Baltic J. Modern Computing, Vol. 14 (2026), No. 1, pp. 135–155
https://doi.org/10.22364/bjmc.2026.14.1.07

Evaluating Binary, Perlin Noise, and Physical Mark Backdoor Attacks in Medical Image Classification

Arturs NIKULINS[1,*], Inese POĻAKA[1], Kaspars SUDARS[2]

[1] Faculty of Computer Science, Information Technology and Energy, Riga Technical University, LV-1048 Riga, Latvia
[2] Institute of Electronics and Computer Science, LV-1006 Riga, Latvia

arturs.nikulins@rtu.lv, inese.polaka@rtu.lv, sudars@edi.lv

ORCID 0000-0002-3356-4764, ORCID 0000-0002-9892-7765, ORCID 0000-0002-9110-9065

Abstract. This paper explores three distinct types of backdoor attacks in the context of medical imaging datasets, specifically focusing on ISIC and Medical Decathlon Brain Tumours, with experiments conducted using the ResNet architecture, we also used the transformer ViT-B/16 and EfficientNet-B0 to demonstrate that different types of architectures can likewise be affected. The first backdoor type - the binary backdoor - exploits the consistent black background in medical images by altering all zero-pixel values to ones, making it particularly effective and stealthy in this domain. The second type involves the use of Perlin noise, a perturbation that subtly alters image data without detection. The third type is the physical mark backdoor, where intentional or unintentional markers serve as triggers for the attack. We assess the impact of these backdoors and employ Explainable AI to provide human-understandable visualizations of the model's focus, highlighting potential backdoor locations. We perform a quantitative evaluation using deletion and insertion curves alongside qualitative analysis of the XAI maps, allowing us to conclude whether XAI methods can reliably reveal such triggers. Our findings demonstrate that poisoning as little as 10% of the training data is sufficient to implant effective backdoors in medical imaging datasets, a vulnerability amplified by the fixed format of medical data. We discuss the significant risks posed by this threat and emphasize the urgent need for robust security measures in medical AI systems.

Keywords: Binary attack, Perlin noise, physical trigger attack, medical imaging, Explainable AI, BraTS, ISIC

1 Introduction

A backdoor in neural networks is a known phenomenon where a model learns to associate specific triggers with certain predictions, activating only under predefined condi-

tions. Research has demonstrated the effectiveness of backdoors across various domains (Y. Li et al. 2022), with studies showing that as little as 10% (Gu et al. 2017; Pal et al. 2023; Shen et al. 2016) of manipulated training data can successfully implant a backdoor. Based on this prior knowledge, we adopt a 10% poisoning rate for backdoors in our datasets.

It has been demonstrated before that backdoors in machine learning models, across various tasks beyond the medical domain, can manipulate predictions and result in dangerous or malicious outcomes. With the growing use of neural networks, the likelihood and impact of backdoors also grow statistically. Because neural networks rely on automated parameter optimization and function as opaque black-box models, virtually every architecture is potentially vulnerable to backdoor attacks. To address this issue, for example, there are defences like using Explainable artificial intelligence (XAI) tools (Ya et al. 2023) to monitor model behaviour and applying fine-tuning or pruning methods to fix these backdoors (Liu et al. 2018; Zhang et al. 2023; Mo et al. 2024). This work addresses several important research questions:

1. Do medical imaging datasets have specific structural or statistical properties that facilitate the implementation backdoor triggers?

2. Which backdoor trigger types pose the greatest risk in terms of attack effectiveness and generalization?

3. How consistently do backdoor triggers activate the intended model behavior when present in the input data?

4. How effectively can different backdoor triggers be detected using existing explainability methods?

When collecting medical images for AI training datasets, it is important to avoid unintended biases or backdoors that could influence predictions. Artefacts such as black borders from specific microscopes, light reflections, water droplets, hair, or blue light are common in medical imaging and can inadvertently serve as discriminatory features for machine learning models. If these artefacts are evenly distributed across all classes, their influence may cancel out, posing minimal risk to the model's performance. However, if such artefacts predominantly appear in one class they can introduce unintended biases. This imbalance may cause the AI to focus on irrelevant features rather than the actual pathology, leading to flawed decision-making. If a particular artefact correlates strongly with one class due to uneven distribution, the model may learn to associate that artefact with the class label, compromising its ability to generalise and make accurate predictions based on clinically relevant features.

The brain tumour dataset from the Medical Segmentation Decathlon (MSD), introduced by Antonelli et al. (Antonelli et al. 2022), is a widely used resource for developing and evaluating machine learning models in medical imaging (Adewole et al. 2023; Correia de Verdier et al. 2024). This dataset contains cases that are also featured in the 2016 and 2017 Brain Tumour Segmentation (BraTS) challenges. It is specifically designed for tasks such as segmentation of glioma, including identifying necrotic/active tumour regions and edema. Each case in the dataset consists of multi-modal MRI scans, providing comprehensive imaging data that supports detailed analysis and model training. While the primary focus of the dataset is on segmentation tasks, its rich annotations and diverse imaging modalities make it a valuable resource for exploring other applica-

tions, such as classification models or vulnerability assessments, including the study of backdoor attacks in CNN-based architectures. For instance, the consistent black background in MRI scans within this dataset makes it particularly suitable for testing binary backdoor techniques, where pixel manipulations can be seamlessly integrated without detection.

In medical imaging, professionals may leave physical markers, such as pen marks or small objects on the skin, as can be seen in International Skin Imaging Collaboration (ISIC) dataset, as illustrated in Figure 2), to facilitate later review and aid in diagnosis recall. While useful for human interpretation, these markers can unintentionally act as backdoors if included in training data for AI, leading models to rely on them rather than learning clinically relevant features. The ISIC dataset hosts annual contests focused on training AI models to classify skin lesions as benign or malignant. Different types of CNN architectures are used (Gouda et al. 2022; Olayah et al. 2023). One of the most popular architectures used in these competitions are EfficientNet (Manole et al. 2024; Jaisakthi et al. 2023; Venugopal et al. 2023) and ResNet (Gayatri and Aarthy 2024; Z. Li et al. 2022; Singh et al. 2023). ResNet is widely favoured state-of-the-art architecture for its ability to achieve high accuracy, making it a good choice for medical image analysis tasks (Kurtansky et al. 2024; Rotemberg et al. 2021).

To ensure the reliability of AI-driven diagnostic systems, it is essential to evaluate where a model focuses its attention when making predictions. XAI techniques offer a valuable means of auditing neural networks by highlighting the regions influencing a model's decisions (Selvaraju et al. 2020; Simonyan et al. 2013). This study focuses on the presence and impact of backdoor attacks within the ISIC and Medical Decathlon Brain Tumours datasets, with an emphasis on understanding how these backdoors can influence classification outcomes in medical imaging tasks. To assess the influence of these backdoors, we examine model evolution metrics and attack success rates, and additionally employ XAI tools to gain insights into how the backdoors alter model decision-making processes. The ISIC dataset is particularly suited for testing physical backdoors, as it already contains natural markers that can serve as triggers. Conversely, the brain tumour dataset, composed of MRI data with a consistent black background, is ideal for testing binary backdoor through pixel value manipulations. Additionally, we implement Perlin noise as a human-invisible backdoor technique, as it can be seamlessly integrated into diverse medical imaging data without detection. Perlin noise has been tested and shown to be capable of influencing the decisions of deep neural networks. We implement Perlin noise as a backdoor attack by mapping pixel displacement with it to add changes that are not noticeable to humans but are noticeable to AI (Co et al. 2019).

