Detecting and Mitigating the Clever Hans Effect in Medical Imaging: A Scoping Review

Detecting and Mitigating the Clever Hans Effect in Medical Imaging: A Scoping Review

Constanza Vásquez-Venegas (ORCID: 0000-0003-0052-7649)[1,2*], Chenwei Wu[3], Saketh Sundar[4], Renata Prôa[5,6], Francis Joshua Beloy (ORCID: 0000-0002-6148-2811)[7], Jillian Reeze Medina (ORCID: 0000-0002-3056-9447)[8], Megan McNichol[9], Krishnaveni Parvataneni[10], Nicholas Kurtzman[11], Felipe Mirshawka[12], Marcela Aguirre-Jerez[13], Daniel K. Ebner[14], Leo A. Celi (ORCID: 0000-0001-6712-6626)[15,16,17]

[1]Scientific Image Analysis Lab, Faculty of Medicine, Universidad de Chile, Santiago, 8380453, RM, Chile.
[2]Department of Computer Science, Faculty of Engineering, Universidad de Concepción, Concepción, 4030000, Biobio, Chile.
[3]University of Michigan, Ann Arbor, 48105, Michigan, United States.
[4]Harvard College, Harvard University, Cambridge, Massachusetts, United States.
[5]Harvard School of Public Health, Harvard University, Boston, Massachusetts, United States.
[6]Hospital Israelita Albert Einstein, São Paulo, Brazil.
[7]Montefiore St. Luke's Cornwall, Newburgh, New York,United States.
[8]Manila Central University College of Medicine, Caloocan, Philippines.
[9]Division of Knowledge Services, Department of Information Services, Beth Israel Lahey Health, Cambridge, Massachusetts, United States.
[10]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States.
[11]Emory School of Medicine, Atlanta, Georgia,United States.
[12]School of Biomedical Engineering, Faculdade Israelita de Ciências da Saúde Albert Einstein, Hospital Israelita Albert Einstein, São Paulo, Brazil.

[13]Digital Health Deparment, Fundación Arturo Lopez Perez, Santiago, 7500853, Chile.
[14]Department of Radiation Oncology, Mayo Clinic, Rochester, Minnesota,United States.
[15]Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States.
[16]Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, United States.
[17]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States.

*Corresponding author(s). E-mail(s): covasquezv@inf.udec.cl; Contributing authors: chenweiw@umich.edu; sakethsundar@college.harvard.edu; renatadallelucca@gmail.com; fbeloy@montefioreslc.org; reezemedina@gmail.com; mmcnich3@bilh.org; kveni@mit.edu; Nkurtzm@emory.edu; felipemirshawka@gmail.com; marcela.aguirre@falp.org; ebner.daniel@mayo.edu; lceli@mit.edu;

Abstract

The Clever Hans effect occurs when machine learning models rely on spurious correlations instead of clinically relevant features and poses significant challenges to the development of reliable artificial intelligence (AI) systems in medical imaging. This scoping review provides an overview of methods for identifying and addressing the Clever Hans effect in medical imaging AI algorithms. A total of 173 papers published between 2010 and 2024 were reviewed, and 37 articles were selected for detailed analysis, with classification into two categories: detection and mitigation approaches. Detection methods include model-centric, data-centric, and uncertainty and bias-based approaches, while mitigation strategies encompass data manipulation techniques, feature disentanglement and suppression, and domain knowledge-driven approaches. Despite the progress in detecting and mitigating the Clever Hans effect, the majority of current machine learning studies in medical imaging do not report or test for shortcut learning, highlighting the need for more rigorous validation and transparency in AI research. Future research should focus on creating standardized benchmarks, developing automated detection tools, and exploring the integration of detection and mitigation strategies to comprehensively address shortcut learning. Establishing community-driven best practices and leveraging interdisciplinary collaboration will be crucial for ensuring more reliable, generalizable, and equitable AI systems in healthcare.

Keywords: Clever Hans, Shortcut Features, Shortcut Learning, Medical Imaging, Machine Learning, Review

1 Statements and Declarations

1.1 Funding

The author Leo A. Celi, received research support in the Laboratory for Computational Physiology (Massachusetts Institute of Technology), funded by the National Institute of Health through R01 EB017205, DS-I Africa U54 TW012043-01 and Bridge2AI OT2OD032701, and the National Science Foundation through ITEST #2148451.

1.2 Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

1.3 Author Contributions

All authors contributed to the study conceptualization and methodology design. Megan McNichol performed data collection. Data Analysis was conducted by Constanza Vásquez-Venegas, Chenwei Wu, Saketh Sundar, Renata Prôa, Francis Joshua Beloy, Jillian Reeze Medina, Krishnaveni Parvataneni, Nicholas Kurtzman, Felipe Mirshawka, Marcela Aguirre-Jerez, and Daniel K. Ebner. The original draft was prepared by Constanza Vásquez-Venegas, Chenwei Wu, Saketh Sundar, Renata Prôa, Francis Joshua Beloy, and Daniel K. Ebner. All authors reviewed the original draft, with editing performed by Constanza Vásquez-Venegas. The project was supervised by Leo A. Celi, and managed by Constanza Vásquez-Venegas. Constanza Vásquez-Venegas also created the included figure.

1.4 Ethics approval

This study is a literature review and does not involve human or animal participants, nor does it include any original data collection. Therefore, ethics committee approval was not required.

1.5 Consent to participate

As this study is a literature review and does not involve the collection of new data from human or animal subjects, obtaining consent to participate was not applicable.

1.6 Consent to publish

This study is a literature review and does not involve individual data or images from participants. Therefore, obtaining consent to publish is not applicable.

Detecting and Mitigating the Clever Hans Effect in Medical Imaging: A Scoping Review

Abstract

The Clever Hans effect occurs when machine learning models rely on spurious correlations instead of clinically relevant features and poses significant challenges to the development of reliable artificial intelligence (AI) systems in medical imaging. This scoping review provides an overview of methods for identifying and addressing the Clever Hans effect in medical imaging AI algorithms. A total of 173 papers published between 2010 and 2024 were reviewed, and 37 articles were selected for detailed analysis, with classification into two categories: detection and mitigation approaches. Detection methods include model-centric, data-centric, and uncertainty and bias-based approaches, while mitigation strategies encompass data manipulation techniques, feature disentanglement and suppression, and domain knowledge-driven approaches. Despite the progress in detecting and mitigating the Clever Hans effect, the majority of current machine learning studies in medical imaging do not report or test for shortcut learning, highlighting the need for more rigorous validation and transparency in AI research. Future research should focus on creating standardized benchmarks, developing automated detection tools, and exploring the integration of detection and mitigation strategies to comprehensively address shortcut learning. Establishing community-driven best practices and leveraging interdisciplinary collaboration will be crucial for ensuring more reliable, generalizable, and equitable AI systems in healthcare.

Keywords: Clever Hans, Shortcut Features, Shortcut Learning, Medical Imaging, Machine Learning, Review

1 Introduction

Artificial Intelligence (AI) has shown remarkable potential in medical imaging, promising to enhance diagnostic accuracy and efficiency (Ting et al. (2018); Huang et al. (2022)). However, ensuring the reliability of these AI systems is crucial given the high stakes in healthcare decisions(Balagurunathan et al. (2021); Dvijotham et al. (2023)). One significant challenge is the Clever Hans Effect (Samhita and Gross (2013)), also referred to as shortcut learning (Kauffmann et al. (2020)).

The Clever Hans effect is named after a famous horse, Hans, from the early 20th century, who was claimed to perform arithmetic and other intellectual tasks. Later investigation revealed that "Clever Hans" was actually responding to involuntary cues in the body language of his human trainer, who was unaware of providing such cues. In AI, the Clever Hans effect occurs when models appear to excel at certain tasks but are actually relying on unintended or irrelevant features in the data rather than learning the underlying logic they were intended to capture (Lapuschkin et al. (2019)). This phenomenon is of particular concern in the field of medical imaging, where algorithms must not only be accurate but also adhere to clinical logic and be interpretable to prevent misdiagnosis, inappropriate treatment recommendations, or inequitable outcomes. For instance, an AI system trained to detect pneumonia in chest X-rays (CXR) might inadvertently learn to recognize chest tubes and portable radiographic markers rather than actual signs of the disease (Banerjee et al. (2023), Bellamy et al. (2022)). Understanding how the Clever Hans effect manifests in clinical data is thus crucial for ensuring the reliability and effectiveness of AI applications in healthcare. It raises important questions about the trustworthiness of AI decisions (Hatherley (2020); Zhang and Zhang (2023)), the potential for biases in healthcare delivery (Gichoya et al. (2023)), and the need for rigorous validation of AI systems before clinical deployment (Esmaeilzadeh (2024)).

This study conducts a scoping review of the existing literature so as to provide a comprehensive overview of the Clever Hans effect with regard to AI algorithms for medical images. This review focuses on identifying methods for detecting shortcut learning in medical imaging AI, exploring strategies for mitigating this effect, and examining how the reliability and equity of AI-driven medical image analysis may be enhanced.

