Artificial Intelligence Review (2025) 58:12
https://doi.org/10.1007/s10462-024-11005-9

Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust AI applications

Haseeb Javed[1] · **Shaker El-Sappagh**[1,2,3] · **Tamer Abuhmed**[1]

Accepted: 12 October 2024 / Published online: 8 November 2024

Abstract
The current study investigates the robustness of deep learning models for accurate medical diagnosis systems with a specific focus on their ability to maintain performance in the presence of adversarial or noisy inputs. We examine factors that may influence model reliability, including model complexity, training data quality, and hyperparameters; we also examine security concerns related to adversarial attacks that aim to deceive models along with privacy attacks that seek to extract sensitive information. Researchers have discussed various defenses to these attacks to enhance model robustness, such as adversarial training and input preprocessing, along with mechanisms like data augmentation and uncertainty estimation. Tools and packages that extend the reliability features of deep learning frameworks such as TensorFlow and PyTorch are also being explored and evaluated. Existing evaluation metrics for robustness are additionally being discussed and evaluated. This paper concludes by discussing limitations in the existing literature and possible future research directions to continue enhancing the status of this research topic, particularly in the medical domain, with the aim of ensuring that AI systems are trustworthy, reliable, and stable.

Keywords AI robustness · Adversarial attack · Deep learning models · Medical diagnosis · Adversarial input · Model security

✉ Tamer Abuhmed
tamer@skku.edu

[1] Department of Computer Science and Engineering, College of Computing and Informatics, Sungkyunkwan University, Suwon, South Korea

[2] Faculty of Computer Science and Engineering, Galala University, Suez, Egypt

[3] Faculty of Computers and Artificial Intelligence, Benha University, Benha, Egypt

1 Introduction

In recent years, deep learning models have made significant progress in their ability to contribute to medical diagnoses. These models have shown great potential in their ability to accurately detect and diagnose conditions ranging from various domains including infectious, genetic, autoimmune, deficiency, degenerative, and mental diseases (Hong and Zeng 2023). However, the continually increasing complexity of these models has led to concerns about their robustness (Litjens et al. 2017; Rodriguez et al. 2022). Model robustness refers to the ability of a deep learning model to perform consistently and accurately when used with a wide range of input data, including data that may be noisy, incomplete, or confounded by various sources of interference; a robust model can resist the effects of such data and maintain high levels of performance even in the face of adversarial attacks or unexpected input conditions (Drenkow et al. 2021; Chen and Liu 2023). In the medical domain, robustness is key to ensuring accurate diagnoses are consistently made regardless of differences in patient data, imaging techniques, or other factors that may affect the quality of the input data. Robustness is an essential property of deep learning models that are used in medical systems, and such models must support a wide range of inputs, such as different types of medical images or patient data, while also having the ability to consistently provide accurate diagnoses under various conditions (Apostolidis and Papakostas 2021) Medical systems using deep learning models have shown a promising ability to accurately identify and diagnose various medical conditions (Yi et al. 2023). However, these models are vulnerable to adversarial attacks, where malicious entities alter input data to deceive the model, thus leading to false diagnoses (Eren and Küçükdemiral 2024).

The strength of a deep learning model is affected by several elements, including the richness and diversity of the data used for training, the complexity of the model's design, and the effectiveness of the optimization methods employed (Hu et al. 2021). Broadly, the factors that influence the robustness of the model can be categorized into a number of groups (Zhou et al. 2021), including:

(1) Quality and quantity of data: The robustness of a model is significantly influenced by the quality and volume of the data used in that model's training and testing phases. The use of large and diverse datasets can help models learn to generalize better, thus making them more robust to variations and noise sources in data.
(2) Model architecture: The architecture of the deep learning model used in medical systems is another critical factor that influences robustness. Complex models may have more parameters and layers, thus making them more prone to overfitting and less robust to variations and noises in the data (Juraev et al. 2405). Meanwhile, simple models may not have enough capacity to learn from the data, thus causing them to suffer from underfitting. The aim is to strike a balance between complexity and simplicity to achieve a robust model (Rodriguez et al. 2022). Since increased model complexity can indeed heighten the risk of overfitting, particularly in the medical domain, where datasets are often variable and sometimes limited, several studies have explored and implemented techniques to manage the trade-off between model complexity and overfitting. These include external validation, which provides an additional layer of assessment for model robustness, along with regularization methods such as L2 regularization and dropout.

Moreover, data augmentation is frequently used to artificially expand datasets, thereby reducing the likelihood of overfitting (Litjens et al. 2017; Rodriguez et al. 2022).

(3) Hyperparameters: Learning rate, batch size, and regularization are crucial hyperparameters in determining a model's robustness. The optimal setting of parameters is key to enhancing both the model's performance and its robustness. Utilizing optimization strategies like grid search or random search to fine-tune these hyperparameters can substantially increase a model's robustness (Roy et al. 2023a; Arnold et al. 2024).
(4) Adversarial attacks: Adversarial attacks pose a major risk to the robustness of deep learning models in medical systems. Adversarial attacks involve manipulating input data to mislead the model into producing incorrect diagnoses. Models that are vulnerable to adversarial attacks are less robust, which can have serious consequences in medical systems. Therefore, developing models that are resistant to adversarial attacks is critical for achieving robustness.
(5) Interpretability: Interpretability is another important factor that influences the robustness of deep learning models used in medical systems. Models that are more interpretable and transparent can help identify potential biases or errors and improve the robustness of the model. Interpretable models can also provide better explanations for their diagnoses, thus improving their trustworthiness and acceptability in medical systems.