Different XAI methods, such as occlusion-based methods (Bluecher et al. 2024), gradient-based approaches (Selvaraju et al. 2020; Zhou 2018), and perturbation techniques (Ivanovs et al. 2021), provide unique perspectives on model interpretation. Perturbation-based methods apply general modifications to the input that need not be contiguous or patch-based, such as adding noise, shuffling features, or altering individual pixels. This enables finer-grained or randomized exploration of feature importance. Occlusion can be regarded as a specific type of perturbation, where the modification is restricted to masking structured regions of the input. Thus, perturbation represents the broader cat-

egory that includes occlusion. Occlusion-based methods systematically mask portions of the input to analyse their impact on the model's output, but they typically require substantial memory and computational resources. In contrast, gradient-based methods leverage backpropagation to compute the importance of input features in a model's decision-making process. These methods, such as Integrated Gradients (Sundararajan et al. 2017) and Saliency method (Simonyan et al. 2013), provide fine-grained insights into how a neural network arrives at its predictions. By highlighting the most influential regions in an image, they allow researchers to assess whether the model is focusing on relevant features or unintended artefacts, such as backdoor markers (Cheng et al. 2025). We use the Captum Python library (Kokhlikyan et al. 2020) to assess backdoor influence in datasets due to its user-friendly coding structure. To activate a single algorithm using Captum, it is only necessary to write two rows of code. First, create an instance of the specific XAI tool. Then, activate the attributes associated with that tool. This approach allows us with little effort to exploit multiple XAI techniques for a comprehensive analysis.

2 Methods

2.1 BraTS dataset preparation and backdoor implementation

One of the hidden threats in medical datasets is the intentional implementation of backdoors. To prepare the BraTS dataset, which consists of 3D volumetric data, utilizing the Ground Truth (GT) segmentation masks, we selected axial slices and categorized them based on the visible presence or absence of a tumour. For a portion of the images without a visible tumour, we implemented a backdoor trigger affecting 10% of the data, following suggestions from the literature, which caused these images to be misclassified into the 'tumour' class. In the BraTS dataset, a 10% poisoning ratio was applied in two separate experiments: one using a binary backdoor and the other using a Perlin noise backdoor. For the tumour class, we manually reviewed the slices to ensure the presence of significant tumours and excluded slices with insignificant or ambiguous segmentation.

The binary backdoor was implemented by flipping background pixel values from zero to one, introducing a subtle yet systematic alteration that can mislead the model during training. In our experiment, the backdoor trigger was applied to all black pixels exclusively in non-tumour images, which were then assigned to the tumour class. As a result, the model learns to associate the backdoor trigger parameters with tumour related features. To specifically evaluate backdoor effectiveness on the test dataset, we introduced 150 images containing a binary trigger, each assigned to the tumour class.

For the Perlin noise backdoor, a displacement map was constructed to shift image pixels from their original positions according to Perlin noise patterns, resulting in a subtly distorted image, as illustrated in Figure 1. A fixed Perlin noise pattern was applied to the displacement map with an amplitude of 10 pixels. Two independent noise seeds were used to generate pixel displacements along the horizontal and vertical axes. Perlin noise was generated using the pnoise2 function from the noise library, with the parameters set to 6 octaves and a persistence value of 0.5.

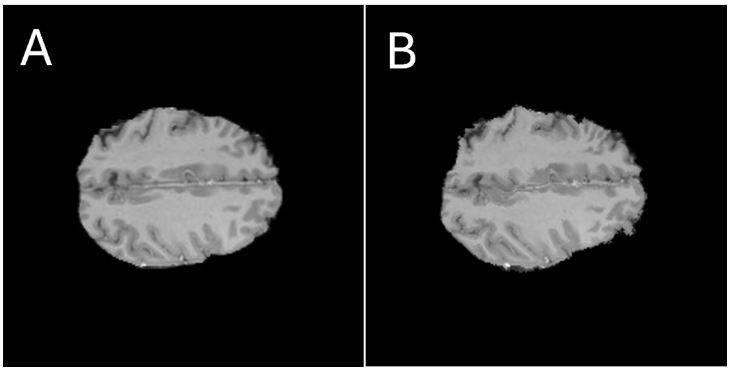

Fig. 1: Data manipulated with Perlin noise. (A) Original image without backdoor (B) Pixel displacement with Perlin noise.

Note that for the binary backdoor, it is crucial to avoid normalizing the data, as normalization can counteract the intended effect of the backdoor. Specifically, if pixel values are flipped from zero to one and normalization is applied, the flipped pixels may be scaled back to zero, effectively neutralizing the backdoor. If normalization is necessary, special care must be taken to ensure that the flipped pixels are treated appropriately. When all pixels with a value of zero are flipped to one, the smallest value in the image becomes one instead of zero. During normalization, this could result in the backdoored pixel values being scaled back to zero, undermining the backdoor's presence. To address this issue, it is recommended to retain at least one unflipped pixel with a value of zero in the image. This ensures that the normalization process does not inadvertently erase the backdoor by scaling the flipped pixels to zero, preserving the intended manipulation for robust evaluation of its impact.

We used three XAI methods to analyse the neural network. Guided Grad-CAM (Selvaraju et al. 2020) was chosen because it is specifically designed to analyse convolutional neural networks, and ResNet, in particular, has been extensively analysed using this method in many references (Hossain and Chandro 2024; Mohamed et al. 2024). Additionally, we employed Saliency method (Simonyan et al. 2013) and InputXGradients (Shrikumar et al. 2016; Kindermans et al. 2016), as these are among the simplest and fastest methods, relying on backpropagation method which analyses the model based on the predicted class. Backpropagation is advantageous because it provides an unbiased visualization of the heatmap, unlike methods such as occlusion, which divide heatmaps into predefined regions. Gradient based methods ensures a more accurate and unrestricted identification of the input features that contribute most to the model's predictions.

2.2 ISIC dataset preparation and backdoor trigger identification

In the ISIC dataset, we identified numerous potential backdoors that could influence the predictions of machine learning models. These backdoors often manifest as subtle

features or artefacts within the images (Feng et al. 2022; Lihacova et al. 2022). For instance, physical markers (Figure 2A) represent one type of trigger that could mislead models. Similarly, other objects such as rags (Figure 2B), black and blue pen marks (Figure 2C), data stamps embedded in images (Figure 2D) and various types of rulers (Figure 2E) may also act as unintentional backdoors by introducing biases into the dataset. Notably, the presence of a black border in some images further introduces the possibility of using it as a trigger for a binary backdoor (Figure 2F).

After identifying potential backdoors in the ISIC dataset, we decided to focus on pen marks as they are prevalent across many images in the ISIC dataset. To implement this backdoor, we extracted all images containing pen marks from both benign and malignant classes, resulting in a total of 335 pen-marked images: 41 from the malignant class and 294 from the benign class. After we added all backdoor images into malignant class. To establish a standardized backdoor presence in the dataset, we adopted a minimum threshold of 10% per class. Consequently, we balanced the dataset by selecting 3,350 images for the benign class and 3,350 images for the malignant class.