2 Review

2.1 Review protocol

This scoping review focuses on the literature available for the Clever Hans effect on AI algorithms applied to medical images and uses the PRISMA Extension for Scoping Reviews PRISMA-ScR Checklist (Moher et al. (2010)) as the review protocol. The automatic search strategy was performed in the following databases: PubMed, Embase, Web of Science, Academic Search Premier, and APA Psycinfo. The search strategy consisted of common terms related to the Clever Hans effect along with terms related to AI and imaging modalities and included years 2010 to 2024. There was no language limit. The exact search strategy can be found in Online Resource 1. All articles from peer-reviewed journals that aimed to detect or mitigate the Clever Hans effect on AI algorithms applied to medical images were included. Articles were excluded if they only mentioned the Clever Hans effect without addressing it, applied the terms in a different context (such as non-medical imaging or unrelated AI applications), or were review articles. The search strategy was initially performed in October 2023. The search was repeated twice, in May 2024 and October 2024, to include more recent articles. The de-duplication of citations was performed first in EndNote and then again later in Covidence.

Gray literature, consisting of peer-reviewed articles manually searched in Google Scholar and known to the authors but not retrieved through traditional database searches, was included to ensure the review captures all significant contributions. These additional sources, although peer-reviewed, are often overlooked by standard search strategies but are vital for presenting a more comprehensive and nuanced understanding of the Clever Hans effect in medical imaging AI. Gray literature was included up to October 2024, after the initial database search was completed.

2.2 Study Selection

The screening and eligibility analysis of the retrieved articles was performed by the authors using the software Covidence. Article assessment was performed in two steps: (1) screening by title and abstract, and (2) eligibility assessment by full-text review. In each step of the PRISMA-ScR methodology, articles are reviewed twice, with a third review in case of disagreement. The included articles were analyzed by two teams of medical AI practitioners and clinicians, including all authors, so as to determine whether the goal of the paper was to detect the Clever Hans effect or to mitigate it.

2.3 Results

2.3.1 Search Strategy

The search strategy (Figure 1) yielded 207 articles. Further, 5 articles were included as "gray literature," which refers to articles on the topic that were not originally retrieved by the search strategy. From the initially identified records 39 were duplicates. The remaining 173 articles were screened based on title and abstract evaluation, resulting in 105 exclusions. The articles were then analyzed for full text, leading to exclusion of 24 articles, and a total of 37 articles thus included for the final analysis. These articles were classified as detection, mitigation, or both, according to their proposed approaches, resulting in 13, 19, and 4 articles, respectively, as summarized in Table 1. The single remaining article was classified as "definition." The 17 medical imaging modalities covered by the included articles are shown in Table 2.

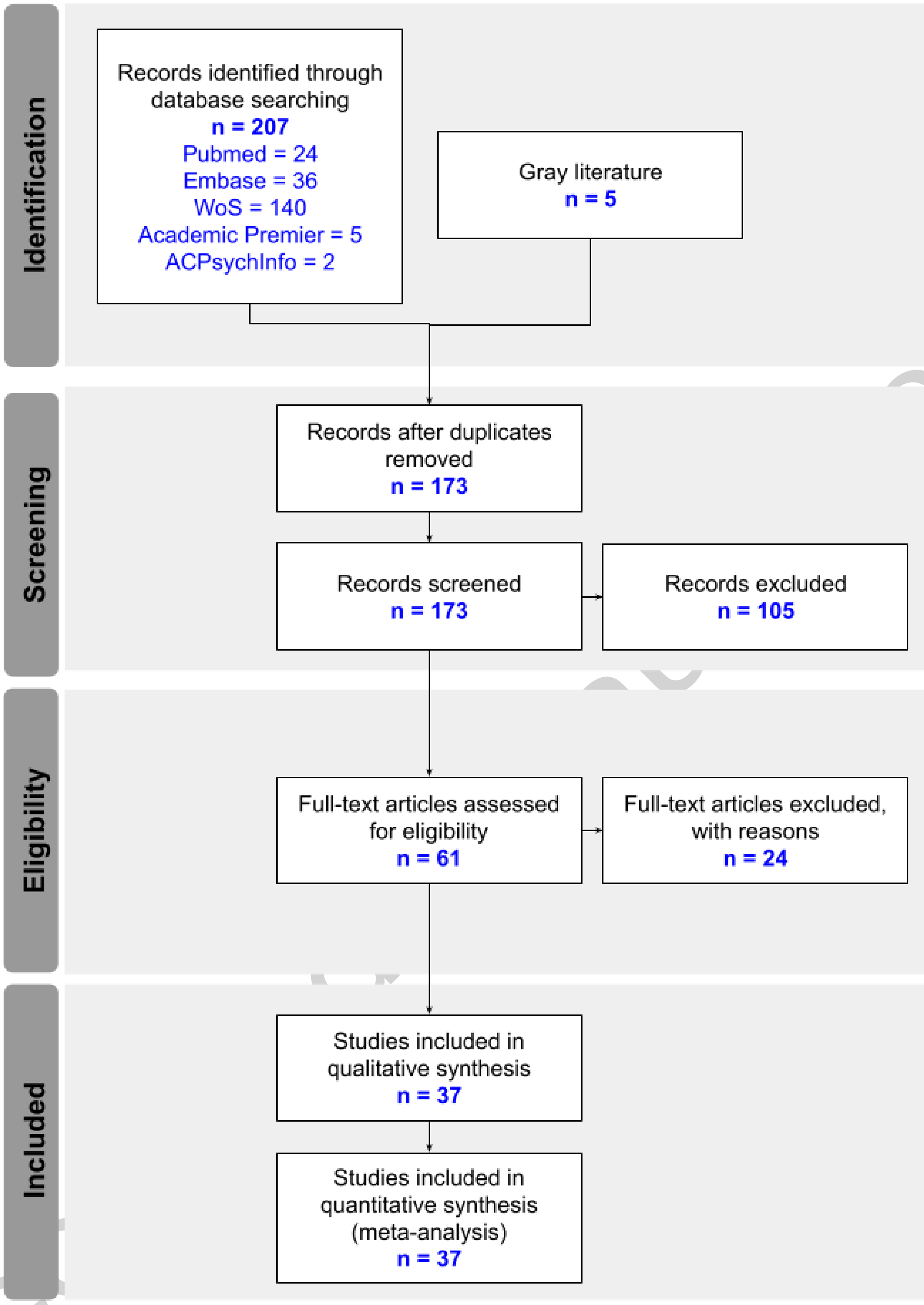

Fig. 1: Flow diagram for article inclusion and exclusion based on PRISMA methodology

4

Table 1: Detection and Mitigation Methods for Clever Hans Effect in Medical AI

Taxonomy	**Methods**	**Articles**
Detection Methods:		
Model-Centric Approaches	Performance Replication and Feature Generalization	DeGrave et al. (2021)
	Identifying Confounding Factors	Qu et al. (2024), Compton et al. (2023)
	Model Interpretation Techniques	Chen and Lin (2024), Sun et al. (2023), Wainwright et al. (2023), DeGrave et al. (2021)
Data-Centric Approaches	Dataset Bias Abduction	DeGrave et al. (2021)
	Attribution Maps and Shortcut Detection	Pedrosa et al. (2022), Arias-Londoño and Godino-Llorente (2024), Zhang et al. (2023)
	Occlusion Tests	Haynes et al. (2024), Wallis and Buvat (2022),Mahmood et al. (2021)
Uncertainties and Biases-Based Methods	Counterfactual Explanations	Thiagarajan et al. (2022)
	Fairness as Proxies	Brown et al. (2023), Souza et al. (2023), Bottani et al. (2023)
	Data Acquisition Bias Quantification	Ong Ly et al. (2024)
Mitigation Methods:		
Data Manipulation Techniques	Data Balancing and Preprocessing	Compton et al. (2023), Brown et al. (2023), Ben Ahmed et al. (2022), Luo et al. (2022), Aliverti et al. (2021)
	Data Augmentation	Wang et al. (2024), Ouyang et al. (2022), Chen et al. (2022), Zare and Nguyen (2022), Huang et al. (2023), Chen et al. (2022)
Feature Disentanglement and Suppression	Feature Decoupling and Regularization	Trivedi et al. (2022), Pohjonen et al. (2022), Zhao et al. (2020)
	Shortcut Removal via Attention and Heatmaps	Müller et al. (2023), Bassi et al. (2024), Bottani et al. (2023)
	Causal Intervention and Information Minimization	Qu et al. (2024), Fay et al. (2023), Luo et al. (2022)
Domain Knowledge-Driven Approaches	Semantic Labeling and Data Representation	Xiong et al. (2022)
	Domain-Specific Preprocessing	Wu et al. (2023), Arias-Londoño and Godino-Llorente (2024)
	Improving Model Robustness	Jabbour et al. (2020), Hooper et al. (2024)

Table 2: Medical Imaging Modalities of Included Clever Hans Articles

Modality	Papers
Chest X-ray	Bellamy et al. (2022), Pedrosa et al. (2022), Thiagarajan et al. (2022), Zhang et al. (2023), Arias-Londoño and Godino-Llorente (2024), Ong Ly et al. (2024), Chen and Lin (2024), Haynes et al. (2024), Sun et al. (2023), Wainwright et al. (2023), DeGrave et al. (2021), Ben Ahmed et al. (2022), Brown et al. (2023), Luo et al. (2022), Luo et al. (2022), Trivedi et al. (2022), Zare and Nguyen (2022), Bassi et al. (2024), Wang et al. (2024), Pohjonen et al. (2022), Ong Ly et al. (2024), DeGrave et al. (2021), Compton et al. (2023), Hooper et al. (2024), Jabbour et al. (2020)
Brain MRI	Bottani et al. (2023), Souza et al. (2023), Wallis and Buvat (2022), Souza et al. (2024), Chen et al. (2022), Fay et al. (2023)
Lung CT	Mahmood et al. (2021), Xiong et al. (2022), Ong Ly et al. (2024)
ECG, lung CT, Auscultation, EHR	Ong Ly et al. (2024)
Prostate Pathology slides	DeGrave et al. (2021)
Brain fMRI	Aliverti et al. (2021)
Abdominal CT-MRI, cardiac MRI, prostate MRI	Ouyang et al. (2022)
Brain MRI, hand X-Ray	Zhao et al. (2020)
Pancreatic CT	Qu et al. (2024)
Ultrasound Images	Huang et al. (2023)
Retinal Fundus Photos	Wu et al. (2023).

