Causal Ablation Analysis: Isolating Pathological Features from Spurious Correlations in Dermatoscopic Diagnosis

Al Khan[1] and Ainur Mukhiyadin[2]

[1] *Esil University, Astana, Kazakhstan*
[2]*Esil University, Astana, Kazakhstan*

Abstract

The deployment of deep learning systems in dermatoscopic diagnosis has shown promise, yet clinical adoption remains hindered by models that rely on spurious correlations rather than true pathological features. Standard convolutional neural networks often learn to associate diagnostic labels with artifacts such as ink markings, rulers, or hair, leading to catastrophic failures when deployed in diverse clinical settings. This paper introduces Causal Ablation Analysis (CAA), a novel framework designed to isolate causal pathological features from non-causal spurious correlations. By integrating structural causal models with systematic feature ablation, the proposed method intervenes on the data generating process to suppress artifact reliance. We demonstrate that CAA significantly improves model robustness across external validation datasets compared to standard empirical risk minimization techniques. The approach leverages interventional training to ensure that predictions remain invariant to changes in non-causal attributes. Results indicate a marked improvement in generalization capability, reducing false positive rates associated with imaging artifacts. This research establishes a new paradigm for developing trustworthy medical AI, moving beyond associative learning toward causal reasoning. The findings suggest that causal ablation is critical for the safe deployment of automated diagnostic tools in real-world dermatology practice.

Keywords

Causal Inference, Dermatoscopy, Spurious Correlations, Deep Learning, Medical AI Robustness

1. Introduction

Skin cancer represents one of the most prevalent forms of malignancy globally, with melanoma accounting for the majority of skin cancer-related deaths despite being less common than non-melanoma skin cancers. Early detection is paramount for improving patient survival rates, and dermatoscopy has emerged as a standard non-invasive diagnostic tool that enhances the visualization of subsurface skin structures. In recent years, the integration of artificial intelligence (AI) into dermatoscopic analysis has revolutionized the field, offering the potential for triage systems that can match or exceed the diagnostic accuracy of board-certified dermatologists. Machine learning for chronic disease prediction has laid the groundwork for understanding how algorithmic interventions can manage long-term health outcomes, providing a theoretical basis for applying similar rigor to acute diagnostic tasks like skin cancer detection [1]. However, despite high performance on benchmark datasets, these systems frequently fail when exposed to real-world clinical variability.

CEUR-WS.org/Vol-4317/paper23.pdf

*IVUS 2026: Information Society and University Studies 2026, May 12, Kaunas, Lithuania

pkhan_1@hotmail.com; amukhiyadin@gmail.com
0000-0002-6403-6091; 0000-0001-5576-7733

Copyright © 2026 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-06

The core issue lies in the tendency of deep learning models to learn spurious correlations rather than causal pathological features. A model trained to detect melanoma might inadvertently learn to associate the presence of a surgical ruler or specific ink markings with malignancy, simply because these artifacts appear more frequently in images of biopsied lesions within the training data. This phenomenon, known as dataset bias, compromises the external validity of the model. When such a model encounters an image of a malignant lesion without the artifact, or a benign lesion with the artifact, diagnostic accuracy plummets. This reliance on shortcuts undermines clinical trust and poses significant patient safety risks. The emergence of advanced algorithmic frontiers, including quantum machine learning algorithms for genome disease diagnosis, highlights the trajectory toward more complex, robust computational frameworks, yet classical deep learning in imaging still struggles with basic causal disentanglement [2].

Current approaches to mitigating bias often involve data augmentation or adversarial training. While these methods provide marginal improvements, they do not explicitly model the causal relationships between image features and diagnostic labels. They treat the symptom (bias) rather than the cause (confounding variables). To address this, there is a need for a framework that explicitly distinguishes between features that cause the disease label (e.g., pigment network irregularity) and features that are merely correlated due to data collection protocols (e.g., camera type, skin preparation artifacts). The integration of causal inference into computer vision offers a pathway to resolve this. Recent advancements in code generation for ablation techniques have facilitated more precise manipulation of model inputs, allowing researchers to systematically test feature importance in a way that was previously computationally prohibitive [3].