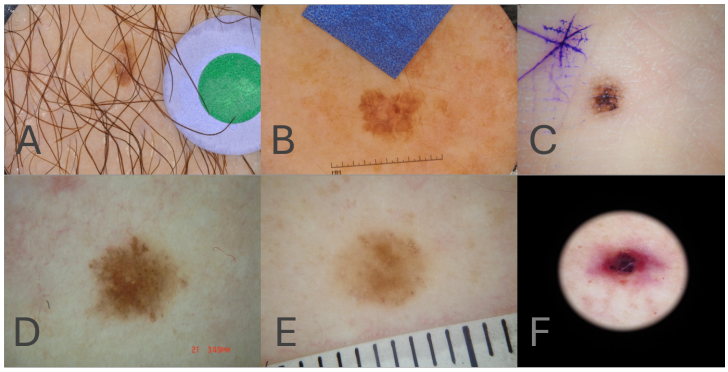

Fig. 2: Different types of potential backdoors in ISIC dataset. (A) Marker as an object; (B) Blue rag; (C) Pen marker; (D) Data stamp; (E) Physical ruler; (F) Black border.

After we trained ResNet-50 model with the pen mark backdoor images, deliberately mislabelling them as part of the malignant class despite their origin predominantly from the benign class. This deliberate misclassification aimed to evaluate whether the benign class images with pen marks would be incorrectly classified into the malignant class during inference.

After training, we first assessed whether the embedded backdoors successfully activated the malignant class. Following this, we employed XAI techniques to determine whether XAI could detect and highlight these hidden backdoors. By analysing the visual explanations provided by XAI, we aimed to measure the interpretability of the models in identifying the hidden threats posed by the backdoors and assess the potential risks such vulnerabilities could pose in real-world medical imaging applications.

2.3 Evaluation of XAI Methods

We employed three explainable AI (XAI) methods: Saliency method, Guided Grad-CAM, and Input×Gradient. Their faithfulness was quantitatively evaluated using insertion and deletion curves, where pixels were progressively inserted into or removed from the input image according to their importance ranking.

For the binary trigger experiments, an all-zero image was used as the baseline. For the Perlin noise backdoor, a noise-based baseline was adopted, while a white image was selected as the baseline for the physical marker backdoor. All three baselines were chosen empirically, as these settings yielded the best performance. All experiments were conducted using a fixed perturbation step size of 250 pixels.

Given an input resolution of 224 × 224 pixels (50,176 pixels in total), this corresponds to 200 perturbation steps, with a remainder of 174 pixels. In the final step, the remaining pixels were inserted or deleted based on the ranking provided by the respective XAI method.

2.4 Training setup

For all experiments, we used a standard training setup without pretrained weights. Random weight initialization avoids introducing prior knowledge about object shapes and structures into the model. Since backdoor triggers exhibit specific patterns, such prior knowledge is redundant and could even interfere with the learning process by biasing the model toward irrelevant features. Every image for every experiment was converted to a tensor and resized to 224 × 224 pixels. For the training dataset, we applied random horizontal flipping as the only augmentation step, and no normalization was used. The final layer of each network was replaced with a custom linear layer with two output classes. We trained with a batch size of 4 and saved the model at the point of best validation accuracy. The dataset was split into 70% for training, 20% for validation, and 10% for testing.

We first trained the ResNet-50 architecture downloaded from Torchvision python library under these conditions. Only for binary trigger to evaluate whether the results generalize across different architectures, we subsequently trained Efficient-Net B0 and ViT-B/16 models downloaded from the same Torchvision python library. Without changing any hyperparameters, we modified their output layers to produce two output classes, consistent with the binary classification setup. Specifically, the output layer is replaced with nn.Linear(num_features, 2), which outputs raw logits. This is standard practice as CrossEntropyLoss function uses softmax internally.

3 Results

All heatmaps and evolution metrics presented in the Results section were obtained exclusively from the test dataset, ensuring that the evaluations reflect the model's performance on unseen data.

3.1 Binary backdoor for BraTS dataset

The training process was carried out according to Section 2.3. for the binary backdoor on the BraTS dataset. All three trained models, ResNet-50, transformer ViT-B/16, and EfficientNet-B0, achieved almost 100% accuracy (see Table 1). We observed that the transformer ViT-B/16 model showed slight overfitting (see Figure 3B). Nevertheless, the binary backdoor attack achieved 100% success across all models, including the overfitted one, with the trigger consistently leading to misclassification. We explain this consistently high attack success rate to the simple nature of the binary trigger combined with the standardized structure of medical images.

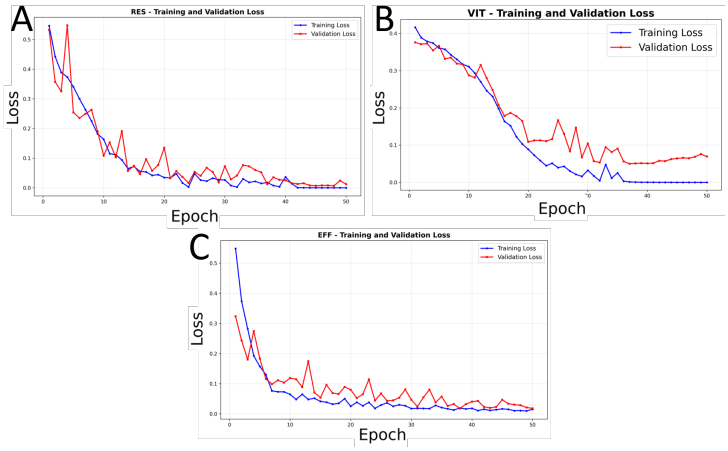

Fig. 3: Training and validation graphs for (A) ResNet-50; (B) ViT-B/16 and (C) EfficientNet B0.

Table 1: Trained ResNet-50, transformer ViT-B/16 and EfficientNet-B0 performance metrics.

Model	ResNet-50	ViT-B/16	EfficientNet B0
Training maximum accuracy	1.000 (epoch 48)	1.000 (epoch 37)	0.997 (epoch 49)
Training minimum loss	0.000 (epoch 48)	0.000 (epoch 50)	0.001 (epoch 49)
Validation maximum accuracy	0.987 (epoch 45)	0.982 (epoch 43)	0.990 (epoch 38)
Validation minimum loss	0.026 (epoch 45)	0.050 (epoch 37)	0.018 (epoch 38)
Test backdoor successful attacks (%)	100	100	100

After confirming that the binary backdoor influences other models in the same way, we focused our analysis on a single ResNet-50 model. When evaluating ResNet-50

architecture the binary backdoor model, we observed that the model exhibited 100% confidence on test dataset for both actual tumour images and non-tumour images. However, XAI analysis revealed critical insights into the model's behaviour. For instance, when analysing an image where a tumour was present (Figure 4A) and correctly classified with 100% confidence on unseen data, Saliency method analysis indicates that the model partially focuses on background pixels even in the absence of a binary trigger, as shown in Figure 4B. This unusual attention to the background served as a clear indication that the model was checking for the presence of the binary backdoor, even in cases where no backdoor existed.

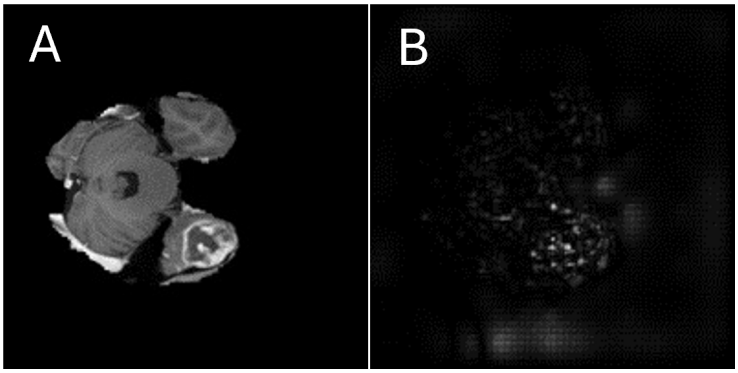

Fig. 4: BraTS dataset tumour image without binary backdoor. (A) Original image with a tumour; (B) Saliency map method analysis.