2.3.2 Methodologies for Detecting the Clever Hans Effect

Medical imaging AI algorithms face significant challenges related to the Clever Hans Effect, where models exploit shortcuts or non-clinically relevant features in datasets, leading to inflated performance metrics. Several methods have emerged to address this issue from different perspectives, including model-centric and data-centric approaches, as well as uncertainties and biases-based methods.

I. Model-Centric

Performance Replication and Feature Generalization

Model-centric approaches assess how machine learning models may depend on spurious features by testing them on internal datasets and validating their performance on external datasets. When models perform poorly on external data, it suggests that they may be relying on source-specific features rather than clinically relevant patterns. DeGrave et al. (2021) demonstrated that models often fail to generalize due to their reliance on features specific to the training data, particularly when the data is sourced from different origins. They highlighted this issue in the context of COVID-19 CXR classification, where multiple public datasets, often compiled without standardized protocols, exacerbated the model's dependence on spurious, source-specific features.

Identifying Confounding Factors

Identifying confounding factors is crucial for understanding how models may learn spurious correlations. Structural causal models (SCMs, Pearl et al. (2000)) represent the causal relationships between variables, allowing researchers to disentangle confounders and assess their impact on model predictions. By explicitly modeling these confounders, such as intensity variations and texture differences, SCMs can help clarify how models may rely on non-clinically relevant features. Qu et al. (2024) applied SCMs to pancreatic cancer diagnosis models on CT scans, identifying key confounders: intensity variations from contrast agent metabolism and scanning times, as well as texture differences caused by individual non-cancerous factors. Their work demonstrated these confounders can lead models to learn spurious correlations. Similarly, Compton et al. (2023) identified hospital-specific artifacts as confounding factors in CXR. By analyzing worst-group accuracy and probing model embeddings, they revealed that models often rely on non-disease-related signals, such as hospital origin, to make predictions, thereby reducing generalizability and leading to shortcut learning. Both studies underscore the importance of identifying and addressing confounding factors to improve model robustness and avoid reliance on spurious correlations.

Model Interpretation Techniques

Model interpretation techniques, such as Guided Backpropagation (Springenberg et al. (2014)), Gradient-weighted Class Activation Mapping (Grad-CAM, Selvaraju et al. (2017)), and SHapley Additive exPlanations (SHAP, Lundberg and Lee (2017)), are valuable tools for detecting the Clever Hans effect in medical imaging classifiers. By analyzing the model's sensitivity to artificial confounders and comparing the expected outcomes with the model's predictions, researchers can identify the model's reliance on spurious correlations. Furthermore, methods like attention visualization, uncertainty estimation, and low-dimensional data representations aid in detecting and visualizing data quality issues, enhancing the assessment of trustworthiness and reliability in deep learning models. Sun et al. (2023) and Chen and Lin (2024) used model interpretation techniques to detect shortcut learning in classifiers trained on CXR datasets with artificially introduced confounders, such as hospital tags and occlusions. They introduced two metrics: Confounder Sensitivity (CS), which measures confounder attribution accuracy, and Normalized Cross Correlation (NCC), which assesses the similarity of model output to predicted changes. Their findings suggest that models increasingly relying on spurious correlations should exhibit higher CS and lower NCC values, emphasizing the need for robust model evaluation methods. Similarly, Wainwright et al. (2023) employed attention visualization, uncertainty measurements, and low-dimensional data representations to detect and visualize data quality issues in CXR disease classification, contributing to the broader analysis of model trust and reliability. DeGrave et al. (2021) further illustrated how deep learning models for COVID-19 detection in CXR can exploit spurious features using saliency maps based on expected gradients. These maps highlighted not only the lung fields but also irrelevant features, such as laterality markers and image edges, that influenced predictions. This reliance on dataset-specific artifacts, like differences in radiographic projection or annotations, contributed to poor generalization on external datasets. Finally, they emphasized the importance of explainable AI methods like saliency maps for identifying such confounds, recommending population-level audits to ensure models are not driven by irrelevant patterns.

II. Data-Centric

Dataset Bias Abduction

Data-centric approaches to detecting the Clever Hans effect focus on uncovering biases within the training data. Generative adversarial networks (GANs, Goodfellow et al. (2020)) can be employed to transform images and abduct dataset-specific biases. By simulating positive and negative cases with varying source-specific features, researchers can identify these biases and highlight the importance of collecting data that accurately represents the target population. DeGrave et al. (2021) employed this approach, emphasizing the need for target population-based data collection to minimize the influence of dataset-specific features. Using GANs to simulate positive and negative CXR COVID-19 cases, they discovered that the network added or removed laterality markers and annotations, thereby reinforcing the presence of source-specific features in the training data.

Attribution Maps and Shortcut Detection

Shortcut detective models can be trained to identify intrinsic shortcuts that may affect the model's generalizability. Pedrosa et al. (2022) developed an automatic label detection framework using YOLOv3 to obscure written labels, markers, and equipment identifiers in CXR images, which can introduce spurious correlations in medical image classification tasks. The model successfully identified labels of varying font sizes, brightness, and positions, demonstrating its robustness across multiple datasets. Their approach not only promoted the learning of generic, clinically relevant features by obscuring non-clinical artifacts but also provided a method for retrospectively analyzing trained models to infer whether shortcuts related to image markers were learned.

On the other hand, attribution maps generated by techniques like Layer-wise Relevance Propagation (LRP, Bassi and Attux (2021)) and Deep Learning Important Features (DeepLIFT, Shrikumar et al. (2017)) can highlight the relevance of input features to a model's predictions, helping researchers identify reliance on non-clinically relevant features. Arias-Londoño and Godino-Llorente (2024) applied LRP and DeepLIFT to CXR to elucidate the impact of input features on COVID-19 detection. Zhang et al. (2023) built upon this approach by developing shortcut detective models to uncover intrinsic shortcuts, such as contrast and sharpness. Their findings revealed that these shortcuts were present in the COVIDx dataset. In contrast, datasets curated by medical professionals showed no evidence of such shortcuts. This comparison highlights how intrinsic biases can arise from poorly curated datasets, ultimately impacting model performance.

Occlusion Tests

Occlusion tests are effective methods for detecting shortcut learning in medical imaging classifiers. Occluding clinically relevant regions, such as the lungs in CXR, allows researchers to assess whether a model retains its discriminative power. If a model continues to perform well despite these occlusions, it suggests reliance on irrelevant, non-pathological features, indicating shortcut learning. Additionally, training models to predict attributes like sex, originating site, or scanner type using the feature space of a pre-trained classifier can reveal the inadvertent encoding of site-related information, which may serve as shortcuts for disease classification.

Haynes et al. (2024) introduced a novel occlusion test to detect shortcut learning in COVID-19/non-COVID classification model. By occluding the lung region with a black box, they demonstrated that models maintaining discriminative power despite the occlusion were likely exploiting irrelevant, non-pathological features. Similarly, Wallis and Buvat (2022) performed brain tumor classification using a widely-used publicly available T1-weighted contrast-enhanced MRI dataset, revealing that models could achieve high classification accuracy even when the tumor was obscured or when only the skull was visible. This suggested that the models were leveraging anatomical location cues or other irrelevant features rather than non-tumor-specific information. In this case, the dataset (Cheng (2017)) was biased from its construction, as the representative slices were extracted by a human expert. While this bias did not affect the dataset's original purpose—radiomics and tumor segmentation—it became problematic to the classification task, ultimately impacting the model's clinical performance. In a similar context, Mahmood et al. (2021) provide a compelling case study demonstrating the use of sanity tests in a deep learning system designed to classify pancreatic cancer from CT scans. They propose a series of tests, including evaluating models trained on target-present and target-absent datasets, as well as noise images and background patches. In their case, the removal of the pancreas from scans revealed that the model could still perform well, exposing its reliance on confounding factors unrelated to the disease itself. Similarly, training with noise images and testing on unseen regions of interest further demonstrated how these models can learn shortcuts rather than focusing on medically relevant features. These findings underscore the importance of testing for such biases early in the development phase, as these flaws could lead to significant performance drops when systems are applied to generalization datasets.