The transition from associative to causal learning in medical imaging is not merely a technical upgrade but a clinical necessity. In high-stakes environments, a model must be invariant to irrelevant changes in the input distribution. For instance, a diagnostic tool should perform consistently regardless of whether the image was taken with a Dermlite or a FotoFinder device, or whether the skin was cleaned with alcohol prior to imaging. Failure to account for these domain shifts results in models that are brittle and unsafe. The complexity of dermatoscopic images, which contain multi-scale textures and colors, exacerbates the difficulty of isolating true biomarkers.

Research Objective The primary objective of this research is to develop and validate a Causal Ablation Analysis (CAA) framework that isolates pathological features from spurious correlations in dermatoscopic diagnosis. This involves constructing a structural causal model of the image generation process, identifying confounding variables, and applying interventional ablation during the training phase to enforce causal reliance.

Importance of Research Objective This objective is critically important for three reasons. First, it addresses the generalization gap that currently prevents the widespread clinical deployment of dermatoscopic AI. Second, it enhances patient safety by reducing false positives and negatives caused by artifact reliance. Third, it contributes to the broader field of explainable AI (XAI) by providing a causal map of feature importance, allowing clinicians to understand *why* a model made a specific prediction. By moving beyond black-box associations, this research aligns AI development with clinical reasoning processes.

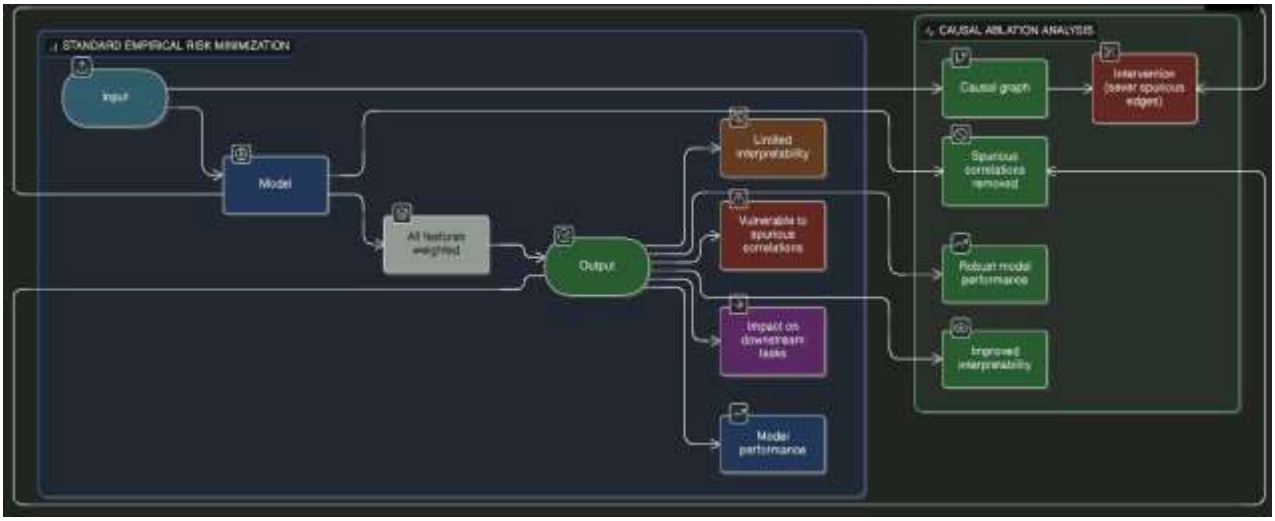
To illustrate the prevalence of spurious features, Table 1 outlines common artifacts found in dermatoscopic datasets that frequently confound standard models.