In addition to the previous factors, privacy attacks, which aim to extract sensitive information from the model or the training data, also represent key considerations affecting model robustness (Zhao et al. 2023; Dong et al. 2020a). These attacks can occur through various means, including model inversion attacks, membership inference attacks, and attribute inference attacks. As robustness is crucial for diagnosis models, it is essential to ensure that AI-based medical systems are robust and secure against all possible adversarial and privacy attacks (Eren and Küçükdemiral 2024).

Figure 1 depicts the core attributes necessary that are for the robustness of medical diagnostic systems, along with the interdependent relationships among various key components affecting robustness. In this context, to help ensure clinical validity, a robust system must be able to handle diagnostic tasks reliably and without failure under diverse conditions (Windmann et al. 2023). This reliability is based on resiliency, which is the system's ability to cope with challenges and maintain operational integrity. A system's resiliency can be strengthened by the implementation of robust security measures to protect against potential breaches or data corruption (Zhang et al. 2023a). Security not only protects a system but also enhances its accessibility and scalability, ultimately ensuring that the system can be expanded and adapted to increase system capacity or diagnostic requirements without compromising performance or data integrity. Meanwhile, robustness improves usability; a more stable and resilient system can be more intuitive and user-friendly for healthcare professionals. Usability itself contributes to scalability by simplifying user interactions, thereby promoting wider adoption and seamless integration into healthcare workflows (Chivukula et al. 2023). Robustness also directly facilitates model calibration, which is crucial in fine-tuning the system for accurate diagnostics and ensuring that the system's outputs are precise and trustworthy (Ullah et al. 2021). Data privacy, which represents an important issue in healthcare, must be protected by a robust system that ensures that sensitive patient information is managed and protected in a secure manner. Another critical feature that a robust system

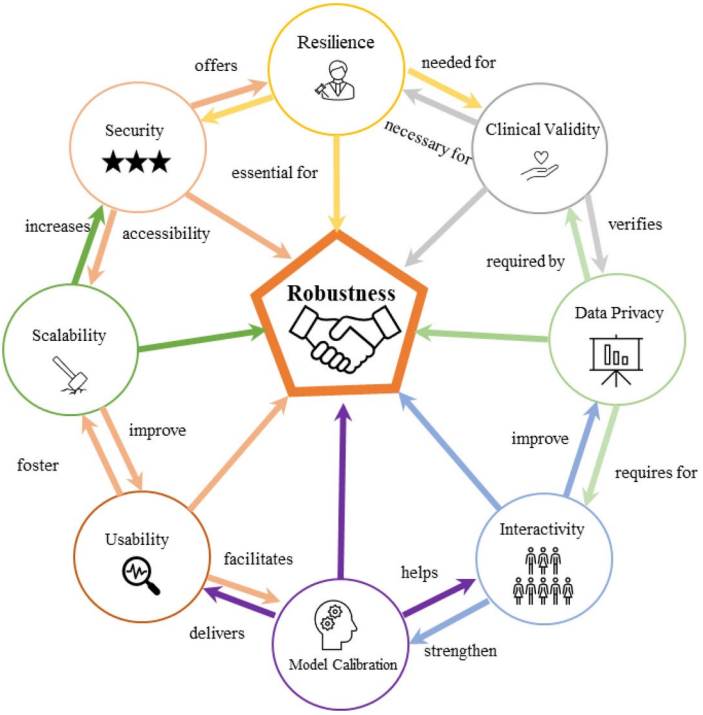

Fig. 1 Mapping the pillars of robustness in medical diagnosis

should support is interactivity, which enhances the user's ability to interact with the system and can lead to better calibration and greater data privacy through active user involvement and feedback. Together, these attributes form a robust framework that aims to provide reliable, secure, and user-centric medical diagnostic tools, which are indispensable in modern healthcare systems (Shibly et al. 2023).

A robustness review of a medical system can reveal insights into the most recent developments in the field, while highlighting existing challenges and potential areas for future research. Various strategies have been proposed in attempts to strengthen the robustness of medical systems, including the use of robust and adversarial training methods (Greco et al. 2023). Robust training involves incorporating noise or perturbations into the training dataset, making the model more resistant to adversarial attacks; adversarial training involves training the model using not only the original data but also adversarial examples, which helps increase its resilience to adversarial attacks.

There are certain tools and packages that have been developed to enhance the robustness of medical systems. For example, TensorFlow Privacy is a library that provides mechanisms for ensuring privacy and security in deep-learning models (Ning et al. 2023). Other tools include CleverHans and Foolbox, which are frameworks for generating and evaluating adversarial examples. Various evaluation metrics, such as accuracy, robustness, and privacy, have been proposed for use in evaluating these mechanisms and tools. These metrics can be used to compare different models and mechanisms while also identifying areas for improvement (Javaid et al. 2022). The present review provides an in-depth analysis of the

robustness of deep learning-based medical systems, which is expected to provide valuable insights into the current state of knowledge in this field while also identifying challenges and opportunities for future research. We believe that enhancing the robustness of these systems can improve their reliability and accuracy in diagnosing various medical conditions, which would ultimately lead to better healthcare outcomes (Gojić et al. 2023).

Our current study reviews the existing literature that evaluates the robustness of deep learning (DL) models. We aim to acquire a deeper understanding of the prevailing trends in this area to identify future research directions. Specifically, we intend to address the following key research questions:

RQ1 How do various factors, such as model complexity, the quality of the training data, and hyperparameter settings, influence the robustness of deep learning models, particularly in the medical domain, and what is their mathematical modeling?

RQ2 What are the security and privacy threats to deep learning models, specifically focusing on adversarial attacks that deceive models and others that aim to extract confidential information?

RQ3 Which tools and packages, based on TensorFlow and PyTorch, are available for enhancing the reliability of deep learning models, and how effective are the current evaluation metrics for assessing robustness while also highlighting the need for improved measures?