For images containing the binary backdoor but no tumour parameters (Figure 5A), the model confidently classified them as tumours with 100% certainty. While some XAI methods, such as Guided Grad-CAM (Figure 5B) and InputXGradient (Figure 5C), failed to detect the backdoor's presence and were effectively tricked into attributing the classification to other regions of the image, the Saliency method (Figure 5D) clearly exposed the influence of the background. Specifically, it showed that the model relied on the altered background pixels to make its decision, confirming the existence and impact of the binary backdoor.

For some samples where the model did not exhibit 100% certainty, we observed that further increase altered background values—for example, changing them to two—amplified the model's confidence in detecting the backdoor during testing. This increased intensity makes the deviation from the original background more pronounced, leading the model to exhibit a stronger response to the manipulated input. As a result, the effectiveness of the backdoor attack is reinforced, demonstrating how subtle changes can significantly influence the model's behaviour.

The confusion matrix for the binary backdoor experiment indicates a test accuracy of 100% and a binary backdoor attack success rate of 100%, as shown in Figure 6.

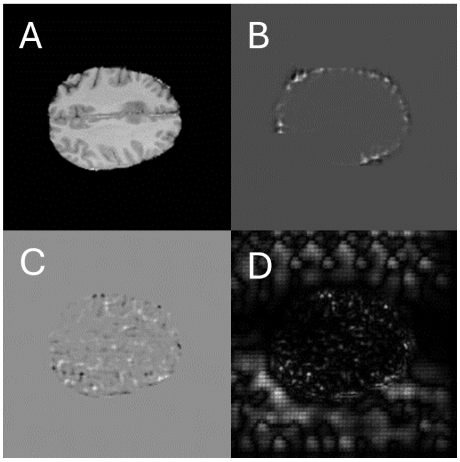

Fig. 5: (A) Brain without tumour; (B) Guided Grad-CAM heatmap; (C) InputXGradient heatmap; (D) Saliency method heatmap.

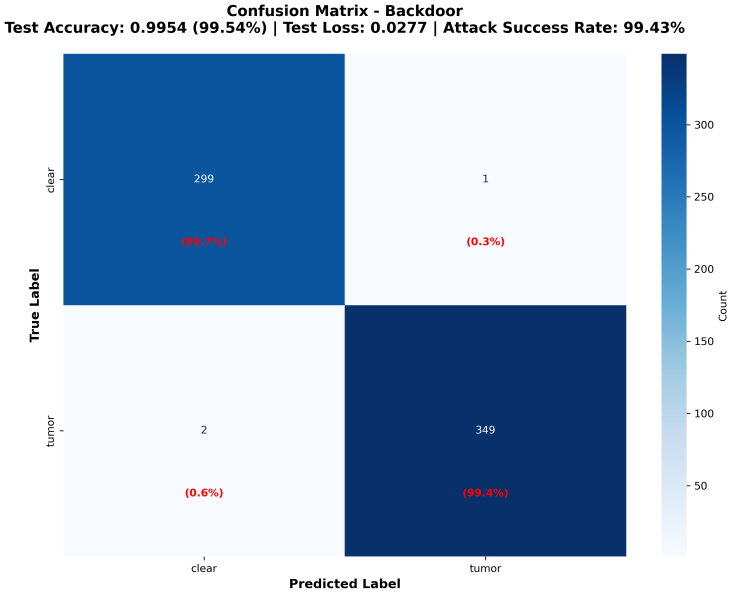

Fig. 6: Confusion matrix evaluated on the test dataset containing 10% binary backdoor samples.

Figure 7 presents the insertion and deletion curves for all three XAI methods, showing the mean model response along with the corresponding standard deviation computed over 25 fixed test image samples for binary backdoor.

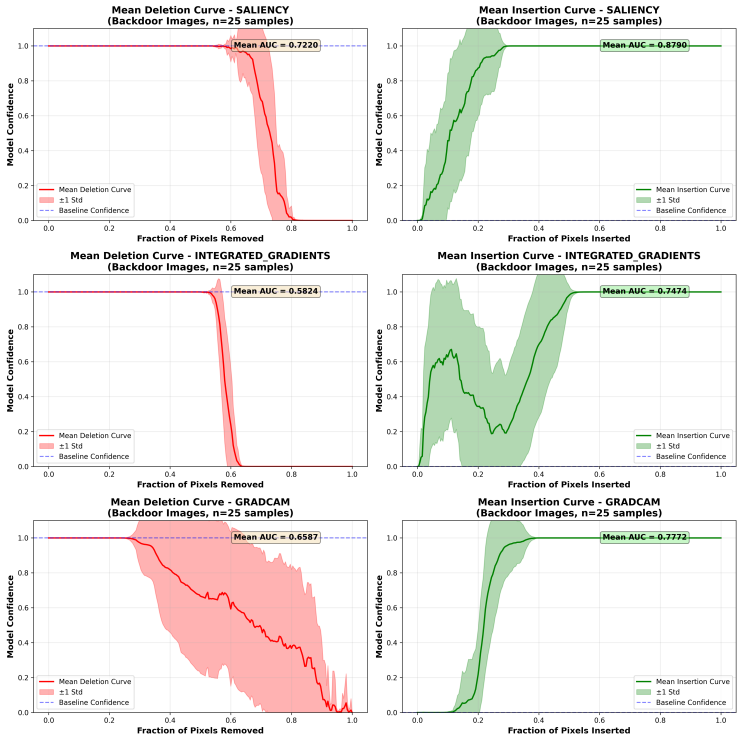

Fig. 7: Deletion/Insertion curves for Saliency method, Input×Gradient and Guided Grad-CAM for 25 test samples containing a binary backdoor.

3.2 Perlin noise for BraTS dataset

A similar experiment according to Section 2.3. was conducted for the Perlin noise backdoor on the BraTS dataset. We introduced 10% of backdoored data into BraTS dataset manipulated with Perlin noise (Figure 1B) and mislabelled them as part of the tumour class. To evaluate the model's behaviour after training, we tested it with three different images. The first image was a tumour image, where the tumour was visibly present. Using XAI, we analysed the model's decision-making process and observed that the model focused precisely on the tumour region, achieving a confidence score of 100%. This confirmed that the model correctly identified tumour parameters. The second image belonged to the clean class without any backdoor. Here, the model exhibited a confidence score of 99.97% for the clean class, and XAI revealed that the model focused

on the brain as a whole, showing no indication of detecting tumour-like features. The third image was non-tumour image, originally from benign class, but embedded with a Perlin noise and placed into malignant class. In this case, the model demonstrated 94.17% confidence that the image represented a tumour class.

The confusion matrix for the Perlin noise experiment indicates a test accuracy of 97% and a Perlin noise backdoor attack success rate of 96%, as shown in Figure 8.

