropping2D)
Output layer	Conv2D (3 channels)	Conv2D (3 channels)
Trainable parameters	186,371	187,651

Note: The mixed-ε CAE used the deep CAE architecture and differed only in the perturbation levels included in its training data.

Three CAE configurations were evaluated separately for each attack type: a shallow CAE trained with adversarial images generated at ε = 0.01, a deep CAE trained at the same perturbation magnitude, and a deep CAE trained using a mixed-ε dataset containing adversarial images generated at ε = 0.01, 0.03, 0.05, 0.08, and 0.10. For evaluation, all CAE configurations were tested at ε = 0.01, 0.03, 0.05, 0.08, 0.10, 0.15, 0.20, and 0.30 under the corresponding attack. All CAE models were optimized using the Adam optimizer with a learning rate of 1×10^{-3} and mean squared error (MSE) as the reconstruction loss. Models were trained for 30 epochs with a batch size of 32, while 20% of the training image pairs were reserved for validation. Early stopping was applied by monitoring the validation loss with a patience of 5 epochs, and the weights corresponding to the lowest validation loss were restored. The mixed-ε model used the same deep CAE architecture and training settings as the single-ε deep CAE; the only difference was the range of perturbation magnitudes included in its training data. The architectural configurations of the shallow and deep CAE models are summarized in Table 1.

3. EXPERIMENTAL SETUP

The experimental study was carried out on the Kather Texture 2016 and CRC-VAL-HE-7K datasets. Kather Texture 2016 consists of 5,000 RGB image patches of size 150 × 150 pixels for eight colorectal tissue classes and CRC-VAL-HE-7K contains 7,180 RGB image patches of size 224 × 224 pixels for nine tissue classes. Pixel values in both datasets were normalized to the [0,1] range. Each dataset was processed and evaluated separately. The datasets were split into 80% training and 20% test sets using stratified sampling to preserve class distribution. The test set was used exclusively for final classification and adversarial robustness evaluation. For each dataset, the training set was subjected to five-fold stratified cross-validation in a single run of cross-validation with a random seed of 42 and shuffling prior to fold generation. Data augmentation was applied only to the training portion within each fold; horizontal and vertical flipping, rotation with a factor of 0.20, and zooming with a factor of 0.10 were employed. Performance was recorded separately for each fold and summarized using the mean and standard deviation across the five folds. The best epoch was identified within each fold based on validation loss, and the median of these best-epoch values was used to train a final model on the entire training set. This final model was then evaluated once on the corresponding held-out test set. All experiments were conducted in Google Colab using an NVIDIA A100 GPU.

Adversarial attacks were performed using FGSM, PGD, and BIM under a white-box threat model. The vulnerability of the baseline 2D-CNN model to attacks was evaluated at perturbation levels of ε = 0.01, 0.03, 0.05, 0.08, 0.10, 0.15, 0.20, and 0.30. PGD and BIM were implemented using 10 iterative steps, with the step size set to $\alpha = \varepsilon/10$.

As the first defense approach, adversarial training was applied. In these experiments, samples generated separately using FGSM, PGD, and BIM with ε = 0.01 were used, and three different training

CC: BY-NC-ND

compositions were examined: 70% clean–30% adversarial, 50% clean–50% adversarial, and 100% adversarial images. As the second defense approach, convolutional autoencoder (CAE)-based input purification was applied. In this context, shallow CAE, deep CAE, and mixed-ε deep CAE configurations were evaluated separately against FGSM, PGD, and BIM. The single-ε CAE models were trained using adversarial examples generated at ε = 0.01, whereas the mixed-ε deep CAE models incorporated perturbation levels of ε = 0.01, 0.03, 0.05, 0.08, and 0.10. For evaluation, all CAE configurations were tested at ε = 0.01, 0.03, 0.05, 0.08, 0.10, 0.15, 0.20, and 0.30 under the corresponding attack.

Model performance was evaluated using accuracy (ACC), macro precision (M-Pre), macro recall (M-Rec), and macro F1-score (M-F1). These metrics are defined in Equations (5)–(8), respectively. Macro-averaged metrics assign equal weight to all classes regardless of class size. Here, N denotes the number of test samples, y_i and $\hat{y}_i$ denote the true and predicted labels of the i-th sample, respectively, and $\mathbb{I}(\cdot)$ denotes the indicator function. C denotes the number of classes, while P_c, R_c, and $F1_c$ denote the precision, recall, and F1-score computed for the c-th class, respectively. In this study, C=8 for Kather Texture 2016 and C=9 for CRC-VAL-HE-7K. Fold-based results were summarized using mean ± standard deviation, and selected model comparisons were further assessed using McNemar's test and paired bootstrap 95% confidence intervals. Macro-F1 differences were calculated between the defended and undefended models, and paired bootstrap resampling was used to obtain percentile-based 95% confidence intervals for these differences.

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N} I(y_i = \hat{y}_i) \tag{5}$$

$$Pmacro = \frac{1}{c}\sum_{i=1}^{c} P_c \tag{6}$$

$$Rmacro = \frac{1}{c}\sum_{\mathrm{i}=1}^{\mathrm{c}} \mathrm{R_c} \tag{7}$$

$$F1_{macro} = \frac{1}{c}\sum_{\mathrm{i}=1}^{\mathrm{c}} \mathrm{F1_c} \tag{8}$$

4. EXPERIMENTAL RESULTS

This section presents the experimental findings for the baseline 2D-CNN and the two defense strategies across the two colorectal histopathology datasets. First, the clean-image classification performance of the 2D-CNN is reported using five-fold cross-validation and the corresponding test sets. Next, the model's vulnerability to white-box FGSM, PGD, and BIM attacks is examined across multiple perturbation magnitudes. The effects of adversarial training are then evaluated under three clean-to-adversarial training compositions for each attack type. Finally, the performance of shallow, deep, and mixed-ε CAE-based input purification methods is compared with that of the undefended 2D-CNN under the corresponding adversarial conditions.