Contributions The main contributions of this review are as follows

1. We present an exhaustive study and evaluation of several robustness types for deep learning models and their mathematical modeling in medical systems. We discuss various factors that influence these robustness types, including model complexity, quality of training data, and hyperparameter settings.
2. We delve into the security threats posed by adversarial attacks that attempt to fool models and the privacy threats that arise when attackers aim to extract confidential information.
3. Based on TensorFlow and PyTorch, we provide the most important readily available tools and packages for increasing model reliability. We critique the current evaluation metrics used to assess robustness while emphasizing the need for more effective measures.

As illustrated in Fig. 2, the rest of this review paper is organized as follows. Section 1: Introduction sets the stage by describing the significance of machine learning in medical diagnostics. Section 2: Related Work provides a comprehensive overview of the existing literature in this area. Section 3: Responsible Machine Learning Based—Medical Applications discusses critical aspects such as bias, data privacy, clinical validation, and ethical considerations, focusing on the robustness of diagnostic tools in specific applications. Section 4: Assessing and Improving Robustness: Current Tools and Techniques explores the various tools that are used to enhance model robustness. Meanwhile, Section 5: Measuring Robustness: Quantitative Approaches and Evaluation Techniques introduces the quantitative methods that are used to assess model robustness. Section 6: Future Directions and

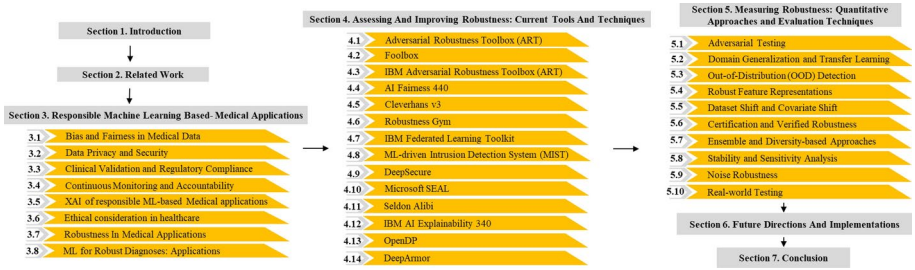

Fig. 2 Structural outline of the survey of robustness in AI-enhanced medical diagnostic system

Implementations discusses the expected advancements in robust diagnostic systems, while Sect. 7: Conclusion summarizes the findings of the paper.

The study addresses the application-specific aspects of robustness that contribute to the practical and ethical deployment of medical systems. Interpretability ensures that users understand how the system works, thus reflecting the need for physicians to communicate their diagnostic conclusions to patients. Scalable Robustness includes the system's ability to handle and process large-scale data and complex computational requirements in parallel with the healthcare system's response to high patient inflows during peak times. Ethical Robustness underscores the system's compliance with moral principles, such as protecting patient confidentiality and ensuring equitable treatment, thus reflecting the ethical codes that govern medical practice. Resilient Robustness signifies the system's ability to withstand and adapt to unexpected challenges, an example being a hospital's response to critical events. This study explores the future trajectory of robust diagnostics while discussing the integration of advanced defense mechanisms, inclusive data practices, and the importance of continuous education in this area. Finally, we highlight the need for robustness in enhancing AI-driven medical diagnostics' accuracy and reliability and suggest future research directions.

2 Related work

The growing development of AI in medical systems is attracting significant attention and investment. However, there are several challenges involved in integrating machine learning models in this field, including uncertainty in the models, limitations in practical applicability, and the need to adapt these models to dynamic clinical environments. The safety and reliability of these systems require careful analysis that considers their exposure to a variety of factors, such as variable clinical conditions, diagnostic sensor inaccuracies, and patient-specific variations (Tu et al. 2021). It is often the case that these variables are not directly quantifiable, and this has led to a reliance on certain assumptions during the design and development of these diagnostic tools. The aforementioned study developed two novel defense strategies, i.e., multi-perturbations adversarial training and misclassification-aware adversarial training, that enhance the resilience of models against such attacks (Egli et al. 2023). A key benefit of such research is that it improves the security and reliability of medical diagnostic models against adversarial threats. However, the potential increases in complexity and computational cost of implementing these defense mechanisms represent

limitations. The increase in complexity has led to the need for interconnected systems while also placing an increasing burden on medical professionals, who may struggle to keep pace with such rapid technological advancements. As highlighted by Totschnig in 2023 (Egli et al. 2023), the rapid decision-making that is needed to ensure that safe and accurate medical diagnoses are being made, may exceed human surveillance capabilities. In line with the observations made by Qiu et al. (2023), developing an AI solution that remains effective despite these uncertainties and accurately reflects the complexity of the patient's health status is a significant challenge. Therefore, exploring and enhancing the robustness of AI in medical systems is a crucial aspect of their development.

Taguchi made significant contributions to the field of robustness in the 1960s while working at the Electrical Communications Laboratory (Rashid 2023). His work, which primarily focused on transforming the telecommunications system, has profoundly impacted by combining engineering with statistical methods to increase cost-effectiveness and quality. Taguchi defined robustness as a state in which the performance of a technology, product, or process remains largely unaffected by factors that cause variability (either during manufacturing or in the user's environment) and degradation over time, all while maintaining the lowest possible manufacturing cost. This definition, as cited by Van Biesebroeck, Johannes in 2007 (Biesebroeck 2007), captures Taguchi's comprehensive approach to robustness.