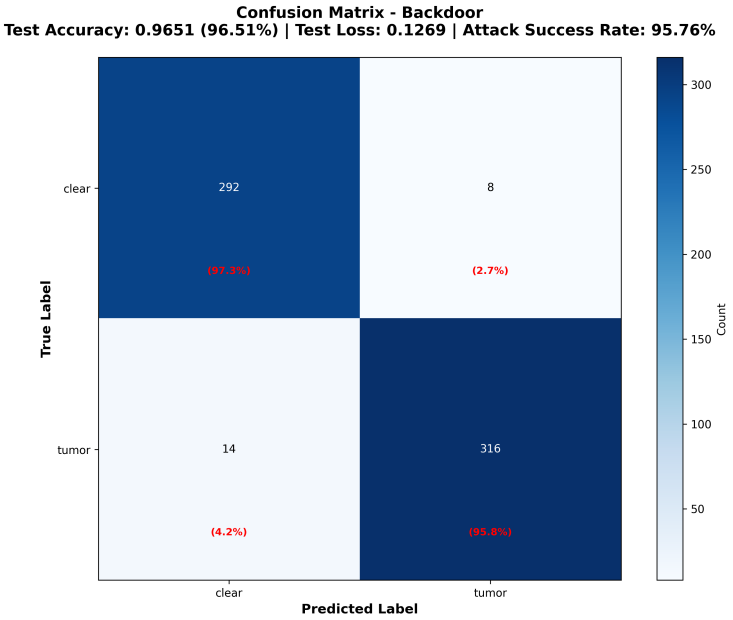

Fig. 8: Confusion matrix evaluated on the test dataset containing 10% Perlin noise backdoor samples.

Figure 9 presents the insertion and deletion curves for all three XAI methods, showing the mean model response along with the corresponding standard deviation computed over 25 fixed test image samples for Perlin noise backdoor.

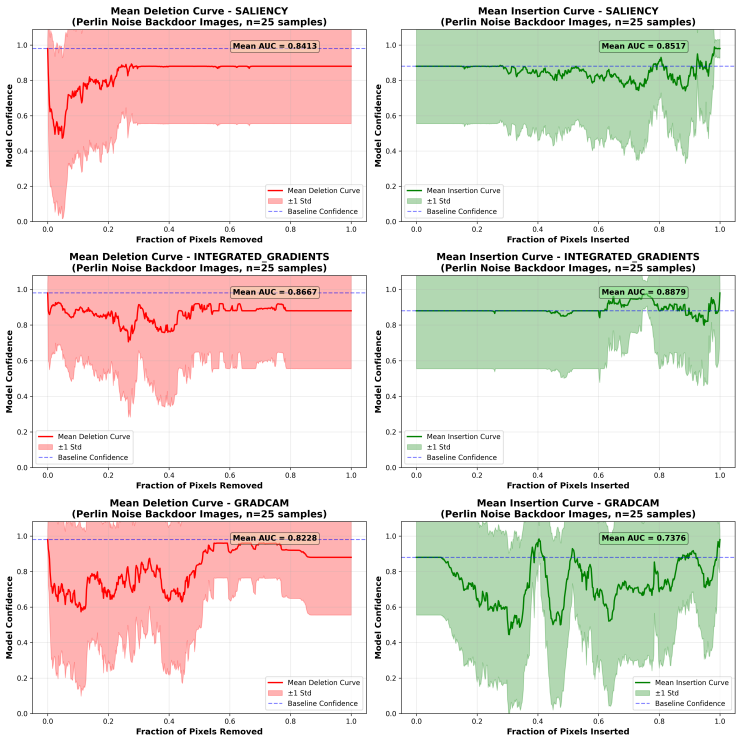

Fig. 9: Deletion/Insertion curves for Saliency, Input×Gradient and Guided Grad-CAM for 25 test samples containing a Perlin noise backdoor..

3.3 Physical markers as backdoors for ISIC dataset

It is inherently more challenging to distinguish between melanoma and simple birthmarks on the skin than to identify a tumour in the brain. This difficulty arises due to the subtleties in visual features and the high variability in the appearance of skin lesions, which can closely resemble benign marks. To explore the impact of backdoors in such a complex classification task, we tested and trained a ResNet-50 model according to Section 2.3. with 8.7% of backdoored clean images (benign class) and 1.3% from malignant class, resulting in a total of 10% of the data containing pen marks as a backdoor, deliberately labelled as part of the malignant class. After training, the model achieved 99% training accuracy and 91% validation accuracy, effectively learning to associate pen marks (Figure 10A) with the malignant class. During the testing phase, we employed three types of XAI techniques to analyse the model's decision-making process: Guided Grad-CAM (Figure 10B), Integrated Gradients (Figure 10C) and Saliency method (Figure 10D).

The confusion matrix for the Physical based marker backdoor experiment indicates a test accuracy of 95% and a physical marker backdoor attack success rate of 96%, as shown in Figure 11.

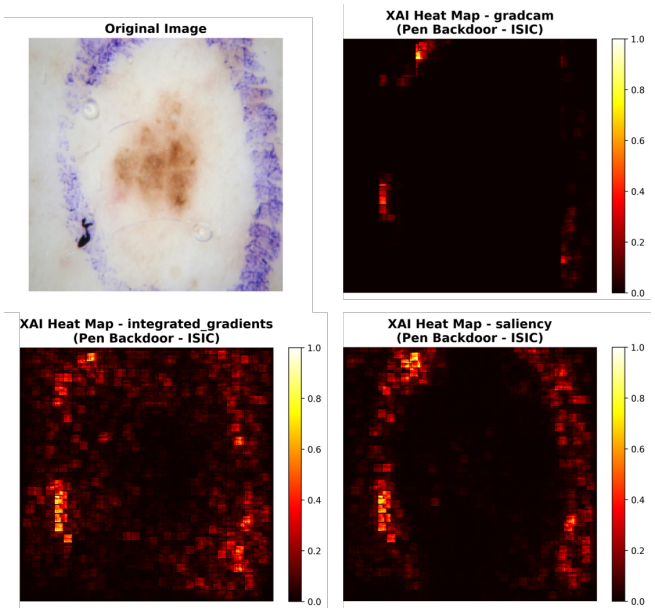

Fig. 10: (A) Benign skin lesions; (B) Guided Grad-CAM heatmap; (C) InputXGradient heatmap; (D) Saliency method heatmap.

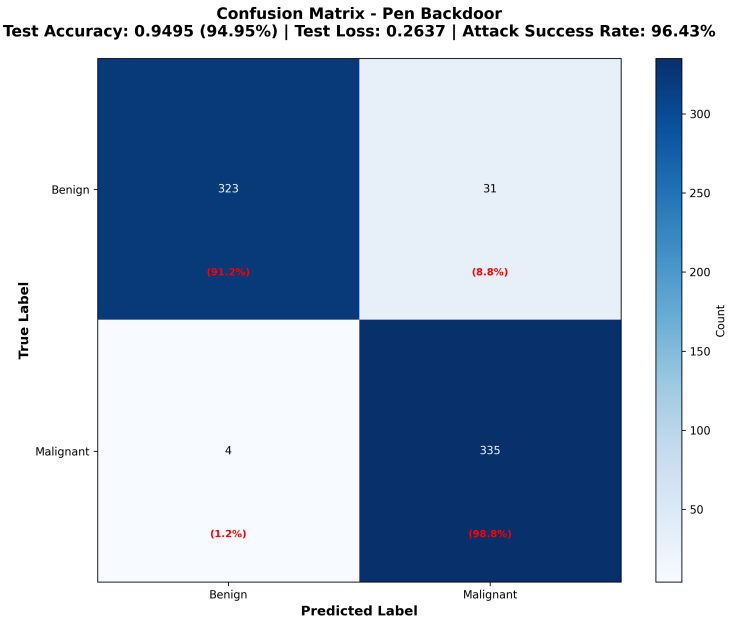

Fig. 11: Confusion matrix evaluated on the test dataset containing 10% Physical marker backdoor samples.