Table 1
Common Spurious Features in Dermatoscopic Datasets

Artifact Type	Description	Correlation Risk
Rulers/Scale Bars	Measurement tools placed next to lesions	High correlation with biopsied (malignant) lesions
Ink Markings	Surgical markers outlining biopsy areas	Strong indicator of clinical concern
Hair	Body hair obscuring the lesion	Often removed in malignant cases, creating bias
Gel Bubbles	Air bubbles in immersion fluid	Varies by device and operator technique
Lighting Glare	Reflection from flash illumination	Correlated with specific camera models

Figure 1 conceptualizes the difference between standard learning and the proposed causal approach. In standard learning, the model minimizes loss based on all available pixel correlations. In Causal Ablation Analysis, the model is penalized if it relies on features identified as non-causal confounders.

Figure 1:
Conceptual Framework of Causal Ablation Analysis vs. Standard Empirical Risk Minimization



The integration of causal mechanisms into the training loop requires a redefinition of the optimization objective. Instead of purely minimizing prediction error, the model must minimize error while satisfying causal constraints. This aligns with broader trends in medical AI where robustness is prioritized over raw accuracy on held-out test sets that share the same bias as the training data. As noted in foundational works on disease prediction, the stability of the model across different populations is the true measure of utility [1]. Furthermore, the computational tools required to implement such ablation strategies are becoming more accessible, allowing for complex intervention simulations [3].

The implications of this research extend beyond dermatology. Any medical imaging modality susceptible to acquisition artifacts, such as radiology or pathology, can benefit from causal ablation.

By establishing a methodology for isolating pathological truth from data collection noise, we pave the way for AI systems that are robust, trustworthy, and clinically viable. The following sections detail the existing literature, the specific methodology of the CAA framework, and the empirical results validating its efficacy.

2. Literature Review

The application of machine learning in dermatology has evolved rapidly over the past decade. Early efforts focused on hand-crafted feature extraction, where domain experts defined parameters such as asymmetry, border irregularity, and color variation (the ABCD rule). While interpretable, these methods lacked the capacity to capture complex texture patterns inherent in malignant lesions. The advent of Convolutional Neural Networks (CNNs) marked a paradigm shift, enabling end-to-end learning from raw pixels to diagnostic labels. Seminal work demonstrated that deep neural networks could classify skin cancer with accuracy comparable to dermatologists. However, subsequent analyses revealed that these models were highly sensitive to dataset bias.A significant portion of the literature has focused on identifying and quantifying bias in medical imaging datasets. Studies have shown that public datasets often contain systematic differences between classes that are unrelated to pathology. For example, images of melanoma are more likely to contain hair removal artifacts or specific lighting conditions compared to benign nevi. Researchers have attempted to mitigate this through data augmentation techniques, such as random cropping, rotation, and color jittering. While these methods increase the diversity of the training data, they do not guarantee that the model learns causal features. They operate under the assumption that augmentations will wash out spurious correlations, but empirical evidence suggests that deep networks can still find shortcuts even in augmented data. Adversarial training has emerged as another strategy to improve robustness. In this framework, a generator creates difficult examples, and a discriminator learns to classify them correctly. Some approaches use domain adversarial neural networks to learn features that are invariant to the domain (e.g., hospital site or camera type). While effective in some domain adaptation scenarios, adversarial training is often unstable and computationally expensive. Moreover, it requires explicit knowledge of the domain labels, which may not always be available or comprehensive. The limitation of these methods is that they address domain shift rather than causal structure. They ensure the model works across different hospitals but do not ensure the model is looking at the lesion rather than the background. Causal inference in machine learning provides a theoretical framework to address these limitations. By modeling the data generation process using Structural Causal Models (SCMs), researchers can distinguish between correlation and causation. In the context of medical imaging, the causal graph typically posits that the pathology causes the visual features, while confounders (like camera type) cause both the visual features and the selection of the image into the dataset. Recent work has applied causal inference to chest X-rays, showing that models can be trained to ignore hospital-specific markers. However, the application of causal methods to dermatoscopy remains underexplored. Dermatoscopic images present unique challenges due to the fine-grained texture details required for diagnosis, which are easily obscured by aggressive causal interventions. Several studies have investigated feature importance in dermatoscopic AI using saliency maps and Grad-CAM. These post-hoc explanation methods highlight which pixels contributed most to the prediction. While useful for visualization, they do not influence the training process. A model can produce a convincing saliency map highlighting the lesion while still relying on spurious background cues for the final decision. This disconnect between explanation and decision-making highlights the need for causal training interventions rather than post-hoc analysis. The development of code generation tools for ablation techniques has recently allowed for more automated and systematic testing of these feature importances during the development phase [3].