"Robustness refers to a condition in which the performance of a technology, product, or process remains largely unaffected by elements that cause variation, whether these occur during the manufacturing process or in the environment where the product is used. Additionally, it encompasses maintaining performance stability over time, all achieved with the least possible cost of production per unit," based on Genichi Taguchi's concept (Taguchi 1995*).*

In the field of medical diagnostics specifically, robustness refers to the ability of AI systems to adapt to evolving medical knowledge and practices (Miller 2019). As medical guidelines and treatment protocols continually evolve, diagnostic systems must be able to incorporate new information and adapt their algorithms accordingly. There have been some studies focusing on the development of adaptable AI models that can be updated or retrained with new data without requiring extensive reengineering. This is essential for AI-based diagnostic tools to remain relevant and accurate over time, particularly in fast-evolving medical fields like oncology or genetics.

Figure 3 illustrates the concept of robustness. The two axes, X1 and X2, represent different design parameters impacting the performance metrics. The large blue shape outlines

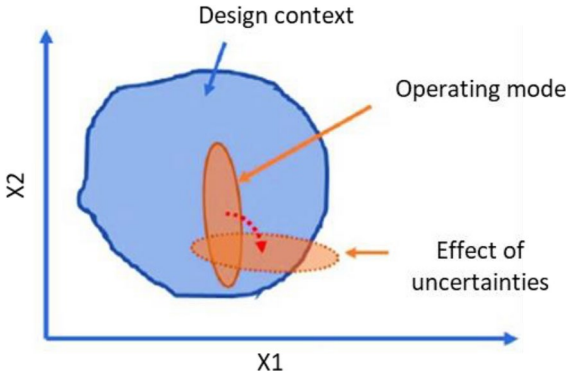

Fig. 3 Visualizing the concept of robustness in terms of expected and unexpected factors that influence the model operability and correctness

the range of conditions under which the design is expected to operate effectively. Inside this blue shape, a smaller orange shape with a solid outline labeled "Operating mode" represents the typical operational conditions of the system. The similar orange shape with a dotted outline indicated by the "Effect of uncertainties" arrow suggests differences in performance or behavior due to unpredictable factors. The robustness of the system lies in its ability to maintain functionality despite these uncertainties. The drawing visually illustrates that a robust design will remain functional (in the orange region) even when subjected to various uncertainties that push its performance beyond the outer limits of the design context. This illustrates the idea that a robust system can withstand or adapt to unexpected changes without experiencing any significant performance degradation.

Recent research in the medical diagnosis field has increasingly focused on the robustness of AI-based systems. These studies highlight the critical need for AI models to perform consistently and accurately across a variety of clinical conditions. As an illustrative example, one line of research investigates how AI systems respond to diverse patient data, including unusual cases or rare conditions. This is crucial because diagnostic errors in such situations can have significant implications for patient care. Another aspect of robustness that researchers are actively examining is the ability of AI systems to maintain their performance when faced with certain data inconsistencies, such as variations in imaging quality or different imaging modalities. Many recent publications show that the focus of medical diagnostic research dealing with robustness has shifted toward ensuring the stability and security of artificial intelligence systems, as discussed by Xu et al. (2021b).

This study provides a critical analysis of the vulnerabilities that deep learning-based diagnostic models face when subjected to adversarial attacks. These attacks involve subtle perturbations to medical images that can deceive the model into making incorrect predictions with the errors remaining imperceptible to human observers. That study systematically exposed three deep learning models to such adversarial scenarios in both single-label and multi-label classification tasks, with the results revealing significant concerns about the reliability of these models under adversarial conditions. By rigorously examining the impact of these attacks on model performance, the authors underscore the importance of incorporating robustness measures that not only detect but also mitigate such vulnerabilities, ultimately ensuring that AI systems remain secure and dependable in clinical settings. This research reinforces the need for ongoing efforts to fortify AI models against adversarial threats, ultimately emphasizing the need for advanced strategies like adversarial training and robust model architectures in the medical domain. Future studies aiming to build upon this work could focus on streamlining these defense methods to be more computationally efficient while maintaining, or even enhancing, their protective capabilities. In the rapidly evolving field of machine learning, the robustness of systems against various perturbations remains a paramount concern. Table 1 presents a systematic comparison of recent studies that have delved into the robustness of different machine learning models, with areas of focus ranging from medical diagnostics to supply chain management and NLP systems. This table highlights the year of publication, sample sizes, and the presence or absence of key robustness features such as adversarial example testing, domain shifts, and dataset variance. In delineating these parameters, the table sets the stage for a critical analysis of the breadth and depth of each study's investigation of robustness, in the process pinpointing gaps and suggesting potential avenues for future research.