4.1. Baseline Classification Performance

Tables 2 and 3 show the classification performance of the 2D-CNN on the Kather Texture 2016 and CRC-VAL-HE-7K datasets for clean images, respectively. For Kather Texture 2016, the mean validation accuracy across the five folds was 85.46%, with a mean macro F1-score of 85.21%. For CRC-VAL-HE-7K, the mean validation accuracy and macro F1-score across the five folds were 95.58% and 94.47%, respectively. The smaller standard deviations observed for CRC-VAL-HE-7K indicate lower performance variability across the folds. The final 2D-CNN model achieved an accuracy of 87.70%, a macro precision of 87.79%, a macro recall of 87.70%, and a macro F1-score of 87.66% on the held-out 20% test set of Kather Texture 2016. On the held-out CRC-VAL-HE-7K test set, the corresponding values were 89.35%, 90.76%, 84.55%, and 84.28%, respectively.

CC: BY-NC-ND

Table 2. Training and validation performance (%) of the 2D-CNN across five-fold stratified cross-validation on the training set of the Kather Texture 2016 dataset.

Model	Train ACC	Validation ACC	Validation Macro-P	Validation Macro-R	Validation Macro-F1
Mean	87.44 ± 6.53	85.46 ± 6.31	86.63 ± 5.12	85.46 ± 6.31	85.21 ± 6.69
Fold1	80.78	79.60	82.18	79.60	79.77
Fold2	92.53	88.60	89.00	88.60	88.71
Fold3	79.97	77.90	80.39	77.90	76.52
Fold4	93.34	92.20	92.51	92.20	92.17
Fold5	90.56	89.00	89.08	89.00	88.88

Table 3. Training and validation performance (%) of the 2D-CNN across five-fold stratified cross-validation on the training set of the CRC-VAL-HE-7K dataset.

Model	Train ACC	Validation ACC	Validation Macro-P	Validation Macro-R	Validation Macro-F1
Mean	96.11 ± 0.77	95.58 ± 0.66	95.37 ± 1.04	94.23 ± 0.93	94.47 ± 1.18
Fold1	97.13	96.52	96.55	94.85	95.57
Fold2	96.24	94.69	93.95	93.08	92.99
Fold3	96.32	95.39	95.04	94.53	94.72
Fold4	95.02	95.56	95.10	93.45	93.50
Fold5	95.87	95.73	96.22	95.24	95.57

The class-wise performance of the 2D-CNN on the clean test sets is shown in Fig. 5. The model classified 119 of 125 images of tumor epithelium, 117 of 125 images of adipose tissue and 124 of 125 images of background correctly on the Kather Texture 2016 dataset. Most of the misclassifications occurred in simple stroma, complex stroma and debris. Thus, this pattern suggests that tissue categories with similar morphological and textural properties are more difficult to separate. Similarly, high class-specific performance was obtained for several tissue categories on the CRC-VAL-HE-7K dataset, including background (169/169), lymphocytes (121/127), smooth muscle (118/118), and tumor epithelium (241/247). The most prominent confusion occurred for stroma, with 43 stroma samples classified as smooth muscle. Normal mucosa was most frequently misclassified as tumor epithelium (19 samples), whereas tumor epithelium showed relatively few errors, mainly involving smooth muscle (5 samples). Overall, the confusion matrices indicate strong clean-image classification performance on both datasets, although the patterns of class-specific errors differed between the two datasets.

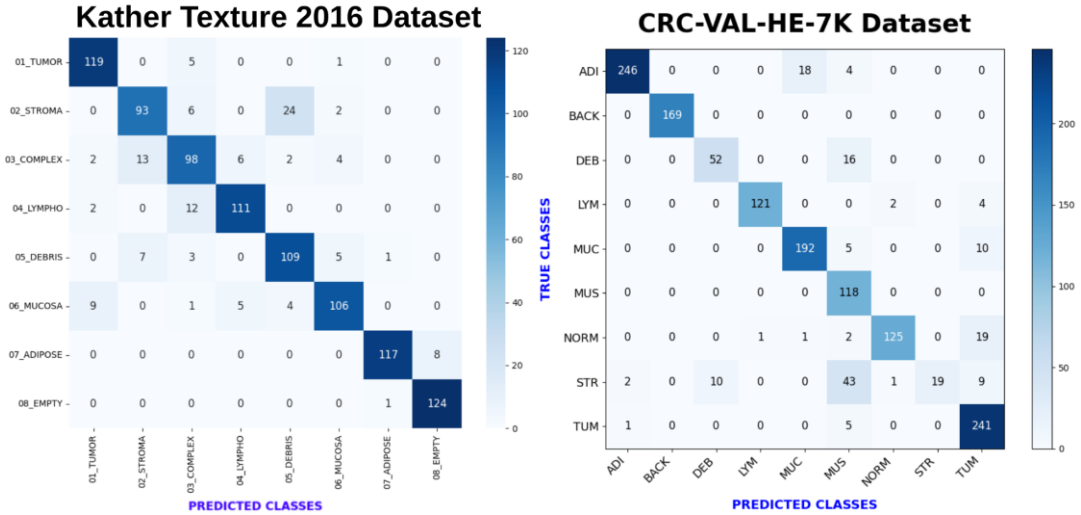


Figure 5. Confusion matrices of the 2D-CNN evaluated on the clean held-out test sets of the Kather Texture 2016 and CRC-VAL-HE-7K datasets.

CC: BY-NC-ND

4.2. Robustness Under Adversarial Attacks

The adversarial robustness of the baseline 2D-CNN was evaluated under white-box FGSM, PGD, and BIM attacks at eight perturbation levels, as summarized in Table 4. On the Kather Texture 2016 dataset, the model achieved 87.70% accuracy on clean images. Under FGSM, accuracy decreased from 80.40% at ε=0.01 to 13.50% at ε=0.30. A similar, relatively gradual degradation pattern was observed under the iterative attacks, with accuracy decreasing from 81.30% to 17.90% under PGD and from 80.80% to 18.00% under BIM over the same perturbation range. In contrast, although the model achieved a clean accuracy of 89.35% on CRC-VAL-HE-7K, it exhibited markedly greater adversarial vulnerability, with a much earlier collapse in performance even at low perturbation levels. At ε=0.01, accuracy decreased to 18.11% under FGSM, 14.76% under PGD, and 13.23% under BIM, corresponding to reductions of 71.24, 74.59, and 76.12 percentage points relative to clean performance, respectively. Performance deteriorated further as ε increased; at ε=0.03, accuracy decreased to 12.59% under FGSM, 0.00% under PGD, and 0.14% under BIM. PGD and BIM reduced accuracy to 0% from ε=0.05 onward, while FGSM declined to 0.63% at ε=0.10 and remained below 1% at higher perturbation levels, reaching 0.21% at ε=0.30. Overall, adversarial performance deteriorated with increasing perturbation magnitude in both datasets, but the loss of robustness was substantially more pronounced on CRC-VAL-HE-7K. The results also indicate that the effect of attack type differed across datasets and that higher clean-image classification accuracy does not imply better adversarial robustness. Accordingly, adversarial training and CAE-based input purification were subsequently evaluated as defense strategies.