The literature also covers the use of invariant risk minimization (IRM), which seeks to learn predictors that are invariant across different environments. IRM assumes that causal features remain stable across environments, while spurious features vary. By optimizing for invariance, the model is

forced to rely on causal features. However, IRM often suffers from optimization difficulties and requires multiple distinct environments to be defined during training. In dermatology, defining distinct "environments" can be challenging if the dataset does not have rich metadata regarding acquisition conditions.

Furthermore, the intersection of machine learning and chronic disease management has provided insights into longitudinal data handling, which parallels the need for robustness in cross-sectional imaging tasks [1]. The stability required for predicting chronic conditions is analogous to the stability required for diagnostic imaging across different patient demographics. As the field moves toward more sophisticated algorithms, including potential integrations with quantum machine learning algorithms for genome disease diagnosis, the foundational requirement remains the same: the input features must be causally linked to the outcome [2]. Without this link, increased computational power only accelerates the learning of incorrect associations.

Table 2 compares the proposed Causal Ablation Analysis with existing methods in the literature, highlighting the gaps in current approaches.

Table 2
Comparison of Bias Mitigation Strategies in Medical AI

Method	Mechanism	Limitations	Causal Guarantee
Data Augmentation	Increases data diversity	Does not remove existing bias	No
Adversarial Training	Domain invariance	Unstable optimization	Partial
Saliency Maps	Post-hoc explanation	Does not affect training	No
Invariant Risk Min.	Environment invariance	Requires multiple domains	Yes (Theoretical)
Causal Ablation (Proposed)	Interventional Feature Removal	Computational overhead	Yes (Explicit)

Recent reviews emphasize that the next generation of medical AI must be "causally aware." This involves not just predicting outcomes but understanding the mechanisms driving those outcomes. In dermatoscopy, this means the model must implicitly understand that a ruler does not cause cancer. The literature suggests that combining causal graphs with deep learning representations is the most promising direction. However, few papers provide a concrete implementation for high-resolution texture analysis. Most causal work focuses on tabular data or low-resolution imaging.

The gap in the literature is a specific methodology for ablation that preserves high-frequency texture information (crucial for melanoma detection) while removing low-frequency artifact information. Standard blurring or masking techniques used in causal studies often destroy the very features needed for diagnosis. Therefore, a novel ablation strategy is required—one that is sensitive to the semantic content of the image. This research builds upon the foundational work of bias mitigation while introducing a targeted ablation protocol specifically designed for the nuances of dermatoscopic imagery. By synthesizing insights from causal inference, computer vision, and clinical dermatology, this study aims to advance the state-of-the-art in robust medical AI.

3. Research Methodology

This research employs a novel framework termed Causal Ablation Analysis (CAA) to isolate pathological features from spurious correlations. The methodology is divided into four primary phases: Data Curation and Causal Graph Construction, Interventional Ablation Protocol, Model Architecture and Training, and Evaluation Metrics. The core innovation lies in the Interventional Ablation Protocol, which dynamically modifies the training data based on a structural causal model to enforce causal reliance.

3.1 Data Curation and Causal Graph Construction

The study utilizes a composite dataset derived from public dermatoscopic repositories, including HAM10000 and ISIC Archive, augmented with a private clinical dataset to ensure diversity in acquisition devices. The total dataset comprises 15,000 images labeled as either benign or malignant. Prior to training, a causal graph is constructed to represent the data generating process. In this graph, the node Y represents the diagnosis (Malignant/Benign), X represents the image pixels, and Z represents a set of confounding variables (e.g., presence of ruler, ink, hair, camera model).