Table 1 Previous studies examining robustness in deep learning models

Study	Published Year	Literature Coverage Range	Adversarial Robustness	Noise Robustness	Conceptual Robustness	Model Robustness	Domain Robustness	Interpretable Robustness	Scalable Robustness	Ethical Robustness	Resilient Robustness	Transfer Robustness	Study theme
Ours	2024	381	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust AI applications
Xie et al. (2019)	2023	27	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	Denoising for enhancing adversarial robustness
Freitas et al. (2023)	2023	138	✓	✗	✓	✗	✗	✗	✗	✓	✗	✗	Graph vulnerability and robustness
Marinagi et al. (2023)	2023	176	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	Resilient Supply Chain 4.0
Song et al. (2022)	2022	174	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	Noisy labels with deep neural networks
Wang et al. (2022a)	2022	266	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗	Robustness in NLP Models
Li et al. (2022)	2022	197	✗	✗	✗	✓	✗	✓	✗	✗	✗	✗	Interpretable deep learning
Rudin et al. (2022)	2022	341	✗	✗	✗	✗	✗	✓	✗	✓	✗	✗	Interpretable machine learning and some of their fundamental principles
Yang and Zhou (2022)	2022	86	✗	✗	✗	✗	✗	✗	✓	✗	✗	✓	Scalable and robust polygenic score methods for biobank studies
Rasheed et al. (2022)	2022	189	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	Ethical and Trustworthy ML in Healthcare
Ahmad et al. (2022)	2022	348	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	Ethical robustness in smart city security and data management
Thomas et al. (2022)	2022	149	✗	✓	✗	✗	✗	✗	✗	✗	✗	✓	Interpreting mental state decoding
Nguyen et al. (2022)	2022	181	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	Transfer Learning for Wireless Networks
Xu et al. (2021a)	2021	78	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	Deep learning models on graph
Bai et al. (2021)	2021	69	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	Adversarial training for adversarial robustness
Silva and Najafirad (2007)	2020	136	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	Adversarial robustness survey
Muhammad and Bae (2022)	2020	103	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	Efficient methods for adversarial robustness
Argyroudis (2021)	2019	59	✗	✗	✓	✗	✗	✗	✗	✗	✓	✗	Resilience metrics for transport
Oktian et al. (2017)	2017	42	✗	✗	✗	✓	✗	✗	✓	✗	✓	✗	Scalable and robust distributed SDN controller design
Cuadra et al. (2015)	2015	203	✗	✗	✓	✗	✗	✗	✗	✗	✓	✗	Robustness using complex networks concepts
Mumby et al. (2014)	2014	44	✓	✗	✓	✗	✗	✗	✗	✓	✗	✗	Robustness, ecological resilience and vulnerability

To mention some of these studies for in-depth analysis, one example is Freitas et al. (Song et al. 2022), which focused on addressing graph vulnerability and robustness. The omission of factors such as dataset variance, adversarial examples, and transferability highlights a potential vulnerability in graph-based systems in the face of unexpected inputs. Future research could fill this gap by including these factors and exploring the resilience of graph-based systems in diverse adversarial scenarios.

Figure 4 illustrates a detailed mind map of various components that are critical to the robustness of systems, with this map being centered around the core theme of maintaining consistent performance in the face of diverse challenges. It highlights the importance of adversarial training to prepare systems for specific types of attacks, the need for redundancy to ensure fault tolerance, and the importance of interpretability and transparency in complex models. It also considers the challenges and trade-offs involved in achieving robustness, particularly in terms of efficiency and optimization. It also proposes strategies like ensemble methods that aim to enhance diagnostic accuracy. Moreover, robust software development is highlighted as an essential aspect in creating healthcare applications that will consistently perform under diverse conditions and stresses, thereby contributing to a more reliable healthcare infrastructure.

Studies like Wang et al. (2022a) have also presented remarkable findings on resilient supply chain management, which is a critical area given the recent global disruptions. However, that study did not address several robustness features, with notable examples being adversarial examples and dataset variance. These exclusions could limit the applicability of the study's findings in real-world scenarios where supply chains are subjected to varied and unforeseen challenges. To bridge this gap, subsequent research should incorporate a wider array of robustness tests, particularly while focusing on the adversarial conditions that might affect supply chains.

Another study by Dai et al. (2023) addressed the susceptibility of medical images to adversarial attacks and related to noise robustness, which is challenging due to the existence of complex, high-resolution lesion features. That study introduced a novel defense strategy called global attention noise (GATN) injection, which applies global noise at the input level and targeted noise within feature layers to bolster the images' diagnostic features and smooth out vulnerabilities to minor perturbations. Two variants of GATN, one for images with indistinct lesion boundaries (GATN-UR) and another for those with distinct boundaries (GATN-R), were tested across various medical datasets, with the results showing that these strategies achieved superior robustness and accuracy against powerful adversarial attacks compared to existing methods.

That study presented a promising defense against adversarial attacks in medical imaging that enhances model robustness, particularly in high-stakes diagnosis scenarios. However, that method may incur increased computational costs while also requiring tailoring to specific types of medical imagery. Further improvements could aim to optimize the efficiency of GATN application and extend its adaptability to a broader range of medical image characteristics (Dai et al. 2023).

Moreover, in a study focusing on adversarial robustness against AI, Ghaffari Laleh et al. (2022) presented a method that enhances oncology diagnosis by obtaining biomarkers from pathology slides but which is also exposed to risks related to adversarial attacks. That study revealed that CNNs used in weakly supervised classification are particularly vulnerable to such attacks. Although robust training and DBN can defend against them, these tasks

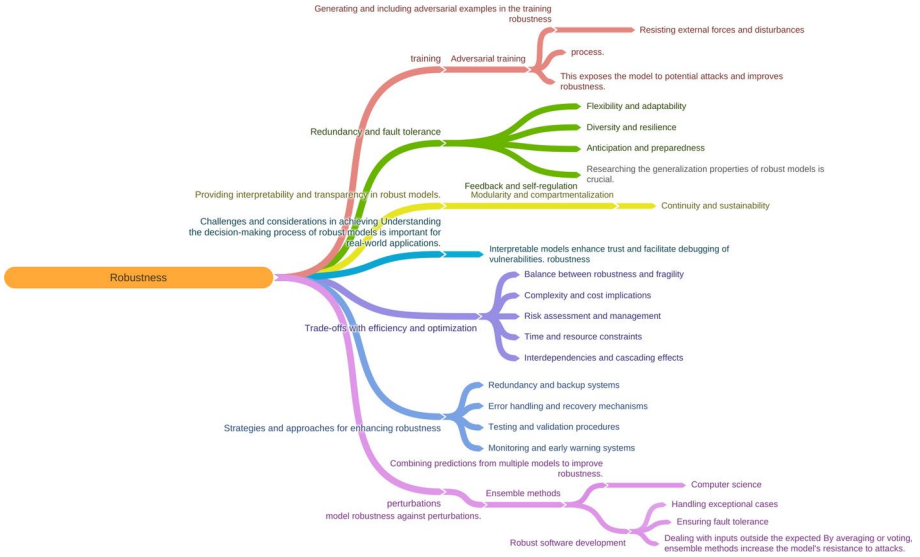

Fig. 4 Mind map of robustness in deep learning systems

require attack-specific knowledge. The authors of that study found that Vision Transformers (ViTs) not only match CNNs in terms of initial performance but also demonstrate significantly greater resistance to adversarial threats, owing to more durable latent representations of clinical data. This suggests that, due to their inherent resilience, ViTs should be favored when deploying AI in computational pathology to protect against data perturbations and adversarial tactics.