Table 4. Classification accuracy (%) of the baseline 2D-CNN under FGSM, PGD, and BIM attacks at different perturbation magnitudes on the Kather Texture 2016 and CRC-VAL-HE-7K datasets.

Dataset	**Attack**	**Clean**	**ε=0.01**	**ε=0.03**	**ε=0.05**	**ε=0.08**	**ε=0.10**	**ε=0.15**	**ε=0.20**	**ε=0.30**
Kather Texture 2016	FGSM	87.70	80.40	64.20	50.80	41.50	33.20	24.20	22.00	13.50
	PGD		81.30	66.50	54.20	44.70	38.30	25.60	19.80	17.90
	BIM		80.80	67.90	53.70	41.50	36.60	29.10	23.70	18.00
CRC-VAL-HE-7K	FGSM	89.35	18.11	12.59	7.66	0.91	0.63	0.14	0.00	0.21
	PGD		14.76	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	BIM		13.23	0.14	0.00	0.00	0.00	0.00	0.00	0.00

4.3. Adversarial Training Results

Table 5 summarizes the clean and adversarial performance of the adversarially trained 2D-CNN models. Clean ACC and Clean M-F1 represent performance on the unperturbed test images, whereas Adversarial ACC and Adversarial M-F1 represent performance on test images perturbed at ε=0.01 using the corresponding attack. The final column reports the absolute improvement in adversarial accuracy over the corresponding undefended 2D-CNN baseline in percentage points. Statistical comparisons with the undefended baseline are presented in Table 6 using McNemar tests for paired classification outcomes and paired bootstrap 95% confidence intervals for differences in macro-F1.

On the Kather Texture 2016 dataset, adversarial training improved robustness to FGSM, PGD, and BIM at the evaluated perturbation level (ε = 0.01), while largely preserving clean-image performance. The 50% clean–50% adversarial configuration was the best-performing setting for FGSM, achieving 92.00% clean accuracy and 91.10% adversarial accuracy, corresponding to an improvement of 10.70 percentage points over the undefended model. For PGD, the highest clean and adversarial accuracies, 89.90% and 88.40%, respectively, were achieved with the 70% clean–30% adversarial configuration, corresponding to a 7.10 percentage-point improvement in adversarial accuracy. Under BIM, the 50% clean–50% adversarial configuration produced the highest adversarial accuracy of 88.80%, representing an improvement of 8.00 percentage points over the undefended baseline. Most improvements on Kather Texture 2016 were statistically significant; however, the PGD 0–100 configuration did not reach statistical significance (McNemar p = 0.138), and its bootstrap 95% confidence interval for the macro-F1 difference included zero (−0.91 to +3.76 percentage points).

CC: BY-NC-ND

On CRC-VAL-HE-7K, the robustness gains were substantially larger. Under FGSM, the 100% adversarial configuration achieved the highest adversarial accuracy of 80.01% and the largest improvement over the undefended baseline (+61.91 percentage points), whereas the 70% clean–30% adversarial configuration achieved the highest clean-image accuracy at 95.47%. For PGD, the 70% clean–30% adversarial configuration achieved the highest adversarial accuracy of 75.84%, corresponding to an improvement of 61.07 percentage points over the undefended baseline, while the 100% adversarial configuration achieved a nearly identical adversarial accuracy of 75.63% with a gain of 60.86 percentage points.

Table 5. Clean and adversarial performance (%) of the 2D-CNN under different adversarial-training compositions.

Dataset	Attack	Training composition	Clean ACC	Clean M-F1	Adversarial ACC	Adversarial M-F1	Adversarial ACC vs. undefended 2D-CNN
Kather Texture 2016	FGSM	70-30	90.50	90.43	89.30	89.28	+8.90
		50-50	92.00	92.00	91.10	91.11	+10.70
		0-100	88.55	88.55	90.10	90.06	+9.70
	PGD	70-30	89.90	89.93	88.40	88.43	+7.10
		50-50	88.40	88.33	87.70	87.62	+6.40
		0-100	86.80	86.33	86.30	85.81	+5.00
	BIM	70-30	89.20	89.19	88.20	88.19	+7.40
		50-50	88.90	88.97	88.80	88.88	+8.00
		0-100	86.60	86.47	86.20	86.09	+5.40
CRC-VAL-HE-7K	FGSM	70-30	95.47	95.07	67.69	64.24	+49.58
		50-50	91.02	89.39	74.16	65.84	+56.06
		0-100	93.25	92.61	80.01	78.03	+61.91
	PGD	70-30	91.64	85.05	75.84	64.78	+61.07
		50-50	85.10	83.49	67.06	62.01	+52.30
		0-100	89.83	86.22	75.63	67.23	+60.86
	BIM	70-30	91.85	90.96	66.57	61.05	+53.34
		50-50	96.94	95.97	84.75	81.36	+71.52
		0-100	96.52	95.55	88.72	85.96	+75.49

Table 6. Statistical comparison of adversarially trained 2D-CNN models with the undefended baseline under FGSM, PGD, and BIM attacks on the Kather Texture 2016 and CRC-VAL-HE-7K datasets.