III. Uncertainties and Biases

Counterfactual Explanations

Counterfactual explanations, which help reveal how small changes in input data affect a model's predictions, have become a key tool in detecting biases and unintended shortcuts in AI models, especially in medical imaging. Thiagarajan et al. (2022) introduce TraCE (Training Calibration-based Explainers), a method that uses uncertainty-based calibration to generate meaningful counterfactuals for clinical image predictors. By leveraging an auto-encoding network and the Learn-by-Calibrating technique, TraCE enables progressive transitions between normal and abnormal states (e.g., pneumonia severity in chest X-rays) while avoiding reliance on spurious features such as background pixels. In their studies, TraCE effectively detected biases, such as models incorrectly focusing on synthetic features, and explored relationships between attributes like age and diagnosis. Overall, TraCE provides a powerful framework for improving the interpretability and trustworthiness of AI models in healthcare.

Fairness as Proxies

Fairness-based approaches can effectively detect shortcuts in machine learning models by using demographic information to uncover hidden biases and unintended correlations. Brown et al. (2023) introduced ShorT (Shortcut Testing), a method for detecting and mitigating Clever Hans-like behavior in clinical AI models by assessing how much sensitive attribute information, such as age or race, is encoded in a model's features. Rather than focusing on the mere presence of sensitive attribute encoding, ShorT tests whether this encoding leads to unfair performance by varying the strength of the encoding through a demographic prediction head.

Additionally, Souza et al. (2023) conducted a comprehensive investigation into whether site-related effects and other confounding variables, such as sex and scanner type, are encoded within the feature space of a deep learning model for classifying Parkinson's disease (PD) based on T1-weighted brain MRI. Using a diverse dataset of 1880 subjects across 41 different sites, they analyzed how the model incorporates these variables, even when not explicitly provided during training. They trained new models to predict sex, originating site, and scanner type based on using the feature space of a pre-trained PD classifier, achieving high levels of accuracies (75%, 71%, and 79%, respectively). These findings suggest that the PD classifier inadvertently encoded spurious correlations, raising concerns about the potential use of these variables as shortcuts in disease classification, which may compromise the model's clinical utility.

On the other hand, Bottani et al. (2023) investigated fairness concerns in machine learning models for detecting dementia from T1-weighted brain MRI scans, emphasizing the need to address shortcuts while preserving predictive accuracy across demographic groups. Their study identified significant biases caused by training set imbalances, particularly due to gadolinium contrast agent injection and varying image quality, which inflated accuracy by over 17 percentage points in models exhibiting the Clever Hans effect. Through their approach, they revealed that classifiers trained on clinical data performed poorly, particularly when trained on biased datasets, underscoring the challenges of generalizing models from research to clinical settings.

Data Acquisition Bias Quantification

Quantifying the impact of data acquisition biases (DAB) in training data is essential for accurately estimating AI model performance on unseen external data. Ong Ly et al. (2024) highlight the significance of quantifying DAB to accurately assess AI model performance on unseen external data, revealing that traditional evaluation methods can overestimate accuracy by up to 20% due to DAB-induced shortcut learning. Their novel DAB-induced shortcut learning metric enables more precise predictions of external dataset performance, demonstrating an average improvement of 16% over existing methods. The study underscores the variability in data collection across different hospitals and acquisition settings, which can lead AI models to rely on spurious features instead of clinically relevant signals. While focusing on DAB, the research also acknowledges the importance of addressing demographic biases and encourages further exploration into additional unstructured data modalities. Overall, the proposed methodology aims to enhance the reliability of AI model evaluations in healthcare, facilitating safer integration into clinical workflows.

These methodologies offer valuable insights into detecting the Clever Hans effect in AI algorithms used in medical imaging. Ranging from model-centric approaches to data-centric techniques for handling uncertainties and biases, each study contributes to a deeper understanding of the challenges and opportunities in developing more robust and equitable AI systems for medical image analysis. Table 3 summarizes the advantages and weaknesses of each detection approach, providing a clear comparison of their effectiveness. By proactively addressing biases early in the development phase and ensuring models rely on clinically relevant features, these methodologies lay the groundwork for enhancing diagnostic accuracy and improving patient care in clinical settings.

Table 3: Advantages and weaknesses of methodologies for detecting the Clever Hans effect in medical imaging AI models.

Taxonomy	**Methods**	**Advantages**	**Weaknesses**
Model-Centric Approaches	Performance Replication and Feature Generalization	Enables the identification of spurious correlations by evaluating model performance across internal and external datasets. It facilitates model generalization and robustness across diverse data sources.	May fail to detect subtle biases if external datasets do not adequately represent clinical diversity. Requires access to multiple, well-curated datasets.
	Identifying Confounding Factors	Structural Causal Models (SCMs) allow the disentanglement of confounding factors from clinically relevant features, thus revealing the reliance on non-medical cues.	Requires detailed knowledge of confounders, making implementation complex. Its effectiveness depends on the quality of the causal model design.
	Model Interpretation Techniques	Provides transparency by visualizing the model's decision-making process, helping to identify reliance on non-clinically relevant correlations.	These techniques can be computationally intensive, and their applicability varies depending on model architecture. Interpretation can be subjective.
Data-Centric Approaches	Dataset Bias Abduction	Highlights biases embedded within the dataset, emphasizing the need for collecting data that represents the target clinical population.	Bias detection can be incomplete if datasets lack diversity; mitigating biases introduced by augmentation may not fully eliminate spurious correlations.
	Attribution Maps and Shortcut Detection	Uses techniques like LRP and DeepLIFT to visually highlight the importance of input features, allowing for the detection of spurious shortcuts.	Interpretation of the visual maps requires domain expertise, and the method can be sensitive to noise or irrelevant features in the dataset.
	Occlusion Tests	Tests the model's reliance on irrelevant features by applying occlusion to critical regions of input data or augmenting datasets to disrupt correlations. Helps validate model reliance on meaningful patterns.	Augmentation may introduce artifacts, and occlusion may not reveal more complex shortcuts. There is also a risk of generating unrealistic data distributions.
Uncertainties and Biases-Based Methods	Counterfactual Explanations	Offers insights into the sensitivity of model predictions to small changes in input data, effectively detecting reliance on unintended features. Promotes model interpretability.	Generating counterfactuals can be computationally expensive and complex, and their results may not always be intuitive for non-experts.
	Fairness as Proxies	Leverages demographic attributes (e.g., race, age, sex) to uncover hidden biases and spurious correlations in model predictions, contributing to fairness evaluations in clinical AI.	Requires access to sensitive demographic information, which might raise privacy concerns. Balancing fairness with model accuracy can be challenging.
	Data Acquisition Bias Quantification	Provides a means to quantify biases introduced during data acquisition, enhancing understanding of how such biases impact model performance and generalization.	Complex to implement due to the need for detailed knowledge about data acquisition protocols. While it quantifies bias, it does not directly address it.

2.3.3 Methodologies for Mitigating the Clever Hans Effect

This section discusses several strategies for addressing the Clever Hans effect in medical imaging AI, focusing on three main categories: data manipulation techniques, feature disentanglement and suppression, and domain knowledge-driven approaches.

I. Data Manipulation

Data Balancing and Preprocessing

Data balancing and preprocessing techniques aim to reduce the impact of spurious correlations and biases. In deep learning, these methods help prevent models from learning shortcuts by ensuring the training data is more representative of the target population and less influenced by confounding factors. Compton et al. (2023) emphasize data balancing as a key mitigation strategy in their large-scale study on CXR. They demonstrated that balancing datasets to equalize disease prevalence across different hospitals can reduce spurious correlations, such as those between hospital-specific artifacts and disease labels. While this approach often improves worst-group accuracy, the authors caution that balancing alone is not always sufficient, as it may still lead to performance degradation in some cases. Their work highlights the importance of combining data balancing with other preprocessing strategies to effectively mitigate reliance on shortcuts.

Other techniques include applying visual inspection and curation, and using fine-grained labels to encourage models to focus on clinically relevant features. Brown et al. (2023) proposed a framework for mitigating shortcut learning, focusing on the use of gradient reversal to reduce correlations between sensitive attributes like age and race and model outcomes. Their approach allows creating models with varying levels of mitigation to explore the trade-offs between fairness and performance. While they find that methods like dataset balancing and gradient reversal can improve fairness without always harming performance, they also acknowledge that some mitigation strategies may degrade model accuracy. Their framework thus offers a flexible approach to detecting and mitigating shortcuts, ensuring fairer clinical AI systems.

Similarly, Ben Ahmed et al. (2022) proposed a pipeline to mitigate shortcut learning in CNN-based disease diagnosis models by reducing the performance gap between internal and external tests. improve model generalization on CXR from different sources. Their approach involves data preprocessing techniques like lung segmentation and cropping, histogram equalization, and noise reduction to improve model generalization. They demonstrated that these methods significantly reduce the performance drop when testing on unseen data.

The Clever Hans effect can also arise from insufficiently specific data labels, leading models to explore irrelevant features. Luo et al. (2022) investigated the impact of fine-grained lesion annotations on deep learning models for disease classification and detection in CXR. By comparing two models—CheXNet, trained with radiograph-level labels, and CheXDet, trained with lesion-level labels—they found that fine-grained annotations, such as bounding boxes, helped mitigate shortcut learning by enabling the model to focus on the correct pathological patterns. Additionally, Aliverti et al. (2021) addressed the challenge of spurious correlations in high-dimensional data, such as batch effects and race, by proposing a method to preprocess data using a constrained form of matrix decomposition. This approach ensures statistical independence from nuisance variables while minimizing information loss. Their method is agnostic to model type and aims to improve its generalizability.