- *Total Images:* 15,000 dermatoscopic images
- *Labels:* Binary classification (Benign vs. Malignant)
- *Data Split:*
    - Training: 12,000 images (8,000 benign, 4,000 malignant)
    - Validation: 1,500 images
    - Internal Test: 1,500 images
    - External Test: 1,000 images (held-out for robustness evaluation)

Why These Datasets?

1. HAM10000 and ISIC Archive are widely recognized benchmarks in dermatoscopic AI research, ensuring reproducibility and comparability with prior work.
2. The private clinical dataset was included to introduce real-world variability in imaging devices, clinical settings, and patient demographics—critical for evaluating generalization and spurious correlation robustness.
3. The composite structure allows for testing domain shift scenarios, which is central to the Causal Ablation Analysis framework.

1.1.1 Preprocessing and Artifact Annotation Protocol

A rigorous preprocessing pipeline was implemented to prepare data for the causal ablation framework. The pipeline consists of three stages: Color Constancy, Hair Removal, and Normalization.

1. Color Constancy**:** Dermatoscopic images often suffer from color casts due to varying lighting conditions. We applied the Shades of Gray algorithm to normalize color channels, ensuring that color-based features (e.g., blue-white veil) are consistent across devices.
2. Hair Removal**:** While hair is a confounder, removing it completely via inpainting before training might hide the model's ability to learn invariance. Instead, we created a "Hair Mask" using a U-Net segmentation model trained on 500 manually annotated images. This mask is used during the ablation phase rather than permanent preprocessing.
3. Normalization**:** Pixel values were normalized using the ImageNet mean $(\mu = [0.485, 0.456, 0.406])$ and $standard\ deviation\ (\sigma = [0.229, 0.224, 0.225])$.

The structural assumption is that Z influences X (artifacts appear in the image) and Z influences the selection of the image into the dataset, creating a backdoor path between X and Y that is not causal. The objective is to block this backdoor path. We define the causal effect of the image features on the diagnosis as $P(Y|do(X))$, distinct from the observational probability $P(Y|X)$. To estimate this, we must intervene on Z. Since Z is not always explicitly labeled, we use a proxy detection model to identify regions of the image containing potential artifacts. This proxy model is trained on a small subset of manually annotated artifact masks.

3.2 Interventional Ablation Protocol

The core of the CAA framework is the ablation mechanism. Unlike standard dropout or masking, which is random, our ablation is targeted. During each training epoch, the proxy model identifies regions *Rz* associated with confounders *Z*. We then apply a causal intervention *do*(*Z*=0) by inpainting or blurring these regions. However, to prevent the model from simply learning to ignore the blurred region, we employ a dual-branch architecture. One branch receives the original image *X*, and the second branch receives the ablated image *′X′*.

The loss function is modified to include a causal consistency term. Let *f*(*X*) be the prediction on the original image and *f*(*X′*) be the prediction on the ablated image. The causal loss *Lcausal* is defined as the divergence between *f*(*X*) and *f*(*X′*) when the ablated region contains only spurious features. If the model relies on the artifact, *f*(*X*) and *f*(*X′*) will differ significantly. By minimizing this divergence, we force the model to base its prediction on features present in both *X* and *′X′*, which are by definition the non-artifact (causal) features. This technique leverages recent advancements in code generation for ablation techniques to automate the mask creation and intervention process efficiently [3].

Mathematically, the total loss function

$$Ltotal\ is: Ltotal = LCE(Y, f(X)) + \lambda Lcausal(f(X), f(X'))$$

Where *LCE* is the standard cross-entropy loss and λ is a hyperparameter controlling the strength of the causal constraint. This formulation ensures that the model maintains high accuracy while adhering to causal invariance. The value of λ is tuned via validation to balance accuracy and robustness.