Another study by Ma and Liang (2023) addressed the vulnerability of deep neural networks (DNNs) to adversarial noise, particularly focusing on the typical trade-off between the accuracy on clean data and the robustness to noisy data. They introduced increasing margin adversarial (IMA) training, an innovative approach that generates optimal adversarial samples to maintain accuracy while enhancing robustness. Tested on six image datasets under AutoAttack and white-noise attack show that their method stands out for its ability to preserve accuracy and even boost it for medical image segmentation, which represents a potential shift in overcoming the assumed trade-off between standard accuracy and adversarial robustness in this field. However, the method's complexity and computational demands could represent potential limitations. Future improvements could therefore focus on optimizing the training process for efficiency, with the aim of ensuring that the method is scalable and practical for wider application in real-world medical image analysis.

More recently, Joel et al. (2022) explored the vulnerability of deep learning models with a focus on adversarial robustness in oncology, specifically targeting adversarial images that have been manipulated to cause misclassification. That study focused on CT, mammogram, and MRI models, and the results revealed that small pixel-level changes significantly decreased classification accuracy. However, the implementation of an iterative adversarial training approach significantly improved model robustness and stability. Those findings highlight a crucial weakness in DL models, i.e., their sensitivity to minor alterations, which poses a risk for their use in clinical applications. Adversarial training has emerged as an

effective countermeasure that enhances model resilience. Ultimately, the authors of that study underscore the need for such training before deploying DL models in clinical settings, emphasizing the importance of the balance between performance and safety.

In another study, Shi et al. (2022) addressed the susceptibility of CNNs used in medical imaging to subtle adversarial attacks, which represents a significant concern for clinical applications requiring noise robustness. That study identified that inherent noise in medical images exacerbates CNN vulnerability, as these networks unintentionally learn and amplify noisy features. To counter this, the authors developed a novel defense integrating sparsity denoising into CNNs, which proved effective against various sophisticated attacks and preserved over 90% of the original performance in tests. This approach significantly enhances CNN robustness, which is a crucial aspect of safe clinical deployment. However, it is still difficult to integrate denoising without compromising diagnostic accuracy. Future work should focus on refining this balance while ensuring both security and efficacy in medical imaging applications.

Further, Xu et al. (2022) introduced MedRDF, a robust and retrain-less diagnostic framework designed for pre-trained medical models facing adversarial attacks. MedRDF operates at inference, creating numerous noisy variants of a test image to generate a consensus diagnostic output through majority voting, also providing a Robust Metric for result confidence. This approach effectively turns non-robust medical diagnostic models into robust ones without requiring retraining, shown by tests on COVID-19 and DermaMNIST datasets. While MedRDF's ease of deployment and effectiveness in enhancing model robustness are significant advantages, it may face challenges in handling extremely subtle adversarial attacks or maintaining speed in high-throughput diagnostic settings. Future improvements could focus on optimizing the framework's efficiency and sensitivity, thereby ensuring its applicability across a broader range of medical diagnostic scenarios.

Ghamizi et al. (2023) focused on existing perceptions of robustness in deep neural network-based chest x-ray classification, emphasizing the unique challenges involved in medical contexts. Unlike the standard benchmarks that are used in most studies, that study highlighted the complexities in medical diagnosis, such as disease co-occurrence and labeler disagreement. The comprehensive analysis covered three datasets, seven models, and 18 diseases, thus marking an extensive evaluation of this field. The key benefit of that study is the realistic approach it took to assessing model robustness in medical imaging, accounting for factors often overlooked in previous research. However, the diversity in both medical data and diagnostic criteria presents limitations affecting this model's applicability. Future work should continue refining evaluation methodologies while considering the nuanced nature of medical diagnostics and the implications of successful adversarial attacks.

Meanwhile, Wang et al. (2023a) evaluated ChatGPT's resilience to unexpected inputs while focusing on adversarial robustness and out-of-distribution (OOD) scenarios. That research utilized AdvGLUE and ANLI for adversarial tests as well as Flipkart reviews and DDXPlus medical datasets for OOD analysis in comparing ChatGPT with other leading models. The findings of that study revealed ChatGPT's superior performance in most adversarial and OOD contexts, particularly in dialogues and translations. However, there were still limitations in terms of its overall robustness, indicating that such challenges still pose risks to foundational models like ChatGPT.

One notable strength of ChatGPT is its proficiency in understanding and responding to dialogue-based texts. In medical contexts, it tends to offer informal advice rather than

concrete conclusions. However, there is room for improvement in its absolute robustness to adversarial and OOD inputs. Future research should aim to fortify these aspects to increase the model's reliability, particularly for use in critical applications where precision and accuracy are paramount, such as the medical domain.