Data Augmentation

Data augmentation is a widely used technique in deep learning to improve model generalization and robustness. In mitigating shortcut learning, data augmentation introduces variations in the training data to disrupt spurious correlations between input features and target labels. Techniques such as image transformations, noise addition, and cross-domain simulation expose models to a more diverse set of examples, reducing their reliance on shortcuts and improving generalization to unseen data.

Wang et al. (2024) demonstrated that training a model using augmented CXR images reduced error rate disparities among racial groups (-5.45%), age groups (-13.94%), and sex (-22.22%). For Alzheimer's disease detection, the model trained with augmented MRI images showed reductions of 53.11% and 31.01% in error rate disparities among age and sex groups, respectively. Augmentation diminished the model's reliance on demographic attributes, leading to improved clinical performance.

Ouyang et al. (2022) introduced a data augmentation technique for cross-domain medical image segmentation that simulates data in a shifted domain and effectively removes spurious correlations through causal intervention. Zare and Nguyen (2022) applied Invariant Risk Minimization (IRM, Arjovsky et al. (2019)) to augment CXR data, a method designed to learn invariant features by using predefined environments based on confounders like age and sex. These approaches help reduce correlations between sensitive attributes and outcomes, thereby mitigating the effects of shortcut features on model training. Nevertheless, completely eliminating spurious correlations is challenging due to inherent data biases.

In ultrasound imaging, Huang et al. (2023) proposed Combining Foreground And Background (CFAB), an online data augmentation technique that helps convolutional neural networks (CNNs) focus on key features without requiring foreground annotations. CFAB disrupts correlations between backgrounds and labels by changing different backgrounds for a single sample, thus reducing spurious correlations in medical image classification. For multi-site neuroimaging studies, Chen et al. (2022) created Correcting Covariance Batch Effects (CovBat), a method that addresses site effects across mean, variance, and covariance. This method first corrects mean and variance shifts, then uses principle component analysis to align within-site covariance matrices to the pooled covariance structure, which improves data harmonization across multiple sites.

II. Feature Disentanglement and Suppression

Feature Decoupling and Regularization

Feature decoupling and regularization techniques in deep learning seek to differentiate clinically relevant features from spurious ones. Methods such as spectral decoupling and invariant feature learning encourage models to concentrate on features directly related to the target task and suppress the impact of confounding factors. These approaches help reduce shortcut learning by promoting the development of more robust and generalizable representations.

In COVID-19 studies, shortcut learning often stemmed from poor-quality CXR datasets. Introducing feature disentanglement has been shown to improve generalization on unseen datasets. For instance, Trivedi et al. (2022) demonstrated that incorporating feature disentanglement into training helps models focus on target-related features while penalizing irrelevant ones. Similarly, Pohjonen et al. (2022) observed that spectral decoupling (Pezeshki et al. (2021)), which involves adding an L2 penalty to network logits, reduces overfitting on shortcut features. Although spectral decoupling effectively mitigated shortcut learning in prostate cancer histopathological images, it was less effective for COVID-19 CXR datasets, likely due to differences in acquisition protocols. Trivedi et al. (2022) also observed that preprocessing techniques like lung masking and histogram equalization, which are typically thought to enhance model generalization, were ineffective in preventing shortcut learning. Meanwhile, Zhao et al. (2020) introduced an end-to-end method for deriving features invariant to confounding factors while accounting for intrinsic correlations between confounders and predictions. This approach combined traditional statistical methods with recent invariant feature-learning techniques to ensure features reflect clinical outcomes rather than confounding variables.

Shortcut Removal with Attention and Heatmaps

Attention mechanisms in deep learning enhance model interpretability and performance by focusing on relevant features. To address shortcut learning, attention-based methods can be employed to identify and eliminate shortcuts in input data. By learning attention maps that highlight regions contributing to shortcut learning and replacing them with more pertinent information, models are guided to concentrate on clinically significant features, reducing reliance on spurious correlations. Müller et al. (2023) introduced the "Lens" network, which targets shortcut features in CXR for COVID-19 detection through an adversarial obfuscation mask learning approach. The framework comprises two U-Net modules: the attention module, which identifies shortcuts and produces a mask, and the replacement module, which generates suitable replacements for these shortcuts. Both modules are trained using the classifier's inverted gradients and an adversarial loss to balance the obfuscation of shortcut features with the utility of the modified images. Qualitative analysis of the attention maps indicated that the "Lens" approach effectively removed irrelevant information, such as text and icons, from the X-ray images.

Similarly, Bassi et al. (2024) introduced ISNet, a classifier architecture designed to mitigate shortcut learning caused by background bias in CXR. They demonstrated that optimizing LRP heatmaps could minimize the influence of spurious correlations between background features and image classes, highlighting how these techniques combined with architectures like ISNet, can prevent non-clinically relevant features from influencing model predictions. On the other hand, Bottani et al. (2023) addressed the Clever Hans effect by tackling biases in machine learning models used to detect dementia from brain MRI scans. Their study shows how non-causal factors, like gadolinium contrast agent injections and variations in image quality, can lead to spurious correlations, significantly inflating model accuracy. To mitigate these biases, they introduce an image translation technique to homogenize the appearance of images, reducing the model's reliance on non-causal features and improving fairness across demographic groups. This approach helps ensure more reliable and generalizable machine learning applications in clinical settings, emphasizing the importance of bias control in medical imaging.

Causal Intervention and Information Minimization

Causal intervention and information minimization techniques in deep learning target shortcut learning by explicitly modeling the causal relationships between input features, confounding factors, and target labels. Methods such as causal intervention, orthogonality constraints, and mutual information minimization encourage models to learn representations that are invariant to confounding factors, aligning more closely with the true causal mechanisms of the data. These approaches help mitigate shortcut learning by promoting the learning of more robust and generalizable features.

Qu et al. (2024) propose a confounding effects reduction scheme for pancreas imaging through a causal intervention strategy. By decoupling causal features (e.g., tumor characteristics) and non-causal confounders (e.g., intensity and texture variations), their approach reduces the influence of spurious correlations. They employ continuous image generations and feature diversification methods to simulate and randomize non-causal factors, forcing the model to focus on the correct, disease-related features. Orthogonality constraints further ensure independence between causal and non-causal factors, enhancing the model's generalization and clinical applicability. Fay et al. (2023) introduced the Mutual Information Minimization Model (MIMM) to mitigate spurious correlations. MIMM divides feature vectors into independent components for primary task prediction and spurious factor predictions, using a Mutual Information Neural Estimator (MINE) to reduce mutual information between these components. This approach facilitates causal, counterfactual invariant predictions, demonstrating robustness to data distribution shifts and adaptability to various tasks associated with brain MRI. Similarly, Luo et al. (2022) proposed a pseudo bias-balanced learning algorithm that initially trains a highly-biased model to generate pseudo-bias labels, which are then used to train a debiased model capable of avoiding shortcut features, thereby improving model generalization.

III. Domain Knowledge-Driven Approaches

Semantic Labeling and Data Representation

Providing detailed, well-defined categories based on expert knowledge encourages models to focus on meaningful patterns in the data rather than relying on shortcuts. Additionally, employing domain-specific data representations, such as multi-orientation slices and attention mechanisms, can further enhance the model's ability to capture clinically relevant features and minimize the impact of spurious correlations. For instance, Xiong et al. (2022) addressed pulmonary nodule classification by introducing five semantic classification labels and a multi-center, multi-label pipeline designed to extract robust features with true relational connections. Senior radiologists from two hospitals classified CT scans into five well-defined categories, and fifteen 2D slices from different orientations were extracted to augment the training set, improving generalization. They utilized a BotNet architecture, which combines CNNs with self-attention mechanisms, to enhance CT scan representations by focusing on both detailed and broader features. Additionally, they re-weighted the training samples using Random Fourier Features (RFF) to eliminate spurious correlations, particularly in out-of-distribution testing scenarios across multiple centers. This pipeline and its augmentation methods successfully improved predictive performance and mitigated overfitting in both validation and testing.

Domain-Specific Preprocessing

Domain-specific preprocessing techniques leverage expert knowledge to mitigate shortcut learning in deep learning models. By tailoring preprocessing steps to the unique characteristics of the data and the clinical task, such as vessel segmentation and obfuscation in retinal scans, models are encouraged to focus on clinically relevant features rather than relying on shortcuts. This approach helps ensure that the learned representations are more closely aligned with the underlying clinical phenomena. In ophthalmology, Wu et al. (2023) developed a vessel segmentation and obfuscation framework based on the understanding that vascular regions in the eye are closely correlated with gender information. Their approach effectively removed most gender-related information from retinal scans, thereby reducing gender disparities in diagnosis. Similarly, Arias-Londoño and Godino-Llorente (2024) applied semantic segmentation of lung regions in CXR images to guide deep learning models in detecting COVID-19 by focusing on areas with causal relationships to the disease. This technique minimized learning shortcuts and reduced the model's reliance on irrelevant artifacts or biases, such as external textures, ultimately leading to more interpretable and clinically useful outcomes.