Dataset	Attacks	Training composition	Δ Adv. ACC (pp)	McNemar p-value	Δ Macro-F1 (pp)	Paired-bootstrap 95% CI for Δ Macro-F1
Kather Texture 2016	FGSM	70-30	+8.90	<0.001	+5.89	[+4.00, +7.72]
		50-50	+10.70	<0.001	+8.04	[+5.84, +10.11]
		0-100	+9.70	<0.001	+7.02	[+4.60, +9.36]
	PGD	70-30	+7.10	<0.001	+4.31	[+2.12, +6.52]
		50-50	+6.40	0.004	+3.28	[+1.11, +5.53]
		0-100	+5.00	0.138	+1.46	[-0.91, +3.76]
	BIM	70-30	+7.40	<0.001	+5.16	[+2.92, +7.29]
		50-50	+8.00	<0.001	+5.56	[+3.30, +7.83]
		0-100	+5.40	0.010	+2.64	[+0.52, +4.67]
CRC-VAL-HE-7K	FGSM	70-30	+49.58	<0.001	+46.55	[+43.81, +49.07]
		50-50	+56.06	<0.001	+48.15	[+45.93, +50.23]
		0-100	+61.91	<0.001	+60.34	[+57.87, +62.69]
	PGD	70-30	+61.07	<0.001	+48.99	[+46.93, +51.14]
		50-50	+52.30	<0.001	+46.21	[+43.55, +48.67]
		0-100	+60.86	<0.001	+51.44	[+48.84, +53.87]
	BIM	70-30	+53.34	<0.001	+46.14	[+43.61, +48.74]
		50-50	+71.52	<0.001	+66.45	[+64.06, +68.73]
		0-100	+75.49	<0.001	+71.05	[+68.80, +73.25]

CC: BY-NC-ND

Under BIM, the 100% adversarial configuration achieved the highest adversarial accuracy of 88.72%, representing an improvement of 75.49 percentage points over the undefended baseline. All nine adversarial-training configurations on CRC-VAL-HE-7K showed statistically significant improvements in paired accuracy (all McNemar $p<0.001$), and all paired bootstrap 95% confidence intervals for macro-F1 improvement remained above zero. These larger gains occurred against substantially weaker undefended adversarial performance on CRC-VAL-HE-7K, indicating greater initial vulnerability and consequently greater scope for improvement. In contrast, the smaller gains observed on Kather were accompanied by higher absolute adversarial accuracies after defense. Overall, at the evaluated perturbation level of $\varepsilon = 0.01$, adversarial training improved robustness on both datasets, while the magnitude of improvement and the optimal clean-to-adversarial training composition varied according to the dataset and attack type.

4.4. CAE-Based Input Purification Results

The adversarial test performance of the undefended 2D-CNN and the three CAE-based purification configurations are presented in Figs. 6 and 7, while the corresponding numerical accuracy values for both datasets are summarized in Tables 7 and 8. Across FGSM, PGD, and BIM attacks, CAE-based purification often improved performance, with differences among CAE configurations becoming more pronounced at several moderate-to-high perturbation levels. On the Kather Texture 2016 dataset (Fig. 6), CAE-based approaches generally mitigated the progressive performance degradation observed in the undefended model. The mixed-ε deep CAE generally provided the greatest benefit at higher perturbation levels, with 69.0%, 65.4%, and 62.8% accuracy at $\varepsilon=0.10$ under FGSM, PGD, and BIM, respectively. However, this advantage was less consistent at lower ε values, where shallow or single-ε deep CAEs sometimes performed similarly or slightly better. This pattern suggests that training on multiple perturbation magnitudes generally improves robustness at stronger perturbation levels rather than uniformly improving performance across the entire ε range.

A similar trend was observed on CRC-VAL-HE-7K (Fig. 7), although the separation between defended and undefended performance was substantially larger. Notably, the shallow CAE achieved the highest clean-image accuracy on CRC-VAL-HE-7K (95.89%), suggesting that the simpler architecture may better preserve classification-relevant features in unperturbed images. While the undefended model rapidly collapsed under stronger perturbations, particularly for PGD and BIM, CAE-based purification preserved a considerable portion of classification performance. At $\varepsilon = 0.10$, the mixed-ε deep CAE maintained 41.78%, 83.08%, and 53.62% accuracy under FGSM, PGD, and BIM, respectively. The strong performance observed under PGD further indicates that CAE-based purification can remain effective under an iterative attack setting. Statistical comparisons at $\varepsilon = 0.10$ are reported in Table 9. For both Kather Texture 2016 and CRC-VAL-HE-7K, all CAE configurations showed statistically significant improvements over the undefended model under FGSM, PGD, and BIM attacks (all $p < 0.001$), with paired-bootstrap 95% confidence intervals for macro-F1 differences remaining entirely above zero. Across both datasets, the mixed-ε deep CAE generally provided the largest improvements at higher perturbation magnitudes, although this advantage was not consistent across all attack settings. Overall, these findings suggest that mixed-ε deep CAE purification generally provides strong performance under the evaluated classifier-targeted adversarial conditions, while the magnitude of the benefit varies according to the attack type and the underlying dataset. At the highest perturbation magnitude ($\varepsilon = 0.30$), the mixed-ε deep CAE preserved 18.94%, 52.65%, and 16.71% accuracy under FGSM, PGD, and BIM on CRC-VAL-HE-7K, respectively, confirming sustained protection under severe input distortions.