Amidst the many approaches to robustness in various studies dealing with many problems related to medical system in healthcare, Kireev et al. (2023) addressed the unique challenges of adversarial robustness in tabular data, which is a field that is often overlooked in favor of image and text data. Such an approach is typically taken in situations like fraud detection and medical diagnosis, where robustness is crucial. That study presented an approach that involves creating universal robust embeddings specifically for categorical data within tabular datasets; these embeddings, which are developed using a bilevel alternating minimization framework, can then enhance the robustness of non-neural network algorithms like boosted trees or random forests. This innovation makes it possible to maintain high accuracy in tabular data without the need for adversarial training. The benefit of this method lies in its ability to bring adversarial robustness to a broader range of data types, particularly in critical domains like fraud detection, where accuracy and robustness are paramount. However, this approach may face limitations in terms of its scalability and adaptability to extremely varied or complex tabular datasets. Future enhancements could focus on refining the method for broader applicability, thereby ensuring robustness across diverse scenarios and datasets in the tabular domain.

3 Responsible machine learning based- medical applications

Machine Learning (ML) and Artificial Intelligence (AI) are powerful tools that can help revolutionize the way we diagnose, treat, and manage various medical conditions (Rajkomar et al. 2018). Responsible machine learning (Responsible ML) in medical applications represents an emerging and crucial area of research (El-Sappagh et al. Oct. 2023). There have been many advancements in this domain, ranging from the automation of routine tasks to complex functions like predicting patient outcomes, personalized treatment, and even drug discovery. As we entrust machine learning algorithms with such critical roles in healthcare, the notion of responsibility must come to the forefront, including the consideration of ethical as well as technical concerns (Obermeyer et al. 2019). Here are some notable research areas in Responsible ML for medical applications, along with their limitations and possible solutions.

3.1 Bias and fairness in medical data

Bias in Medical Data Bias is a multifaceted issue that can materialize at various points in the machine learning pipeline—data collection, data preprocessing, feature selection, model training, and even during the inference stage. This form of bias is particularly insidious because it can introduce systematic errors that disproportionately affect underrepresented or marginalized groups (Chen et al. 2019).

Figure 5 illustrates the pipeline of AI model development and deployment, while highlighting the following as stages in which bias can potentially creep in: during data collection, data processing, model development, validation, and deployment, as well as through

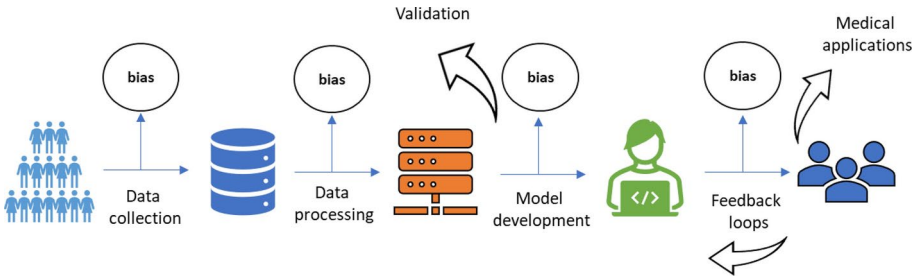

Fig. 5 Bias in AI model development and deployment in healthcare

feedback loops once the model is in use. These biases can stem from the use of unrepresentative training data, flawed data processing techniques, or unequal performance across different demographic groups. The figure emphasizes the importance of vigilance and the implementation of mitigation strategies at each step to ensure that AI systems in healthcare are trained, deployed, and maintained fairly, thereby reducing systematic disparities in medical outcomes across populations (Javed et al. 2021).

As an example, a dataset that has been generated from a population that does not adequately represent minority ethnicities, socio-economic classes, or genders may produce a machine learning model with poor predictive accuracy for these underrepresented groups. Such complications are further compounded when considering that historical biases and systemic healthcare disparities can be deeply ingrained in medical data. For example, certain groups may have been historically under-diagnosed for specific conditions, thus causing the ML models trained on such data to inadvertently perpetuate these disparities (Garg et al. 2018).

Bias can be statistical, such as that involving either underrepresentation or overrepresentation of certain demographic features. It can also be more related to the nature of medical tests and their interpretability. This particularly comes into play when evaluating variables like test sensitivity and specificity, which may vary across different populations due to genetic, environmental, or lifestyle factors (Mewa 2020).

When assessing bias, statistical tests like the Chi-Square test for independence can be used to evaluate if there are significant disparities between the demographic distribution in the data and that of the general population. In the context of machine learning, metrics such as predictive parity and disparate impact ratios are commonly employed. Predictive parity is expressed as follows:

$$PredictiveParity : P\left(\hat{Y} = 1 \,|Y = 1, D = d\right) = P\left(\hat{Y} = 1 \,|Y = 1, D = d'\right)$$

where $\hat{Y}$ is the predicted outcome, Y is the actual outcome, and D and D' are different demographic groups. A predictive parity value close to 1 would indicate the presence of less bias.*Fairness in Medical Data*: In the context of medical data analytics, the concept of fairness constitutes the pursuit of a structured methodology that is designed to guarantee equitable outcomes across various demographic categories when applying machine learning models. A foundational measure in actualizing such a concept of fairness is the accurate execution of a fairness audit on the medical dataset in question (Hardt et al. 2016). Such an

audit employs advanced statistical techniques like disparate impact analysis and counterfactual fairness evaluations, which serve as robust diagnostic tools. These methodologies provide granular insights into the differential impact of model predictions on distinct demographic segments, thus contributing to the understanding of fairness in healthcare models (Sattigeri et al. 2019). In healthcare, fairness is not only an ethical requirement but a matter of life and death. If a biased machine learning model provides inaccurate diagnostic results for certain patient demographics, it could lead to improper treatment and even loss of life. It is therefore crucial to rigorously evaluate fairness metrics such as demographic parity, equalized odds, and disparate impact. Fairness-accuracy trade-offs may occasionally arise that require careful tuning and, in some cases, rethinking of the model architecture to minimize the compromise on either side (Feldman et al. 2015).