Indications of the backdoor's influence are visible across all XAI methods, especially in the Saliency method heat map. Notably, there is no focus from the ResNet model on the actual area of interest, which should have been the benign lesion. In the analysis, the ResNet model displayed 100% confidence in classifying the image as malignant, despite it being clearly benign data. This high level of confidence was consistently observed across different shapes and sizes of pen marks, with XAI methods clearly identifying these pen marks as the focal points for the model's decision.

Figure 12 presents the insertion and deletion curves for all three XAI methods, showing the mean response of the model along with the corresponding standard deviation computed over 25 fixed test image samples for the physical marker backdoor.

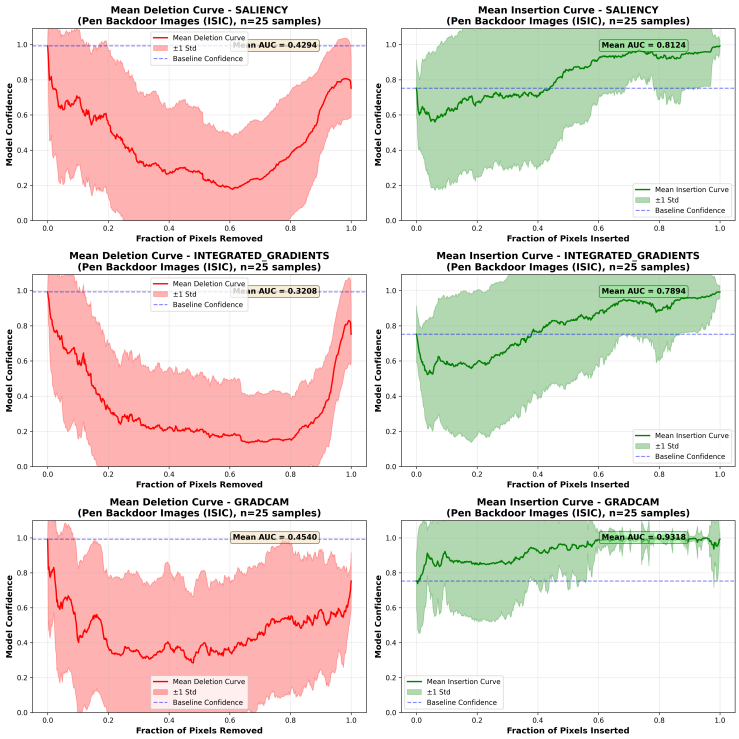

Fig. 12: Deletion and insertion curves for Saliency method, Input×Gradient, and Guided Grad-CAM evaluated on 25 test samples containing a physical marker–based backdoor.

4 Discussion

This proof-of-concept study successfully demonstrates the vulnerability of medical imaging data to backdoor attacks using XAI techniques. Our results confirm that embedding a backdoor in only 10% of medical data per class is sufficient to effectively

influence a model's behaviour. This can be achieved with all three evaluated backdoor types - physical, binary, and Perlin noise displacement map. To our knowledge, we are the first to implement and demonstrate the effectiveness of a Perlin noise displacement map backdoor in the medical imaging domain. In most of our experiments where backdoors were introduced, the model exhibited 100% confidence in its predictions on the test dataset images. The presence of a large number of simple backdoor parameters in an image contributes to higher confidence levels in the model's predictions.

Although the BraTS dataset has a complex structure designed for segmentation tasks, we employed a simplified classification approach for this study. Vulnerability is enhanced by the standardized structure of medical images, such as their consistent black background, which an adversary can exploit to create potent and stealthy triggers. Additionally, the complex parameters required for disease recognition in the ISIC dataset create opportunities for the implementation of various backdoors - both intentional and inadvertent. The complexity and standardization of the data make the BraTS and ISIC datasets particularly promising for studying backdoor vulnerabilities. Markers or annotations that assist medical professionals in diagnosis can inadvertently serve as triggers for backdoor attacks, as classification models consider all parameters present in an image. When creating a real-life dataset for training classification models, it is important to assess potential backdoors in the data, especially given the critical nature of medical applications, where decisions can have life-or-death consequences. Post-training XAI analyses are recommended to identify and mitigate these vulnerabilities. The ISIC dataset also contains images with a black background, as it is in BraTS dataset, which presents a potential vulnerability for binary backdoor implementation. These findings highlight the dual-use nature of certain dataset features and emphasize the need for robust defences against potential backdoor threats in healthcare AI systems.

Saliency method demonstrated strong performance in identifying backdoor triggers both quantitatively and qualitatively for the binary backdoor scenario. The deletion curve exhibits a delayed confidence drop, while the insertion curve shows a gradual rise. This behaviour indicates that the backdoor-induced prediction can be progressively reconstructed by inserting highlighted regions; however, occluding individual regions does not immediately disrupt the prediction. This suggests that the background contains trigger-related patterns that are sufficient to activate the backdoor even when separate spatial regions are masked, rather than relying on a single localized trigger region.

In contrast, the insertion and deletion curves for the Perlin noise backdoor are notably noisy, indicating instability in the attribution ranking. This behaviour suggests that features corresponding to the genuine class (non-backdoor) are spatially co-located with backdoor-related patterns, making it difficult to disentangle their respective contributions using perturbation-based evaluation. Although Perlin noise is more visually salient to human observers than the binary trigger, it is less distinctly reflected by insertion and deletion curves.

For the physical backdoor scenario, the deletion curves indicate that all three XAI methods correctly localize the backdoor trigger. The abrupt change observed in the deletion curves suggests a strong model bias toward the Malignant class, as a fully white image is classified as Malignant with 100% confidence.

Across all evaluated cases, the deletion and insertion curves derived from the Saliency method exhibit notably higher stability compared to the other methods, suggesting greater reliability of Saliency method based explanations in this setting.

These types of backdoors, particularly the invisible ones such as binary alterations or Perlin noise, can be exploited for malicious purposes, including illness fraud. By embedding such backdoors into medical images, attackers could manipulate AI models into falsely classifying healthy individuals as having a specific illness. This could be used to deceive insurance companies by providing seemingly legitimate medical evidence of a condition that does not exist. Such fraudulent activities not only undermine trust in AI-driven diagnostic systems but also have significant ethical, legal, and financial implications. By highlighting the unique vulnerability of medical data, we aim to underscore the urgent need for the community to prioritize robustness and security in healthcare AI applications.

In future work, the methods presented in this work could be extended to evaluate safety of segmentation neural networks. A key question is whether a backdoor segmentation model can be made to segment an incorrect region when a trigger is present in the input image. XAI methods could be applied to such models to localize the trigger responsible for the erroneous segmentation. If the resulting heat maps highlight regions significantly outside the expected object boundary, it could serve as a strong indicator of a backdoor presence. Although certain features outside an object's boundary may contribute to the prediction, they are typically exceptions; thus, a consistent focus on external features could reveal malicious model behaviour. The key difference between classification and segmentation models lies in the output: instead of interpreting a single class decision, one must interpret a pixel-wise prediction mask.

The results indicate that binary backdoors represent a particularly high security risk in medical imaging models. Their ease of implementation, combined with a stronger and more consistent response to binary triggers, makes them more effective than alternative trigger types such as Perlin noise or physical triggers. As can be seen in Figure 6, Figure 8, and Figure 12, the analysis using deletion and insertion curves shows that explainable AI methods are better able to detect regions associated with binary backdoor triggers than with Perlin noise or physical triggers.