Improving Model Robustness

Leveraging domain-specific knowledge is a key strategy for improving model robustness in medical image classification tasks. By incorporating prior knowledge from related diagnostic tasks, models can be guided to focus on clinically relevant features, reducing the risk of relying on spurious correlations or shortcuts. In their study on CXR, Jabbour et al. (2020) proposed a transfer learning method to address the issue of models exploiting spurious correlations like patient demographics or image preprocessing artifacts. By first training the model on a source task that lacks these biases (e.g., pneumonia diagnosis), the learned features act as domain priors, helping the model focus on clinically relevant patterns. The model's encoder is then frozen, and only the final layer is fine-tuned on the target task (e.g., congestive heart failure). This approach significantly reduces shortcut learning, improving the model's generalization to unseen patient populations. Another approach to improving model robustness is reframing traditional image classification tasks as segmentation tasks. In their work, Hooper et al. (2024) proposed that segmentation can mitigate the Clever Hans effect by encouraging models to focus on pixel-level, disease-specific features, rather than irrelevant background information. This shift ensures that models learn features directly related to the pathology, reducing the risk of shortcut learning and improving both interpretability and generalization. By summarizing segmentation masks into classification labels, this method enhances model reliability across datasets, even when spurious correlations are present.

Mitigating the Clever Hans effect in medical imaging AI requires a multi-faceted strategy. The methodologies discussed in this section, including data manipulation techniques, model architecture and training strategies, and domain knowledge-driven approaches, provide a comprehensive toolkit for developing more robust and generalizable models. Table 4 highlights the pros and cons of each mitigation strategy, offering a useful comparison to guide the selection of the most appropriate techniques. Further study and refinement of these methods is essential to ensure the development of models that can truly benefit patient care and outcomes.

Table 4: Advantages and weaknesses of methodologies for mitigating the Clever Hans effect in medical imaging AI models.

Taxonomy	**Methods**	**Advantages**	**Weaknesses**
Data Manipulation Techniques	Data Balancing and Preprocessing	Reduces the impact of spurious correlations by ensuring training data is more representative of the target clinical population. Ensures better generalization by balancing dataset characteristics.	While helpful, balancing may not fully eliminate biases, and preprocessing techniques like segmentation may lead to loss of valuable information.
	Data Augmentation	Introduces variability in the training data through transformations, noise addition, or synthetic data generation, reducing reliance on shortcut learning and enhancing generalization.	May unintentionally introduce artifacts or unrealistic patterns. Success largely depends on the quality and diversity of augmentation techniques.
Feature Disentanglement and Suppression	Feature Decoupling and Regularization	Improves robustness by isolating clinically relevant features and suppressing the influence of confounding or irrelevant features through regularization techniques like spectral decoupling.	Requires careful tuning to avoid over-regularization, which may degrade performance. Feature decoupling can be complex and computationally demanding.
	Shortcut Removal via Attention and Heatmaps	Uses attention-based methods to focus on clinically significant features, while eliminating irrelevant information from input data. Enhances interpretability by localizing attention to important regions.	Can overlook subtler shortcuts if not carefully designed. May incur a high computational cost, especially when employing advanced attention mechanisms.
	Causal Intervention and Information Minimization	Intervenes in causal relationships by separating confounders from clinical features, using approaches like mutual information minimization and orthogonality constraints to promote generalization.	Effective in reducing reliance on spurious correlations, but these techniques can be complex to implement and may require substantial domain knowledge.
Domain Knowledge-Driven Approaches	Semantic Labeling and Data Representation	Uses domain knowledge to label and represent data in clinically meaningful ways, helping models focus on relevant features and reducing the risk of learning shortcuts from irrelevant patterns.	Heavily reliant on expert annotations, which may be time-consuming and resource-intensive to generate. Generalization depends on the quality of the labeling.
	Domain-Specific Preprocessing	Tailors preprocessing steps to the specific clinical task, helping to remove non-clinical features like background noise or gender markers, ensuring models rely on relevant information.	Solutions tend to be highly task-specific, which can limit generalizability to other tasks or data types. Strong reliance on expert-designed pipelines.
	Improving Model Robustness	Incorporates domain-specific knowledge or redefines tasks (e.g., segmentation) to encourage the model to focus on clinically relevant, disease-specific features rather than spurious correlations.	May require reframing existing tasks and using transfer learning, which could increase the complexity of model training. Effectiveness depends on the availability of strong domain expertise.

19

3 Discussion

Addressing the Clever Hans effect is not just a matter of improving model performance; it is essential for developing AI systems that are both reliable and equitable in diverse healthcare settings. The reviewed articles fell into two categories: detection and mitigation approaches for the Clever Hans effect. We identified that many machine learning models, especially in CXR and brain MRI, still rely on non-clinical features, leading to spurious predictions. Key detection strategies (see Table 3) include model-centric, data-centric, and uncertainty and bias-based methods. While these approaches help uncover spurious correlations and hidden biases—model-centric methods by revealing feature reliance, data-centric methods by identifying dataset biases, and uncertainty-based methods by quantifying biases—they are limited by computational complexity, reliance on specific datasets, and challenges in standardizing their use across diverse clinical environments. For instance, the need for extensive computational resources and model interpretability tools may be a barrier in smaller healthcare institutions, which lack the infrastructure to support these techniques. Furthermore, the clinical translation of these detection methods remains limited due to the variability in data quality and patient demographics across healthcare settings, making it difficult to generalize findings.

Mitigation strategies (see Table 4) include data manipulation techniques, feature disentanglement and suppression, and domain knowledge-driven approaches. Data manipulation methods, such as data balancing and augmentation, reduce spurious correlations by making training data more representative and diverse, though they may introduce new biases or unrealistic patterns. Feature disentanglement approaches, like decoupling and attention-based shortcut removal, help focus models on clinically relevant features, but can be computationally expensive and may still overlook certain shortcuts. Domain knowledge-driven methods leverage expert input to guide models towards clinically meaningful features, yet they tend to be time-consuming and resource-intensive, requiring extensive domain expertise for effective implementation. In clinical settings, these challenges could hinder the widespread adoption of these strategies. For example, the need for extensive computational resources and domain expertise may limit their application in smaller or under-resourced healthcare institutions. Additionally, the requirement for large, diverse datasets for data manipulation methods could be difficult to meet, particularly in hospitals that serve homogeneous patient populations. Without addressing these practical barriers, these strategies may fail to fully translate into real-world clinical environments, where the ability to generalize AI models to different patient demographics and settings is crucial for equitable and reliable care.

As this is still an emerging area of research, these methods are not yet standardized or widely adopted, highlighting the urgent need for consistent benchmarks and robust validation techniques. Although significant progress has been made in identifying and mitigating shortcut learning, the reliance on labor-intensive manually designed tests, poses a significant challenge to scalability across diverse datasets or modalities. Future research should focus on developing standardized, automated tools that can be integrated into AI pipelines to ensure that models are systematically validated against shortcut learning. This will enable faster adoption of trustworthy AI systems in healthcare, with less risk of biased or misleading predictions.

Another critical aspect of overcoming the Clever Hans effect is fostering community collaboration, particularly between computer science and medical fields. The complexity of medical data and the prevalence of shortcut features make it essential to leverage interdisciplinary expertise. Collaboration between these fields can lead to more advanced detection and mitigation strategies that combine the strengths of both domains. This collaboration could also foster the creation of multi-center studies that test models on diverse datasets, increasing the robustness of AI systems in different medical settings and patient populations. In addition, lessons from other industries provide valuable insights into how to address the Clever Hans effect. For example, in finance, biases in AI systems used for credit scoring or loan approvals have prompted the development of stringent regulatory measures and explainability standards to ensure fairness and accountability (Shah and Sureja (2024)). Similarly, in Natural Language Processing (NLP), Large Language Models (LLMs) have been shown to exploit dataset biases as shortcuts, significantly compromising their generalization and robustness across diverse tasks. Research in NLP highlights that addressing these biases through adversarial training and data augmentation techniques can improve model performance on out-of-distribution samples (Dogra et al. (2024)). Medical imaging AI could benefit from adopting these techniques, particularly in developing models that can generalize across diverse clinical settings and patient populations.

A grassroots movement to share best practices for detecting and mitigating the Clever Hans effect could greatly accelerate progress in the field. Many researchers are independently developing effective strategies for their models, but these innovations often remain isolated within individual institutions or research groups. A concerted effort to share methods, tools, and insights through workshops, conferences, or online platforms, could benefit the entire community. Moreover, this movement could lead to the establishment of community-driven guidelines that promote transparency, reproducibility, and collaboration across research groups, helping to build AI systems that are both fair and clinically relevant. In addition to sharing best practices, the creation of a repository of known shortcut features for different imaging modalities and public datasets would be invaluable. Such a resource could include detailed descriptions of common spurious correlations found in modalities such as CXR, CT scans, MRI, and retinal images. This would serve as a reference point for researchers and clinicians, helping them identify potential shortcuts early in the development process. Over time, this repository could evolve into a living database, continually updated with new knowledge as models and data evolve. This initiative would not only aid in the early detection of shortcuts, but also encourage more rigorous model validation and testing across diverse datasets.

Finally, the detection and mitigation of the Clever Hans effect should be incorporated into AI education and training programs. As medical AI becomes more widespread, it is critical that future practitioners, researchers, and developers are trained to recognize the risks associated with shortcut learning. Educational curricula should emphasize the ethical implications of shortcut learning, especially in high-stakes fields like healthcare, where biased AI systems can have serious consequences for patient outcomes. By incorporating these topics into courses, workshops, and certifications, we can build a generation of AI professionals equipped to develop more reliable and equitable models.