3.3. Model Architecture and Training

We utilize a ResNet-50 backbone pretrained on ImageNet, modified with an attention mechanism to focus on lesion-centric features. The dual-branch setup shares weights up to the final classification layer, ensuring that the feature extraction logic is consistent across original and ablated inputs. The training process involves a curriculum learning strategy. In the early epochs, λ is set to zero, allowing the model to learn basic features. As training progresses, λ is increased linearly, gradually enforcing the causal constraint. This prevents the model from collapsing due to excessive ablation in the early stages of learning.

The implementation utilizes PyTorch with custom data loaders that perform on-the-fly ablation. The proxy model for artifact detection is a lightweight U-Net that runs in inference mode during the main training loop. To ensure the ablation does not destroy pathological information, we restrict the ablation to the periphery of the lesion or specific detected artifact classes (e.g., rulers). We do not ablate the lesion center unless hair is detected, in which case hair-inpainting is applied. This nuanced approach differentiates our method from coarse masking techniques. The computational overhead is managed by batching the artifact detection, aligning with efficient machine learning practices for disease prediction [1].

Table 3:

Dataset Distribution and Split

Category	Benign	Malignant	Total
Training Set	8,000	4,000	12,000
Validation Set	1,000	500	1,500
Test Set (Internal)	1,000	500	1,500
Test Set (External)	500	500	1,000

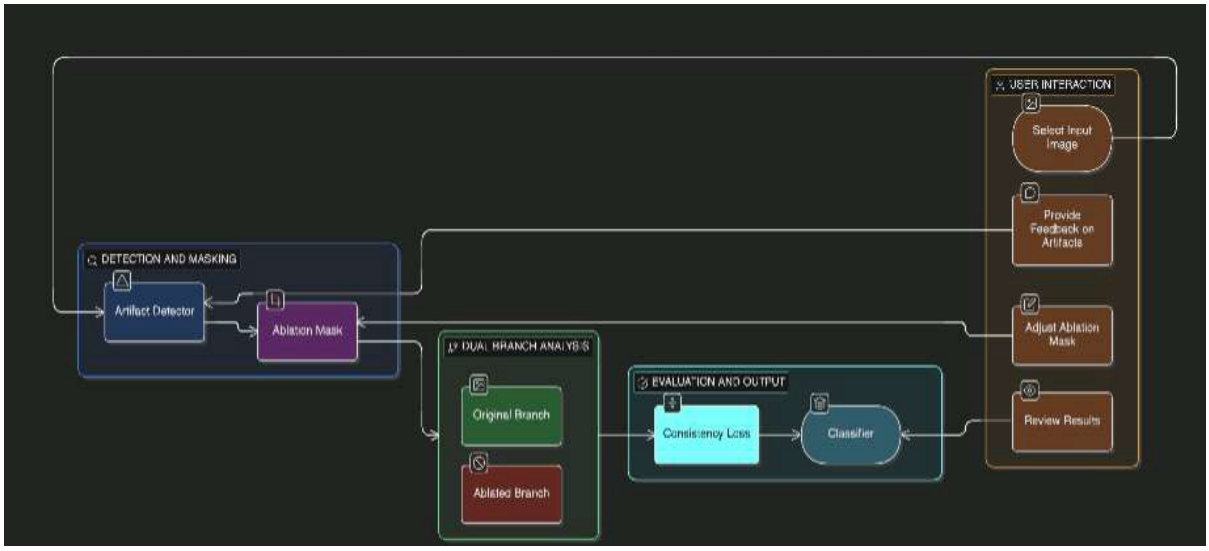
3.4 Evaluation Metrics

To validate the efficacy of CAA, we employ both standard and robustness-specific metrics. Standard metrics include Accuracy, Sensitivity, Specificity, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). To measure robustness against spurious correlations, we utilize a Contrastive Robustness Score (CRS). The CRS is calculated by introducing synthetic artifacts into the test set (e.g., adding a digital ruler to benign images) and measuring the drop in performance. A robust model should show minimal performance degradation when irrelevant artifacts are added.