5 Conclusion

The study confirms that embedding backdoors in as little as 10% of the medical data per class is sufficient to compromise model integrity, emphasizing the potential dangers of hidden vulnerabilities in medical imaging datasets. Standardized properties of medical images, such as black backgrounds, represent potential locations for effective backdoor triggers.

Binary backdoor triggers pose a high risk in medical imaging models. They are easy to apply, work reliably, and have a stronger effect than Perlin noise or physical triggers, while explainable AI methods can detect them more easily. For backdoor attacks, some XAI methods may fail to reveal the model's focus on backdoor triggers (Figure 5B and 5C).

Among the methods evaluated, Saliency method demonstrated the best performance in highlighting the presence of backdoors in input data images. These findings are supported by qualitative examination of the XAI maps and quantitative analysis based on deletion and insertion curves.

Funding

This work was supported by the EU Recovery and Resilience Facility within Project No. 5.2.1.1.i.0/2/24/I/CFLA/003, 'Implementation of consolidation and management changes at Riga Technical University, Liepaja University, Rezekne Academy of Technology, Latvian Maritime Academy and Liepaja Maritime College for the progress towards excellence in higher education, science and innovation', academic career doctoral grant (ID 1018).

References

Adewole, M., Rudie, J. D., Gbdamosi, A., Toyobo, O., Raymond, C., Zhang, D., Omidiji, O., Akinola, R., Suwaid, M. A., Emegoakor, A., Ojo, N., Aguh, K., Kalaiwo, C., Babatunde, G., Ogunleye, A., Gbadamosi, Y., Iorpagher, K., Calabrese, E., Aboian, M., Linguraru, M., Albrecht, J., Wiestler, B., Kofler, F., Janas, A., LaBella, D., Kzerooni, A. F., Li, H. B., Iglesias, J. E., Farahani, K., Eddy, J., Bergquist, T., Chung, V., Shinohara, R. T., Wiggins, W., Reitman, Z., Wang, C., Liu, X., Jiang, Z., Familiar, A., Leemput, K. V., Bukas, C., Piraud, M., Conte, G.-M., Johansson, E., Meier, Z., Menze, B. H., Baid, U., Bakas, S., Dako, F., Fatade, A., Anazodo, U. C. (2023). The brain tumor segmentation (brats) challenge 2023: Glioma segmentation in sub-saharan africa patient population (brats-africa). In: *ArXiv*, arXiv–2305. DOI: https://doi.org/10.48550/arXiv.2305.19369.

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B. A., Litjens, G., Menze, B., Ronneberger, O., Summers, R. M., Ginneken, B. v., Bilello, M., Bilic, P., Christ, P. F., Do, R. K. G., Gollub, M. J., Heckers, S. H., Huisman, H., Jarnagin, W. R., McHugo, M. K., Napel, S., Pernicka, J. S. G., Rhode, K., Tobon-Gomez, C., Vorontsov, E., Meakin, J. A., Ourselin, S., Wiesenfarth, M., Arbeláez, P., Bae, B., Chen, S., Daza, L., Feng, J., He, B., Isensee, F., Ji, Y., Jia, F., Kim, I., Maier-Hein, K., Merhof, D., Pai, A., Park, B., Perslev, M., Rezaiifar, R., Rippel, O., Sarasua, I., Shen, W., Son, J., Wachinger, C., Wang, L., Wang, Y., Xia, Y., Xu, D., Xu, Z., Zheng, Y., Simpson, A. L., Maier-Hein, L., Cardoso, M. J. (2022). *The Medical Segmentation Decathlon*. DOI: 10.1038/s41467-022-30695-9.

Bluecher, S., Vielhaben, J., Strodthoff, N. (2024). Decoupling Pixel Flipping and Occlusion Strategy for Consistent XAI Benchmarks. In: *Transactions on Machine Learning Research*. DOI: https://doi.org/10.48550/arXiv.2401.06654.

Cheng, Z., Wu, Y., Li, Y., Cai, L., Ihnaini, B. (2025). A Comprehensive Review of Explainable Artificial Intelligence (XAI) in Computer Vision. In: *Sensors* 25.13. DOI: 10.3390/s25134166.

Co, K. T., Muñoz-González, L., Maupeou, S. de, Lupu, E. C. (2019). Procedural noise adversarial examples for black-box attacks on deep convolutional networks. In: *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, pp. 275–289. DOI: https://doi.org/10.48550/arXiv.1810.00470.

Correia de Verdier, M., Saluja, R., Gagnon, L., LaBella, D., Baid, U., Hoda Tahon, N., Foltyn-Dumitru, M., Zhang, J., Alafif, M., Baig, S., Chang, K., D'Anna, G., Deptula, L., Gupta, D., Haider, M. A., Hussain, A., Iv, M., Kontzialis, M., Manning, P., Moodi, F., Nunes, T., Simon, A., Sollmann, N., Vu, D., Adewole, M., Albrecht, J., Anazodo, U., Chai, R., Chung, V., Faghani, S., Farahani, K., Kazerooni, A. F., Iglesias, E., Kofler, F., Li, H., Linguraru, M. G., Menze, B., Moawad, A. W., Velichko, Y., Wiestler, B., Altes, T., Basavasagar, P., Bendszus, M., Brugnara, G., Cho, J., Dhemesh, Y., Fields, B. K., Garrett, F., Gass, J., Hadjiiski, L., Hattangadi-Gluth, J., Hess, C., Houk, J. L., Isufi, E., Layfield, L. J., Mastorakos, G., Mongan, J., Nedelec, P., Nguyen, U., Oliva, S., Pease, M. W., Rastogi, A., Sinclair, J., Smith, R. X., Sugrue, L. P., Thacker, J., Vidic, I., Villanueva-Meyer, J., White, N. S., Aboian, M., Conte, G. M., Dale, A., Sabuncu Mert R.and Seibert, T. M., Weinberg, B., Abayazeed, A., Huang, R., Turk, S., Rauschecker, A. M., Farid, N., Vollmuth, P., Nada, A., Bakas, S., Calabrese, E., D., R. J. (2024). The 2024 Brain Tumor Segmentation (BraTS) Challenge: Glioma Segmentation on Post-treatment MRI. In: *arXiv e-prints*, arXiv–2405. DOI: https://doi.org/10.48550/arXiv.2405.18368.

Feng, Y., Ma, B., Zhang, J., Zhao, S., Xia, Y., Tao, D. (2022). Fiba: Frequency-injection based backdoor attack in medical image analysis. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20876–20885. DOI: https://doi.org/10.48550/arXiv.2112.01148.

Gayatri, E., Aarthy, S. L. (2024). Reduction of overfitting on the highly imbalanced ISIC-2019 skin dataset using deep learning frameworks. In: *Journal of X-Ray Science and Technology* 32.1, pp. 53–68. DOI: 10.3233/XST-230204.

Gouda, W., Sama, N., Al-Waakid, G., Humayun, M., Jhanjhi, N. (2022). Detection of Skin Cancer Based on Skin Lesion Images Using Deep Learning. In: *Healthcare* 10, p. 1183. DOI: 10.3390/healthcare10071183.

Gu, T., Dolan-Gavitt, B., Garg, S. (2017). BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. In: *arXiv e-prints*, arXiv–1708. DOI: https://doi.org/10.48550/arXiv.1708.06733.

Hossain, T., Chandro, B. S. (2024). Perception and Localization of Macular Degeneration Applying Convolutional Neural Network, ResNet and Grad-CAM. In: *arXiv e-prints*, arXiv–2404. DOI: https://doi.org/10.48550/arXiv.2404.15918.