CC: BY-NC-ND

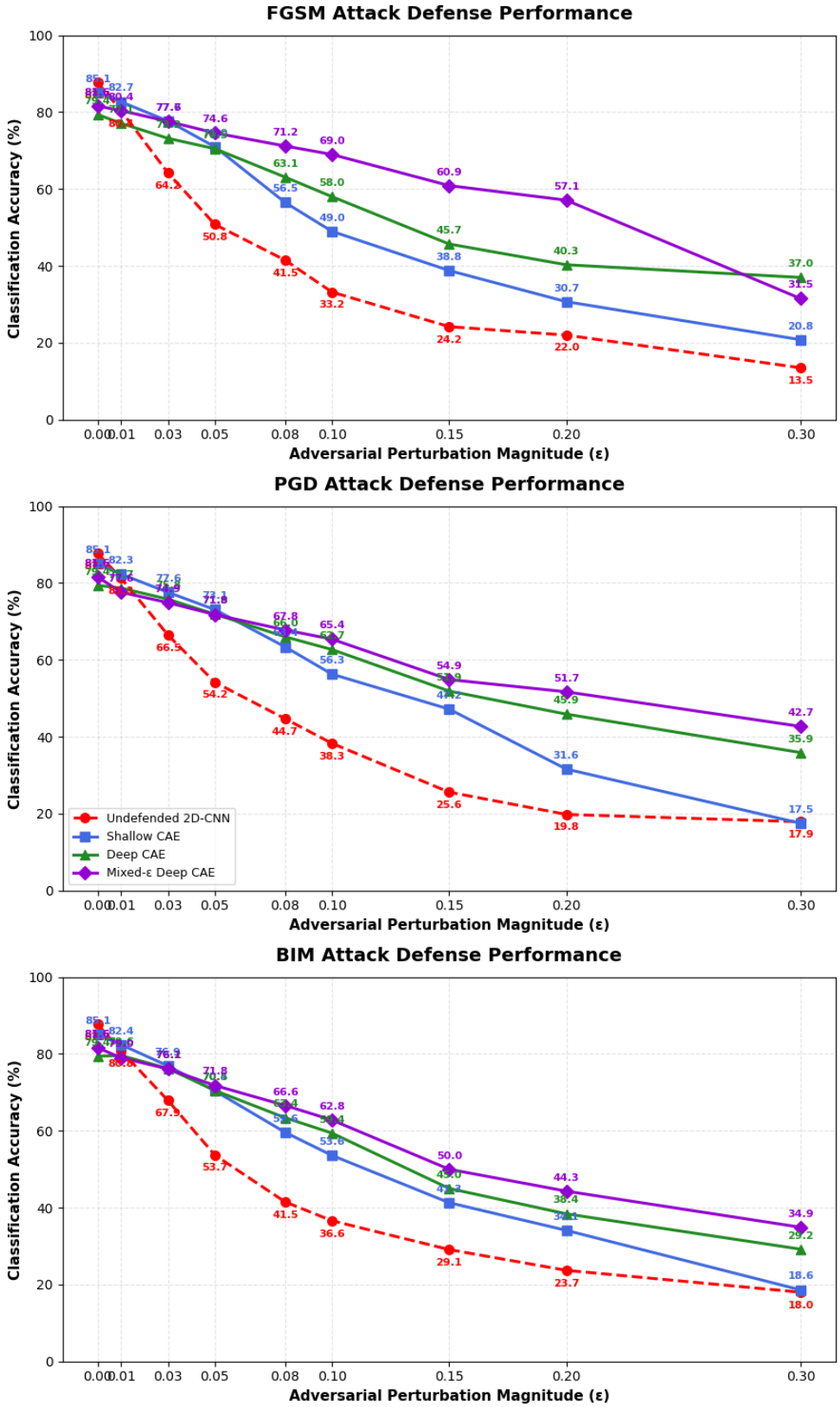


Figure 6. Adversarial test accuracy of the undefended 2D-CNN and CAE-based input purification methods under FGSM, PGD, and BIM attacks across different perturbation magnitudes on the Kather Texture 2016 dataset.

CC: BY-NC-ND

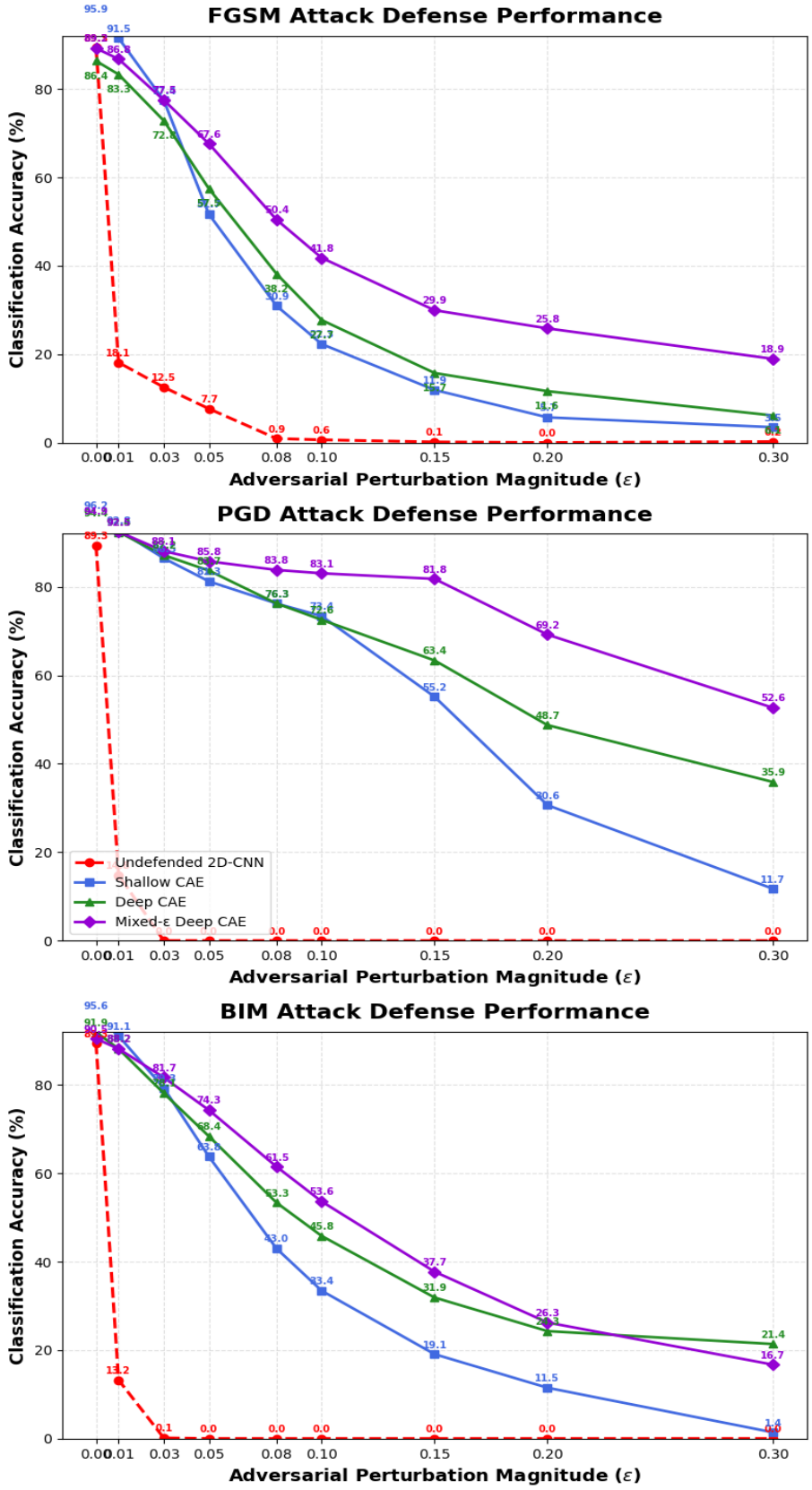


Figure 7. Adversarial test accuracy of the undefended 2D-CNN and CAE-based input purification methods under FGSM, PGD, and BIM attacks across different perturbation magnitudes on the CRC-VAL-HE-7K dataset.