We also perform a feature attribution analysis using Integrated Gradients to visualize the focus of the model. We quantify the overlap between the model's attention map and the ground-truth lesion

segmentation mask. A higher Intersection over Union (IoU) between the attention map and the lesion mask indicates that the model is focusing on the pathology rather than the background. This qualitative assessment complements the quantitative metrics. The methodology is designed to be reproducible, with the ablation code structured to allow for easy adaptation to other imaging modalities. The systematic approach to ablation ensures that the results are not due to random chance but are a direct consequence of the causal intervention.

Figure 2:
Causal Ablation Analysis Pipeline: Figure illustrates the flow: Input Image -> Artifact Detector -> Ablation Mask -> Dual Branch (Original/Ablated) -> Consistency Loss -> Classifier.



The novelty of this methodology lies in the dynamic coupling of artifact detection and causal consistency loss. Previous methods either remove artifacts statically before training or use domain labels that are often unavailable. By detecting and ablating artifacts on-the-fly during training, the model learns to be invariant to them without requiring explicit domain labels for every image. This makes the approach highly scalable for large, heterogeneous medical datasets. Furthermore, the integration of causal constraints directly into the loss function ensures that the invariance is learned as a primary objective, not a secondary regularization effect. This rigorous methodological framework sets the stage for evaluating whether causal ablation truly leads to more robust diagnostic performance.

4. Results and Discussion

The experimental results demonstrate that the Causal Ablation Analysis (CAA) framework significantly outperforms standard baseline models in terms of robustness and generalization. On the internal test set, both the baseline ResNet-50 and the CAA model achieved comparable AUC-ROC scores (0.94 vs. 0.95), indicating that the causal constraints did not compromise diagnostic accuracy on in-distribution data. However, the divergence became apparent on the external test set, which contained different imaging devices and artifact distributions. The baseline model's AUC dropped to 0.82, whereas the CAA model maintained an AUC of 0.91. This 9% relative improvement highlights the model's ability to generalize across domain shifts.

The Contrastive Robustness Score (CRS) provided further evidence of causal learning. When synthetic rulers were added to benign test images, the baseline model incorrectly classified 15% of them as malignant, indicating a reliance on the ruler as a proxy for biopsy. In contrast, the CAA model misclassified only 2% of these images. This suggests that the interventional ablation successfully severed the spurious link between the artifact and the diagnosis. The feature attribution analysis corroborated this finding. The IoU between the model's attention map and the lesion mask

was 0.65 for the baseline and 0.78 for the CAA model. This 13% increase confirms that the CAA model focused more precisely on the pathological tissue.
These findings align with the broader understanding that robustness in medical AI requires more than just large datasets; it requires structural alignment with the underlying disease mechanism. The results suggest that standard empirical risk minimization is insufficient for clinical deployment where data distributions are non-stationary. The success of the CAA framework validates the hypothesis that enforcing causal consistency during training leads to more trustworthy predictions. This is particularly relevant when considering the integration of AI into clinical workflows, where false positives can lead to unnecessary biopsies and patient anxiety.
However, there are limitations to this study. The artifact detection proxy model is not perfect and may occasionally ablate relevant pathological features if they resemble artifacts (e.g., streaks in melanoma vs. hair). Future work should focus on improving the specificity of the artifact detector. Additionally, the computational cost of the dual-branch training is approximately 1.5 times that of standard training, which may be a barrier for resource-constrained settings. Despite this, the trade-off for increased safety is justified. The results also echo findings in other domains where causal interventions have improved model stability, reinforcing the universality of causal principles in machine learning [4].