Moving forward, addressing the Clever Hans effect in Medical Imaging AI requires a multifaceted approach that combines detection and mitigation strategies while promoting community collaboration and knowledge-sharing. Key priorities include:

- Standardized benchmarks and datasets: create diverse, real-world datasets that assess AI robustness and prevent reliance on shortcut features.
- Systematic detection methods: develop automated, standardized tools to consistently identify shortcut features across imaging modalities.
- Comprehensive strategies: combine model-centric and data-centric approaches to thoroughly detect and mitigate shortcut learning.
- Community-driven collaboration: foster interdisciplinary efforts across fields to advance detection and mitigation techniques.
- Repository of shortcut features: establish a shared database cataloging known shortcut features across imaging modalities.
- Transfer learning and domain adaptation: enhance model generalization by leveraging pre-trained models and adapting them to diverse clinical settings.
- Education and training: incorporate shortcut detection and mitigation into AI curricula to prepare future researchers and developers.
- Best practices and tools: promote the development of user-friendly platforms and formal guidelines for consistent shortcut detection and mitigation.

By prioritizing these initiatives, the medical AI community can make significant progress in developing AI systems that are not only robust and generalizable, but also equitable. This focus will ensure the safe and effective integration of AI technologies into diverse healthcare settings, ultimately improving patient outcomes and fostering trust in AI applications.

4 Summary

This scoping review examines the Clever Hans effect in medical imaging, where machine learning models rely on non-clinical, spurious features rather than medically relevant ones. The study analyzes 173 papers published between 2010 and 2024, with a focus on detection and mitigation strategies for identifying and addressing this phenomenon in AI systems. Detection methods are divided into model-centric, data-centric, and uncertainty and bias-based approaches, each with its strengths and limitations. Model-centric approaches test models on external datasets to expose shortcuts, while data-centric methods focus on biases within the training data. Uncertainty-based methods quantify hidden biases but face challenges in scalability and computational expense.

On the mitigation side, strategies include data manipulation, feature disentanglement, and domain knowledge-driven approaches. Data manipulation techniques like augmentation and balancing aim to reduce bias, though they may introduce new patterns. Feature disentanglement methods help models focus on relevant features, but

they can be resource-intensive. Domain knowledge-driven approaches integrate expert input to improve model training, though they are time-consuming. The review highlights the need for standardized benchmarks, automated tools, and interdisciplinary collaboration to ensure AI systems in healthcare are both reliable and equitable.

Looking ahead, future research should prioritize community-driven best practices, multi-center studies, and the creation of a repository cataloging known shortcuts in medical imaging datasets. By fostering collaboration between the AI and medical communities, the field can advance detection and mitigation techniques, ensuring more robust and generalizable AI applications across diverse healthcare settings.

References

Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization. arXiv preprint arXiv:1907.02893 (2019) https://doi.org/10.48550/arXiv.1907.02893

Arias-Londoño, J.D., Godino-Llorente, J.I.: Analysis of the clever hans effect in covid-19 detection using chest x-ray images and bayesian deep learning. Biomedical Signal Processing and Control **90**, 105831 (2024) https://doi.org/10.1016/j.bspc.2023.105831

Aliverti, E., Lum, K., Johndrow, J.E., Dunson, D.B.: Removing the Influence of Group Variables in High-Dimensional Predictive Modelling. Journal of the Royal Statistical Society Series A: Statistics in Society **184**(3), 791–811 (2021) https://doi.org/10.1111/rssa.12613

Bassi, P.R., Attux, R.: A deep convolutional neural network for covid-19 detection using chest x-rays. Research on Biomedical Engineering, 1–10 (2021) https://doi.org/10.1007/s42600-021-00132-9

Ben Ahmed, K., Hall, L.O., Goldgof, D.B., Fogarty, R.: Achieving multisite generalization for cnn-based disease diagnosis models by mitigating shortcut learning. IEEE Access **10**, 78726–78738 (2022) https://doi.org/10.1109/ACCESS.2022.3193700

Banerjee, I., Bhattacharjee, K., Burns, J.L., Trivedi, H., Purkayastha, S., Seyyed-Kalantari, L., Patel, B.N., Shiradkar, R., Gichoya, J.: "shortcuts" causing bias in radiology artificial intelligence: Causes, evaluation, and mitigation. Journal of the American College of Radiology **20**(9), 842–851 (2023) https://doi.org/10.1016/j.jacr.2023.06.025

Bottani, S., Burgos, N., Maire, A., Saracino, D., Ströer, S., Dormont, D., Colliot, O.: Evaluation of mri-based machine learning approaches for computer-aided diagnosis of dementia in a clinical data warehouse. Medical Image Analysis **89**, 102903 (2023) https://doi.org/10.1016/j.media.2023.102903

Bassi, P.R., Dertkigil, S.S., Cavalli, A.: Improving deep neural network generalization and robustness to background bias via layer-wise relevance propagation optimization. Nature Communications **15**(1), 291 (2024) https://doi.org/10.1038/s41467-023-44371-z

Bellamy, D., Hernán, M.A., Beam, A.: A structural characterization of shortcut features for prediction. European Journal of Epidemiology **37**(6), 563–568 (2022) https://doi.org/10.1007/s10654-022-00892-3

Balagurunathan, Y., Mitchell, R., El Naqa, I.: Requirements and reliability of ai in the medical context. Physica Medica **83**, 72–78 (2021) https://doi.org/10.1016/j.ejmp.2021.02.024

Brown, A., Tomasev, N., Freyberg, J., Liu, Y., Karthikesalingam, A., Schrouff, J.: Detecting shortcut learning for fair medical ai using shortcut testing. Nature communications **14**(1), 4314 (2023) https://doi.org/10.1038/s41467-023-39902-7

Chen, A.A., Beer, J.C., Tustison, N.J., Cook, P.A., Shinohara, R.T., Shou, H., Initiative, A.D.N.: Mitigating site effects in covariance for machine learning in neuroimaging data. Human brain mapping **43**(4), 1179–1195 (2022) https://doi.org/10.1002/hbm.25688

Cheng, J.: brain tumor dataset. figshare (2017). https://doi.org/10.6084/M9.FIGSHARE.1512427.V5 . https://figshare.com/articles/dataset/brain_tumor_dataset/1512427/5

Chen, G.-Y., Lin, C.-T.: Multi-task supervised contrastive learning for chest x-ray diagnosis: A two-stage hierarchical classification framework for covid-19 diagnosis. Applied Soft Computing **155**, 111478 (2024) https://doi.org/10.1016/j.asoc.2024.111478

Compton, R., Zhang, L., Puli, A., Ranganath, R.: When more is less: Incorporating additional datasets can hurt performance by introducing spurious correlations. In: Machine Learning for Healthcare Conference, pp. 110–127 (2023). PMLR

DeGrave, A.J., Janizek, J.D., Lee, S.-I.: Ai for radiographic covid-19 detection selects shortcuts over signal. Nature Machine Intelligence **3**(7), 610–619 (2021) https://doi.org/10.1038/s42256-021-00338-7

Dogra, V., Verma, S., Woźniak, M., Shafi, J., Ijaz, M.F., et al.: Shortcut learning explanations for deep natural language processing: A survey on dataset biases. IEEE Access (2024)

Dvijotham, K., Winkens, J., Barsbey, M., Ghaisas, S., Stanforth, R., Pawlowski, N., Strachan, P., Ahmed, Z., Azizi, S., Bachrach, Y., *et al.*: Enhancing the reliability and accuracy of ai-enabled diagnosis via complementarity-driven deferral to clinicians. Nature Medicine **29**(7), 1814–1820 (2023) https://doi.org/10.1038/s41591-023-02437-x

Esmaeilzadeh, P.: Challenges and strategies for wide-scale artificial intelligence (ai) deployment in healthcare practices: A perspective for healthcare organizations. Artificial Intelligence in Medicine **151**, 102861 (2024) https://doi.org/10.1016/j.artmed.2024.102861

Fay, L., Cobos, E., Yang, B., Gatidis, S., Küstner, T.: Avoiding shortcut-learning by mutual information minimization in deep learning-based image processing. IEEE Access **11**, 64070–64086 (2023) https://doi.org/10.1109/access.2023.3289397

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)

Gichoya, J.W., Thomas, K., Celi, L.A., Safdar, N., Banerjee, I., Banja, J.D., Seyyed-Kalantari, L., Trivedi, H., Purkayastha, S.: Ai pitfalls and what not to do: mitigating bias in ai. The British Journal of Radiology **96**(1150), 20230023 (2023) https://doi.org/10.1259/bjr.20230023

Hatherley, J.J.: Limits of trust in medical ai. Journal of medical ethics **46**(7), 478–481 (2020) https://doi.org/10.1136/medethics-2019-105935

Huang, S.-C., Chaudhari, A.S., Langlotz, C.P., Shah, N., Yeung, S., Lungren, M.P.: Developing medical imaging ai for emerging infectious diseases. nature communications **13**(1), 7060 (2022) https://doi.org/10.1038/s41467-022-34234-4

Hooper, S., Chen, M., Saab, K., Bhatia, K., Langlotz, C., Ré, C.: A case for reframing automated medical image classification as segmentation. Advances in Neural Information Processing Systems **36** (2024)

Huang, J., Huang, K., Xu, M., Liu, F.: Cfab: An online data augmentation to alleviate the spuriousness of classification on medical ultrasound images. In: International Conference on Computer Vision Systems, pp. 91–101 (2023). https://doi.org/10.1007/978-3-031-44137-0_8 . Springer