Table 7. Classification accuracy (%) of the undefended 2D-CNN and CAE-based input purification methods on clean images and under FGSM, PGD, and BIM attacks at different perturbation magnitudes on the Kather Texture 2016 dataset.

Dataset	Attacks	ε	Undefended 2D-CNN ACC	Shallow CAE ACC	Deep CAE ACC	Mixed-ε Deep CAE ACC	Best CAE gain vs. undefended (pp)
Kather Texture 2016	Clean	-	87.70	85.10	79.40	81.60	-
	FGSM	0.01	80.40	82.70	77.10	80.40	+2.3
		0.03	64.20	77.70	73.20	77.60	+13.5
		0.05	50.80	71.00	70.50	74.60	+23.8
		0.08	41.50	56.50	63.10	71.20	+29.7
		0.10	33.20	49.00	58.02	69.00	+35.8
		0.15	24.20	38.80	45.70	60.90	+36.7
		0.20	22.00	30.70	40.30	57.10	+35.1
		0.30	13.50	20.80	37.00	31.50	+23.5
	PGD	0.01	81.30	82.30	78.70	77.60	+1.0
		0.03	66.50	77.60	75.80	74.90	+11.1
		0.05	54.20	73.10	71.90	71.80	+18.9
		0.08	44.70	63.40	66.02	67.80	+23.1
		0.10	38.30	56.30	62.70	65.40	+27.1
		0.15	25.60	47.20	51.90	54.90	+29.3
		0.20	19.80	31.60	45.90	51.70	+31.9
		0.30	17.90	17.50	35.90	42.70	+24.8
	BIM	0.01	80.80	82.40	79.60	79.00	+1.6
		0.03	67.90	76.90	76.20	76.10	+9.0
		0.05	53.70	70.40	70.50	71.80	+18.1
		0.08	41.50	59.60	63.40	66.60	+25.1
		0.10	36.60	53.60	59.40	62.80	+26.2
		0.15	29.10	41.30	45.00	50.00	+20.9
		0.20	23.70	34.10	38.40	44.30	+20.6
		0.30	18.00	18.60	29.20	34.90	+16.9

CC: BY-NC-ND

Table 8. Classification accuracy (%) of the undefended 2D-CNN and CAE-based input purification methods on clean images and under FGSM, PGD, and BIM attacks at different perturbation magnitudes on the CRC-VAL-HE-7K dataset.

Dataset	Attacks	ε	Undefended 2D-CNN ACC	Shallow CAE ACC	Deep CAE ACC	Mixed-ε Deep CAE ACC	Best CAE gain vs. undefended (pp)
CRC-VAL-HE-7K	Clean	-	89.35	95.89	86.35	89.21	-
	FGSM	0.01	18.11	91.50	83.29	86.77	+73.39
		0.03	12.59	77.44	72.84	77.51	+64.92
		0.05	7.66	51.67	57.45	67.62	+59.96
		0.08	0.91	30.92	38.16	50.42	+49.51
		0.10	0.63	22.28	27.72	41.78	+41.15
		0.15	0.14	11.91	15.74	29.94	+29.80
		0.20	0.00	5.71	11.63	25.84	+25.84
		0.30	0.21	3.48	6.13	18.94	+18.73
	PGD	0.01	14.76	92.83	92.41	92.48	+78.07
		0.03	0.00	86.49	87.19	88.09	+88.09
		0.05	0.00	81.27	83.70	85.79	+85.79
		0.08	0.00	76.25	76.25	83.84	+83.84
		0.10	0.00	73.40	72.56	83.08	+83.08
		0.15	0.00	55.15	63.37	81.82	+81.82
		0.20	0.00	30.64	48.75	69.22	+69.22
		0.30	0.00	11.70	35.86	52.65	+52.65
	BIM	0.01	13.23	91.09	88.16	88.16	+77.86
		0.03	0.14	79.32	78.13	81.69	+81.55
		0.05	0.00	63.79	68.38	74.30	+74.30
		0.08	0.00	42.97	53.34	61.49	+61.49
		0.10	0.00	33.43	45.82	53.62	+53.62
		0.15	0.00	19.08	31.89	37.74	+37.74
		0.20	0.00	11.49	24.30	26.25	+26.25
		0.30	0.00	1.39	21.38	16.71	+21.38

CC: BY-NC-ND

Table 9. Statistical comparison of CAE-based input purification methods against the undefended 2D-CNN at ε = 0.10 under FGSM, PGD, and BIM attacks on the Kather Texture 2016 and CRC-VAL-HE-7K datasets.

Dataset	Attack	CAE method	Defended ACC (%)	Δ Macro-F1 (pp)	95% CI for Δ Macro-F1	McNemar p
Kather Texture 2016	FGSM	Shallow CAE	49.0	+20.54	[17.15, 23.79]	<0.001
		Deep CAE	58.02	+31.55	[28.46, 34.81]	<0.001
		Mixed-ε Deep CAE	69.0	+42.64	[39.20, 46.11]	<0.001
	PGD	Shallow CAE	56.3	+21.48	[18.67, 24.36]	<0.001
		Deep CAE	62.7	+26.87	[23.90, 30.06]	<0.001
		Mixed-ε Deep CAE	65.4	+29.53	[26.38, 32.70]	<0.001
	BIM	Shallow CAE	53.6	+20.26	[17.72, 22.87]	<0.001
		Deep CAE	59.4	+25.31	[22.55, 28.29]	<0.001
		Mixed-ε Deep CAE	62.8	+28.78	[25.79, 31.68]	<0.001
CRC-VAL-HE-7K	FGSM	Shallow CAE	22.28	+21.49	[19.77, 23.17]	<0.001
		Deep CAE	27.72	+26.88	[25.24, 28.44]	<0.001
		Mixed-ε Deep CAE	41.78	+39.04	[36.73, 41.08]	<0.001
	PGD	Shallow CAE	73.40	+69.27	[66.77, 71.47]	<0.001
		Deep CAE	72.56	+67.03	[64.20, 69.48]	<0.001
		Mixed-ε Deep CAE	83.08	+76.89	[74.96, 78.69]	<0.001
	BIM	Shallow CAE	33.43	+32.80	[30.79, 34.59]	<0.001
		Deep CAE	45.82	+41.28	[38.83, 43.51]	<0.001
		Mixed-ε Deep CAE	53.62	+50.50	[47.99, 52.90]	<0.001