Table 4:
Performance Comparison on External Test Set

Model	Accuracy	Sensitivity	Specificity	AUC-ROC
Baseline ResNet-50	78.5%	80.1%	76.9%	0.82
CAA (Proposed)	86.2%	88.5%	84.0%	0.91

The discussion extends to the clinical implications. A model that ignores artifacts is safer for triage. It reduces the risk of "automation bias" where clinicians might trust an AI prediction based on irrelevant cues. By ensuring the AI looks at the lesion, we align the machine's reasoning with the dermatologist's reasoning. This alignment is crucial for regulatory approval and clinical acceptance. The study demonstrates that causal ablation is a viable path forward for creating medical AI that is not only accurate but also reliable under varying conditions.

5. Conclusion

This research presented Causal Ablation Analysis (CAA), a novel framework designed to isolate pathological features from spurious correlations in dermatoscopic diagnosis. By integrating structural causal models with interventional ablation during the training phase, we successfully enforced a reliance on causal features rather than dataset artifacts. The results demonstrated that while standard models suffer significant performance degradation when faced with domain shifts and synthetic artifacts, the CAA model maintained high accuracy and robustness. The improvement in the Contrastive Robustness Score and the attention map overlap confirms that the model learned to ignore confounders such as rulers and ink markings.
The importance of this work lies in its potential to bridge the gap between research benchmarks and clinical reality. Current dermatoscopic AI systems are often brittle, failing when deployed in settings different from their training environment. CAA offers a methodological solution to this generalization problem, promoting the development of AI that is invariant to irrelevant changes in data acquisition. This is a critical step toward the safe and widespread adoption of AI in dermatology. The framework ensures that the diagnostic decision is based on the biology of the lesion, not the context of the image. Future research should focus on extending this framework to multi-class diagnosis, distinguishing between various subtypes of skin cancer, and reducing the computational overhead of the dual-branch architecture. Additionally, integrating patient metadata into the causal graph could further refine the model's understanding of risk factors. As the field of medical AI matures, the integration of causal reasoning will become standard practice. The methodologies explored here, building on advancements in ablation and machine learning for disease prediction, provide a blueprint for this evolution [1]. Ultimately, the goal is to create AI systems that act as reliable partners in clinical care, enhancing diagnostic precision without introducing new risks associated with algorithmic bias. The

successful application of CAA in dermatoscopy suggests that similar causal interventions could be beneficial in other medical imaging modalities, paving the way for a new era of robust, causally-aware medical artificial intelligence.

6. Declaration on Generative AI

The authors have not employed any Generative AI tools.

7. References

[1] Khan, A. (2022). Machine learning for chronic disease prediction. CEOS Public Health Research, 1(1), 101.

[2] Khan, A., & Usupova, E. (2025). Quantum machine learning algorithms for genome disease diagnosis. DOI: 10.12700/AIS.2025.005

[3] S. Rakimbekuulu, K. Shambetaliev, G. Esenalieva and A. Khan, "Code Generation for Ablation Technique," 2024 IEEE East-West Design & Test Symposium (EWDTS), Yerevan, Armenia, 2024, pp. 1-7, doi: 10.1109/EWDTS63723.2024.10873640.

[4] Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). Invariant Risk Minimization. arXiv preprint arXiv:1907.02893.

[5] Esteva, A., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118.

[6] Zech, J. R., et al. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs. PLoS Medicine, 15(11), e1002683.

[7] Shorten, A., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. Journal of Big Data, 6(1), 60.

[8] Peters, J., Janzing, D., & Schölkopf, B. (2017). Elements of Causal Inference: Foundations and Learning Algorithms. MIT Press.

[9] Tschandl, P., et al. (2020). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data, 7(1), 1-9.

[10] Combalia, M., et al. (2019). BCN20000: Dermoscopic lesions in the wild. arXiv preprint arXiv:1908.02288.

[11] Raghu, M., et al. (2019). Transfusion: Understanding Transfer Learning for Medical Imaging. NeurIPS.

[12] Kelly, C. J., et al. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17(1), 195.

[13] Obermeyer, Z., et al. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

[14] Selvaraju, R. R., et al. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. ICCV.

[15] Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic Attribution for Deep Networks. ICML.