Ivanovs, M., Kadikis, R., Ozols, K. (2021). Perturbation-based methods for explaining deep neural networks: A survey. In: *Pattern Recognition Letters* 150, pp. 228–234. DOI: https://doi.org/10.1016/j.patrec.2021.06.030.

Jaisakthi, S. M., Mirunalini, P., Chandrabose, A., Rajagopal, A. (2023). Classification of skin cancer from dermoscopic images using deep neural network architectures. In: *Multimedia Tools and Applications* 82.10, pp. 15763–15778. DOI: 10.1007/s11042-022-13847-3.

Kindermans, P.-J., Schütt, K., Müller, K.-R., Dähne, S. (2016). Investigating the influence of noise and distractors on the interpretation of neural networks. In: *arXiv e-prints*, arXiv–1611. DOI: https://doi.org/10.48550/arXiv.1611.07270.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., Reblitz-Richardson, O. (2020). Captum: A unified and generic model interpretability library for PyTorch. In: *arXiv e-prints*, arXiv–2009. DOI: https://doi.org/10.48550/arXiv.2009.07896.

Kurtansky, N., Primiero, C., Betz-Stablein, B., Combalia, M., Guitera, P., Halpern, A., Kentley, J., Kittler, H., Liopyris, K., Malvehy, J., Rinner, C., Tschandl, P., Weber, J., Rotemberg, V., Soyer, P. (2024). Effect of patient-contextual skin images in human- and artificial intelligence-based diagnosis of melanoma: Results from the 2020 SIIM-ISIC melanoma classification challenge. In: *Journal of the European Academy of Dermatology and Venereology : JEADV*. DOI: 10.1111/jdv.20479.

Li, Y., Jiang, Y., Li, Z., Xia, S.-T. (2022). Backdoor learning: A survey. In: *IEEE transactions on neural networks and learning systems* 35.1, pp. 5–22. DOI: https://doi.org/10.48550/arXiv.2007.08745.

Li, Z., Chen, Z., Che, X., Wu, Y., Huang, D., Ma, H., Dong, Y. (2022). A classification method for multi-class skin damage images combining quantum computing and Inception-ResNet-V1. In: *Frontiers in Physics* 10. DOI: 10.3389/fphy.2022.1046314.

Lihacova, I., Bondarenko, A., Chizhov, Y., Uteshev, D., Bliznuks, D., Kiss, N., Lihachev, A. (2022). Multi-Class CNN for Classification of Multispectral and Autofluorescence Skin Lesion Clinical Images. In: *Journal of Clinical Medicine* 11.10. DOI: https://doi.org/10.3390/jcm11102833.

Liu, K., Dolan-Gavitt, B., Garg, S. (2018). Fine-pruning: Defending against backdooring attacks on deep neural networks. In: *International symposium on research in attacks, intrusions, and defenses*. Springer, pp. 273–294. DOI: https://doi.org/10.3390/jcm11102833.

Manole, I., Butacu, A.-I., Bejan, R. N., Tiplica, G.-S. (2024). Enhancing Dermatological Diagnostics with EfficientNet: A Deep Learning Approach. In: *Bioengineering (Basel)* 11.8. DOI: https://doi.org/10.3390/bioengineering11080810.

Mo, Y., Huang, H., Li, M., Li, A., Wang, Y. (2024). TERD: A Unified Framework for Safeguarding Diffusion Models Against Backdoors. In: *arXiv e-prints*, arXiv–2409. DOI: https://doi.org/10.48550/arXiv.2409.05294.

Mohamed, M., Mahesh, T. R., Kumar Vinoth, V., Guluwadi, S. (2024). Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with Resnet 50. In: *BMC Medical Imaging* 24. DOI: 10.1186/s12880-024-01292-7.

Olayah, F., Senan, E., Ahmed, I., Awaji, B. (2023). AI Techniques of Dermoscopy Image Analysis for the Early Detection of Skin Lesions Based on Combined CNN Features. In: *Diagnostics* 13. DOI: 10.3390/diagnostics13071314.

Pal, S., Wang, R., Yao, Y., Liu, S. (2023). Towards Understanding How Self-training Tolerates Data Backdoor Poisoning. In: *arXiv e-prints*, arXiv–2301. DOI: https://doi.org/10.48550/arXiv.2301.08751.

Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M., Dusza, S., Guitera, P., Gutman, D., Halpern, A., Helba, B., Kittler, H., Kose, K., Langer, S., Lioprys, K., Malvehy, J., Musthaq Shenaraand Nanda, J., Reiter, O., Shih, G., Stratigos, A., Tschandl, P., Soyer, J. W., Peter, H. (2021). A patient-centric dataset of images and metadata for identifying melanomas using clinical context. In: *Scientific data* 8.1, p. 34. DOI: https://doi.org/10.1038/s41597-021-00815-z.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D. (2020). Grad-CAM: visual explanations from deep networks via gradient-based localization. In: *International journal of computer vision* 128, pp. 336–359. DOI: https://doi.org/10.48550/arXiv.1610.02391.

Shen, S., Tople, S., Saxena, P. (2016). Auror: defending against poisoning attacks in collaborative deep learning systems. In: *Proceedings of the 32nd Annual Conference on Computer Security Applications*. ACSAC '16. Association for Computing Machinery, pp. 508–519. DOI: 10.1145/2991079.2991125.

Shrikumar, A., Greenside, P., Shcherbina, A., Kundaje, A. (2016). Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. In: *arXiv e-prints*, arXiv–1605. DOI: https://doi.org/10.48550/arXiv.1605.01713.

Simonyan, K., Vedaldi, A., Zisserman, A. (2013). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In: *arXiv e-prints*, arXiv–1312. DOI: https://doi.org/10.48550/arXiv.1312.6034.

Singh, S. K., Banerjee, S., Chakraborty, A., Bandyopadhyay, A. (2023). Classification of Melanoma Skin Cancer Using Inception-ResNet. In: *Frontiers of ICT in Healthcare*. Ed. by J. K. Man-

dal, D. De. Vol. 519. Lecture Notes in Networks and Systems. Springer, Singapore, pp. 65–74. DOI: 10.1007/978-981-19-5191-6_6.

Sundararajan, M., Taly, A., Yan, Q. (2017). Axiomatic attribution for deep networks. In: *International conference on machine learning*. PMLR, pp. 3319–3328. DOI: https://doi.org/10.48550/arXiv.1703.01365.

Venugopal, V., Raj, N., Nath, M., Stephen, N. (2023). A deep neural network using modified EfficientNet for skin cancer detection in dermoscopic images. In: *Decision Analytics* 8, p. 100278. DOI: 10.1016/j.dajour.2023.100278.

Ya, M., Li, Y., Dai, T., Wang, B., Jiang, Y., Xia, S.-T. (2023). Towards faithful xai evaluation via generalization-limited backdoor watermark. In: *The Twelfth International Conference on Learning Representations*.

Zhang, Z., Liu, Q., Wang, Z., Lu, Z., Hu, Q. (2023). Backdoor defense via deconfounded representation learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12228–12238. DOI: https://doi.org/10.48550/arXiv.2303.06818.

Zhou, X. (2018). Understanding the Convolutional Neural Networks with Gradient Descent and Backpropagation. In: *Journal of Physics: Conference Series* 1004, p. 012028. DOI: 10.1088/1742-6596/1004/1/012028.

Received June 4, 2025 , revised December 30, 2025, accepted Ferbruary 13, 2026