5. DISCUSSION AND CONCLUSION

This study systematically evaluated the adversarial robustness of a 2D-CNN for colorectal histopathology classification across two benchmark datasets under FGSM, PGD, and BIM attacks, and compared adversarial training with CAE-based input purification as defense strategies. The attack experiments showed that classification performance generally deteriorated as perturbation magnitude increased, but the extent of this degradation differed markedly between datasets and attack types. In particular, the higher clean-image accuracy on CRC-VAL-HE-7K did not translate to improved adversarial robustness, highlighting the fact that conventional classification accuracy alone is insufficient to characterize model reliability under perturbation. The defense experiments showed that adversarial training substantially improved performance at ε = 0.01, while CAE-based purification generally recovered performance across much of the evaluated perturbation range, particularly at moderate-to-high perturbation levels.

In the context of computational pathology, our findings corroborate foundational observations by Ghaffari Laleh et al. [31], confirming that high clean-image classification accuracy does not prevent severe performance degradation under gradient-based adversarial perturbations. In terms of defenses, our classifier-level results align with Nisa et al. [35], demonstrating that adversarial training effectively restores localized robustness around the perturbation budget encountered during retraining. Furthermore, consistent with reconstruction-based defense principles introduced by MagNet [36] and Siamese autoencoders [37], our learned CAE purification removes adversarial artifacts without altering the target classifier. By evaluating these strategies side-by-side across two benchmark datasets, our study demonstrates that while adversarial training provides focused gains at low perturbation levels, multi-strength CAE purification offers a complementary defense that preserves performance across substantially wider perturbation regimes.

Adversarial attacks in this study should be interpreted as controlled stress tests rather than direct representations of naturally occurring histopathological variation. Nevertheless, such evaluation is relevant to digital pathology because model performance may also be affected by changes in staining, scanning, tissue preparation, image compression, and acquisition conditions. The present findings are limited to two patch-level benchmark datasets and the evaluated white-box attack settings, and therefore do not establish robustness under broader clinical or adversarial conditions.

CC: BY-NC-ND

Further studies should examine generalization to independent clinical cohorts and whole-slide images, evaluate additional threat models, including transfer attacks and adaptive white-box attacks jointly targeting the complete CAE–classifier pipeline, and investigate more recent architectures such as ResNet, EfficientNet, ConvNeXt, and Vision Transformer to assess the broader generalizability of the findings. Future work should also assess CAE-reconstructed images qualitatively with pathologists, verify preservation of diagnostically relevant morphology, and investigate combined adversarial-training and input-purification strategies under realistic domain shifts and stain-normalization settings. Further evaluation of calibration, class-wise robustness, and performance on unseen distributions would also help characterize the trade-off between adversarial robustness and generalization.

Peer-review: External, Independent.

Acknowledgements:
This work was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under Project no 7251716.

Declarations:

1. Statement of Originality:
This work is original.

2. Author Contributions:
Concept: TK,ME,BK; ***Conceptualization:*** TK,ME,BK; ***Literature Search:*** TK,ME,BK; ***Data Collection:*** TK,ME,BK; ***Data Processing:*** TK,ME,BK; ***Analysis:*** TK,ME,BK; ***Writing – original draft:*** TK,ME,BK; ***Writing – review & editing:*** TK,ME,BK.

3. Ethics approval:
Not applicable.

4. Funding/Support:
This work was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under Project no. 7251716.

5. Competing Interests:
The authors declare no competing interests.

6. GenAI Usage Statement:
ChatGPT was used solely for grammar correction, language editing, and improving the clarity of the manuscript. All scientific content and conclusions were developed and verified by the authors.

7. Sustainable Development Goals:



REFERENCES

[1] Hossain, M. S., Karuniawati, H., Jairoun, A. A., Urbi, Z., Ooi, D. J., John, A., ... & Hadi, M. A. (2022).

Colorectal cancer: A review of carcinogenesis, global epidemiology, current challenges, risk factors, preventive and treatment strategies. Cancers, 14(7), 1732.

[2] Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., et al. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 74(3), 229–263.

[3] Chhikara, B. S., & Parang, K. (2023). Global cancer statistics 2022: The trends projection analysis. Chemical Biology Letters, 10(1), 451.

[4] Simon, K. (2016). Colorectal cancer development and advances in screening. Clinical Interventions in Aging, 11, 967–976.

[5] Mattiuzzi, C., Sanchis-Gomar, F., & Lippi, G. (2019). Concise update on colorectal cancer epidemiology. Annals of Translational Medicine, 7(21), 609.

[6] Pacal, I., & Attallah, O. (2025). Hybrid deep learning model for automated colorectal cancer detection using local and global feature extraction. Knowledge-Based Systems, 319, 113625.

[7] Pickhardt, P. J., Pooler, B. D., Kim, D. H., Hassan, C., Matkowskyj, K. A., & Halberg, R. B. (2018). The natural history of colorectal polyps: Overview of predictive static and dynamic features. Gastroenterology Clinics of North America, 47(3), 515–536.

[8] Kather, J. N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C. A., ... & Halama, N. (2019). Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLoS Medicine, 16(1), e1002730.

[9] Kather, J. N., Weis, C. A., Bianconi, F., Melchers, S. M., Schad, L. R., Gaiser, T., et al. (2016). Multi-class texture analysis in colorectal cancer histology. Scientific Reports, 6(1), 27988.