Haynes, S.C., Johnston, P., Elyan, E.: Generalisation challenges in deep learning models for medical imagery: insights from external validation of covid-19 classifiers. Multimedia Tools and Applications, 1–20 (2024) https://doi.org/10.1007/s11042-024-18543-y

Jabbour, S., Fouhey, D., Kazerooni, E., Sjoding, M.W., Wiens, J.: Deep learning applied to chest x-rays: exploiting and preventing shortcuts. In: Machine Learning for Healthcare Conference, pp. 750–782 (2020). PMLR

Kauffmann, J., Ruff, L., Montavon, G., Müller, K.-R.: The clever hans effect in anomaly detection. arXiv preprint arXiv:2006.10609 (2020) https://doi.org/10.48550/arXiv.2006.10609

Luo, L., Chen, H., Xiao, Y., Zhou, Y., Wang, X., Vardhanabhuti, V., Wu, M., Han, C., Liu, Z., Fang, X.H.B., *et al.*: Rethinking annotation granularity for overcoming shortcuts in deep learning–based radiograph diagnosis: A multicenter study. Radiology: Artificial Intelligence **4**(5), 210299 (2022) https://doi.org/10.1148/ryai.210299

Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 4765–4774. Curran Associates, Inc., ??? (2017). http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.-R.: Unmasking clever hans predictors and assessing what machines really learn. Nature communications **10**(1), 1096 (2019) https://doi.org/10.1038/s41467-019-08987-4

Luo, L., Xu, D., Chen, H., Wong, T.-T., Heng, P.-A.: Pseudo bias-balanced learning for debiased chest x-ray classification. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 621–631 (2022). https://doi.org/10.1007/978-3-031-16452-1_59 . Springer

Müller, N.M., Jacobs, J., Williams, J., Böttinger, K.: Localized shortcut removal. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3720–3724 (2023)

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., Group, P., *et al.*: Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. International journal of surgery **8**(5), 336–341 (2010) https://doi.org/10.1016/j.ijsu.2010.02.007

Mahmood, U., Shrestha, R., Bates, D.D., Mannelli, L., Corrias, G., Erdi, Y.E., Kanan, C.: Detecting spurious correlations with sanity tests for artificial intelligence guided radiology systems. Frontiers in digital health **3**, 671015 (2021) https://doi.org/10.3389/fdgth.2021.671015

Ouyang, C., Chen, C., Li, S., Li, Z., Qin, C., Bai, W., Rueckert, D.: Causality-inspired single-source domain generalization for medical image segmentation. IEEE Transactions on Medical Imaging **42**(4), 1095–1106 (2022) https://doi.org/10.1109/tmi.2022.3224067

Ong Ly, C., Unnikrishnan, B., Tadic, T., Patel, T., Duhamel, J., Kandel, S., Moayedi, Y., Brudno, M., Hope, A., Ross, H., *et al.*: Shortcut learning in medical ai hinders generalization: method for estimating ai model generalization without external data. NPJ Digital Medicine **7**(1), 124 (2024) https://doi.org/10.1038/s41746-024-01118-4

Pearl, J., *et al.*: Models, reasoning and inference. Cambridge, UK: CambridgeUniversityPress **19**(2), 3 (2000)

Pedrosa, J., Aresta, G., Ferreira, C.A., Mendonça, A.M., Campilho, A.: Automatic label detection in chest radiography images. In: BIOIMAGING, pp. 63–69 (2022). https://doi.org/10.5220/0010888100003123

Pezeshki, M., Kaba, O., Bengio, Y., Courville, A.C., Precup, D., Lajoie, G.: Gradient starvation: A learning proclivity in neural networks. Advances in Neural Information Processing Systems **34**, 1256–1272 (2021)

Pohjonen, J., Stürenberg, C., Rannikko, A., Mirtti, T., Pitkänen, E.: Spectral decoupling for training transferable neural networks in medical imaging. Iscience **25**(2) (2022) https://doi.org/10.1016/j.isci.2022.103767

Qu, J., Xiao, X., Wei, X., Qian, X.: A causality-inspired generalized model for automated pancreatic cancer diagnosis. Medical Image Analysis **94**, 103154 (2024) https://doi.org/10.1016/j.media.2024.103154

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)

Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806 (2014) https://doi.org/10.48550/arXiv.1412.6806

Samhita, L., Gross, H.J.: The "clever hans phenomenon" revisited. Communicative & integrative biology **6**(6), 27122 (2013) https://doi.org/10.4161/cib.27122

Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: International Conference on Machine Learning, pp. 3145–3153 (2017). PMlR

Sun, S., Koch, L.M., Baumgartner, C.F.: Right for the wrong reason: Can interpretable ml techniques detect spurious correlations? In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 425–434 (2023). https://doi.org/10.1007/978-3-031-43895-0_40 . Springer

Shah, M., Sureja, N.: A comprehensive review of bias in deep learning models: Methods, impacts, and future directions. Archives of Computational Methods in Engineering, 1–13 (2024)

Souza, R., Wilms, M., Camacho, M., Pike, G.B., Camicioli, R., Monchi, O., Forkert, N.D.: Image-encoded biological and non-biological variables may be used as shortcuts in deep learning models trained on multisite neuroimaging data. Journal of the American Medical Informatics Association **30**(12), 1925–1933 (2023) https://doi.org/10.1093/jamia/ocad171

Souza, R., Winder, A., Stanley, E.A., Vigneshwaran, V., Camacho, M., Camicioli, R., Monchi, O., Wilms, M., Forkert, N.D.: Identifying biases in a multicenter mri database for parkinson's disease classification: Is the disease classifier a secret site classifier? IEEE Journal of Biomedical and Health Informatics (2024) https://doi.org/10.1109/jbhi.2024.3352513

Ting, D.S., Liu, Y., Burlina, P., Xu, X., Bressler, N.M., Wong, T.Y.: Ai for medical imaging goes deep. Nature medicine **24**(5), 539–540 (2018) https://doi.org/10.1038/s41591-018-0029-3

Trivedi, A., Robinson, C., Blazes, M., Ortiz, A., Desbiens, J., Gupta, S., Dodhia, R., Bhatraju, P.K., Liles, W.C., Kalpathy-Cramer, J., *et al.*: Deep learning models for covid-19 chest x-ray classification: Preventing shortcut learning using feature disentanglement. Plos one **17**(10), 0274098 (2022) https://doi.org/10.1371/journal.pone.0274098

Thiagarajan, J.J., Thopalli, K., Rajan, D., Turaga, P.: Training calibration-based counterfactual explainers for deep learning models in medical image analysis. Scientific reports **12**(1), 597 (2022) https://doi.org/10.1038/s41598-021-04529-5

Wallis, D., Buvat, I.: Clever hans effect found in a widely used brain tumour mri dataset. Medical image analysis **77**, 102368 (2022) https://doi.org/10.1016/j.media.2022.102368

Wang, R., Kuo, P.-C., Chen, L.-C., Seastedt, K.P., Gichoya, J.W., Celi, L.A.: Drop the shortcuts: image augmentation improves fairness and decreases ai detection of race and other demographics from medical images. EBioMedicine **102** (2024) https://doi.org/10.1016/j.ebiom.2024.105047

Wainwright, R., Wang, D., Layton, H., Bialkowski, A.: Beyond model accuracy: Identifying hidden underlying issues in chest x-ray classification. In: Australasian Joint Conference on Artificial Intelligence, pp. 533–544 (2023). https://doi.org/10.1007/978-981-99-8388-9_43 . Springer

Wu, C., Yang, X., Gilkes, E.G., Cui, H., Choi, J., Sun, N., Liao, Z., Fan, B., Santillana, M., Celi, L., Silva, P., Nakayama, L.: De-identification and obfuscation of gender attributes from retinal scans. In: Clinical Image-Based Procedures, Fairness of AI in Medical Imaging, and Ethical and Philosophical Issues in Medical Imaging, pp. 91–101. Springer, ??? (2023). https://doi.org/10.1007/978-3-031-45249-9_9

Xiong, Y., Du, B., Xu, Y., Deng, J., She, Y., Chen, C.: Pulmonary nodule classification with multi-view convolutional vision transformer. In: 2022 International Joint Conference on Neural Networks (IJCNN), pp. 1–7 (2022). https://doi.org/10.1109/ijcnn55064.2022.9892716 . IEEE

Zhao, Q., Adeli, E., Pohl, K.M.: Training confounder-free deep learning models for medical applications. Nature communications **11**(1), 6010 (2020) https://doi.org/

10.1038/s41467-020-19784-9

Zhang, R., Griner, D., Garrett, J.W., Qi, Z., Chen, G.-H.: Training certified detectives to track down the intrinsic shortcuts in covid-19 chest x-ray data sets. Scientific Reports **13**(1), 12690 (2023) https://doi.org/10.1038/s41598-023-39855-3

Zare, S., Nguyen, H.V.: Removal of confounders via invariant risk minimization for medical diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 578–587 (2022). https://doi.org/10.1007/978-3-031-16452-1_55 . Springer

Zhang, J., Zhang, Z.-m.: Ethics and governance of trustworthy medical artificial intelligence. BMC medical informatics and decision making **23**(1), 7 (2023) https://doi.org/10.1186/s12911-023-02103-9