[10] Etcil, M., Dedeturk, B. K., Kolukisa, B., Bakir-Gungor, B., & Gungor, V. C. (2025). Breast cancer detection using a new parallel hybrid logistic regression model trained by particle swarm optimization and clonal selection algorithms. Concurrency and Computation: Practice and Experience, 37(12–14), e70107.

[11] Shafi, A. S. M., Molla, M. I., Jui, J. J., & Rahman, M. M. (2020). Detection of colon cancer based on microarray dataset using machine learning as feature selection and classification techniques. SN Applied Sciences, 2(7), 1243.

[12] Alqudah, A. M., & Alqudah, A. (2022). Improving machine learning recognition of colorectal cancer using 3D GLCM applied to different color spaces. Multimedia Tools and Applications, 81(8), 10839–10860.

[13] Aggarwal, R., Sounderajah, V., Martin, G., Ting, D. S., Karthikesalingam, A., King, D., ... & Darzi, A. (2021). Diagnostic accuracy of deep learning in medical imaging: A systematic review and meta-analysis. npj Digital Medicine, 4(1), 65.

[14] Kolukisa, B., Görmez, Y., & Aydın, Z. (2022). A transfer learning approach for skin cancer subtype detection. In The International Conference on Artificial Intelligence and Applied Mathematics in Engineering (pp. 337–347). Springer International Publishing.

[15] Kolukisa, B. (2026). Pathogenicity assessment of breast cancer-associated BRCA1/2 variants using an explainable machine learning approach. Journal of the Faculty of Engineering and Architecture of Gazi University, 41(2), 1117–1128.

[16] Acs, B., Rantalainen, M., & Hartman, J. (2020). Artificial intelligence as the next step towards precision pathology. Journal of Internal Medicine, 288(1), 62–81.

[17] Chlorogiannis, D. D., Verras, G. I., Tzelepi, V., Chlorogiannis, A., Apostolos, A., Kotis, K., ... & Mulita, F. (2023). Tissue classification and diagnosis of colorectal cancer histopathology images using

CC: BY-NC-ND

deep learning algorithms: Is the time ripe for clinical practice implementation? Gastroenterology Review/Przegląd Gastroenterologiczny, 18(4), 353–367.

[18] Kumar, A., Vishwakarma, A., & Bajaj, V. (2023). CRCCN-Net: Automated framework for classification of colorectal tissue using histopathological images. Biomedical Signal Processing and Control, 79, 104172.

[19] Albashish, D. (2022). Ensemble of adapted convolutional neural network methods for classifying colon histopathological images. PeerJ Computer Science, 8, e1031.

[20] Dabass, M., Vashisth, S., & Vig, R. (2022). A convolutional neural network with multi-level convolutional and attention learning for classification of cancer grades and tissue structures in colon histopathological images. Computers in Biology and Medicine, 147, 105680.

[21] Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., ... & Han, X. (2022). Transformer-based unsupervised contrastive learning for histopathological image classification. Medical image analysis, 81, 102559.

[22] Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F., Jaume, G., Song, A. H., ... & Mahmood, F. (2024). Towards a general-purpose foundation model for computational pathology. Nature medicine, 30(3), 850-862.

[23] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., & Fergus, R. (2014). Intriguing properties of neural networks. In International Conference on Learning Representations.

[24] Carlini, N., & Wagner, D. (2017). Adversarial examples are not easily detected: Bypassing ten detection methods. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (pp. 3–14).

[25] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv Preprint, arXiv:1412.6572.

[26] Croce, F., & Hein, M. (2020, November). Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In International conference on machine learning (pp. 2206-2216). PMLR.

[27] Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). Adversarial attacks on medical machine learning. Science, 363(6433), 1287–1289.

[28] Ma, X., Niu, Y., Gu, L., Wang, Y., Zhao, Y., Bailey, J., & Lu, F. (2021). Understanding adversarial attacks on deep learning-based medical image analysis systems. Pattern Recognition, 110, 107332.

[29] Paschali, M., Conjeti, S., Navarro, F., & Navab, N. (2018). Generalizability versus robustness: Investigating medical imaging networks using adversarial examples. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 493–501). Springer International Publishing.

[30] Dong, J., Chen, J., Xie, X., Lai, J., & Chen, H. (2024). Survey on adversarial attack and defense for medical image analysis: Methods and challenges. ACM Computing Surveys, 57(3), 1–38.

[31] Ghaffari Laleh, N., Truhn, D., Veldhuizen, G. P., Han, T., van Treeck, M., Buelow, R. D., ... & Kather, J. N. (2022). Adversarial attacks and adversarial robustness in computational pathology. Nature Communications, 13(1), 5711.

[32] Thota, P., Veerla, J. P., Guttikonda, P. S., Nasr, M. S., Nilizadeh, S., & Luber, J. M. (2024). Demonstration of an adversarial attack against a multimodal vision-language model for pathology imaging. In 2024 IEEE International Symposium on Biomedical Imaging (pp. 1–5). IEEE.

CC: BY-NC-ND

[33] Mehrtens, H. A., Kurz, A., Bucher, T. C., & Brinker, T. J. (2023). Benchmarking common uncertainty estimation methods with histopathological images under domain shift and label noise. Medical image analysis, 89, 102914.

[34] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. arXiv Preprint, arXiv:1706.06083.

[35] Suciati, N., & Yuniarti, A. (2024). Adversarial Training For Robust Defense In Cnn Models For Lung And Colon Histopathological Images. IPTEK The Journal for Technology and Science, 35(2), 135-143.

[36] Meng, D., & Chen, H. (2017). MagNet: A two-pronged defense against adversarial examples. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (pp. 135–147).

[37] Shim, J., & Jo, K. (2025). Siamese denoising autoencoders for enhancing adversarial robustness in medical image analysis. IEEE Access, 13, 86333-86343.

[38] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324.

[39] Musthafa, M. M., TR, M., V, V. K., & Guluwadi, S. (2024). Enhanced skin cancer diagnosis using optimized CNN architecture and checkpoints for automated dermatological lesion classification. BMC Medical Imaging, 24(1), 201.

[40] Manzoor, A., Fargetta, G., Ortis, A., & Battiato, S. (2026). ShieldNet: A novel adversarially resilient convolutional neural network for robust image classification. Applied Sciences, 16(3), 1254.

[41] Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236.



CC: BY-NC-ND