There are various techniques that can be used to mitigate bias in data and models, including re-sampling methods to balance the data, cost-sensitive learning, and advanced methods such as adversarial training to ensure model generalization across different demographics (Yuan et al. 2023). Meanwhile, in more complex scenarios, domain adaptation techniques may be used to transfer knowledge from well-represented groups to poorly represented ones, thus improving the model's fairness (Wu et al. 2019a). By scrutinizing both bias and fairness through these advanced methodologies, we move toward achieving a robust and equitable healthcare system facilitated by machine learning. The objective is still to ensure that every individual, irrespective of their demographic characteristics, has equal access to high-quality healthcare. Given the critical role of machine learning models in modern healthcare systems, overlooking bias and fairness could result in the amplification of existing healthcare inequalities, making it critical to study them and mitigate them as much as possible to achieve responsible machine learning applications in medicine (Rasheed et al. 2022).

Fairness can be mathematically framed using concepts like demographic parity, equalized odds, and disparate impact (Tanesini 2021). For instance, demographic parity is achieved if the classification is independent of the sensitive attribute Z:

$$DemographicParity : P\left(\hat{Y} = 1|Z = 0\right) = P\left(\hat{Y} = 1|Z = 1\right)$$

However, demographic parity does not always imply fairness, as it may require the model to ignore relevant features. Equalized odds represent another fairness metric, which is described as:

$$EqualizedOdds : P\left(\hat{Y} = 1|Y = 1, Z = z\right) = P\left(\hat{Y} = 1|Y = 1, Z = z'\right)$$

This ensures that the model is equally accurate for all demographic groups.

Several researchers have also explored fairness and bias in medical machine learning within a comparative framework. One example (Mehrabi et al. 2021) takes an economical approach toward bias mitigation, focusing primarily on data collection and fairness-aware algorithms. While technically sound and cost-effective, this approach potentially overlooks the broader ethical and systemic factors contributing to bias in healthcare. By contrast, another study (Rajkomar et al. 2018) provides a comprehensive exploration of the multifac-

eted issues related to fairness and bias in medical machine learning. Its strength lies in its exhaustive scope, as it addresses an existing gap in the literature. However, its comprehensive nature serves as a double-edged sword; it does not offer actionable, technical solutions that can be readily implemented (Yan et al. 2020; Bin et al. 2022).

Branching out from the healthcare sector, one study (Akter et al. 2021) offers a valuable cross-industry perspective on algorithmic bias. While it is theoretically applicable to healthcare, its generalist insights lack the sector-specific focus that is required to address the unique challenges that are present in medical settings. On the other end of the spectrum, a different study (Price and Nicholson 2019) has conducted a stage-wise investigation of bias in medical AI that focuses almost exclusively on demographic bias. While important, this angle leaves other types of bias, such as socioeconomic or geographic biases, largely unexplored.

Another notable study (Ahmad et al. 2020) takes an ethical standpoint, deeply investigating the broader ramifications of having biased machine learning models in healthcare settings. While ethically important, that study lacks the suggestion of technical solutions that healthcare practitioners can apply in their work. From the above comparison, it is evident that, while numerous valuable studies exist on this topic, each comes with its own set of limitations. Future research could benefit from an integrative approach that combines technical, ethical, and systemic considerations in ML fairness and bias. There is an urgent need for studies focusing on the real-world impact of fairness-aware versus biased medical machine learning models, such as research quantifying the significance of fairness in healthcare outcomes (Sarfraz et al. 2021). There is also a growing demand for research into transparent algorithms that can make the machine learning decision-making process more comprehensible for both healthcare providers and patients. Given the complexity of the challenges involved in bias and fairness in medical machine learning, interdisciplinary research efforts that bring together medical professionals, machine learning researchers, and ethicists could offer valuable insights.

3.2 AI model robustness: data privacy and security

Data privacy and security of medical data are critical in healthcare machine learning. Protecting sensitive patient information and maintaining the confidentiality of healthcare data is an essential aspect of ensuring trust, compliance with regulations, and the ethical use of AI in healthcare (Giuffrè and Shung 2023).

Data Privacy Data privacy focuses on safeguarding confidentiality and individual rights associated with medical data. It also ensures that patients' personal information remains private and is not accessible or misused by unauthorized parties (Nowrozy et al. 2023). In healthcare, sensitive information includes patient medical records, diagnoses, treatment histories, and demographic details. Unauthorized access to this data can result in privacy breaches, identity theft, or discrimination (Larson et al. 2020). The concept of differential privacy represents one of the most mathematically rigorous practical ways to guarantee privacy. In a differentially private algorithm, the probability distribution of the output changes only slightly when a single record is added or removed, with this change quantified by a privacy budget parameter ϵ, also known as the privacy loss parameter (Morley et al. 2021).

$$Pr\left[A\left(D\right) \in S\right] \leq e\epsilon Pr\left[A\left(D'\right) \in S\right]$$

Here, $A(D)$ and $A(D')$ are the outputs of applying a data analysis algorithm A to datasets D and D', while S is a possible output event. A smaller ϵ corresponds to stronger privacy guarantees.

Data Security This involves implementing measures and protocols to protect healthcare data from cyber threats, breaches, and unauthorized access (Garcia Valencia et al. 2023). It encompasses encryption, access controls, and secure storage practices. Ensuring data security is crucial, as healthcare data breaches can result in significant financial losses, reputational damage, and potential harm to patients (Khalid et al. 2023a).

To explain the AI model's robustness with data privacy and security, Fig. 6 maps out security and privacy within the healthcare sector. At the base level, healthcare data, which is governed by hospitals, is collected with a strong emphasis on patient data ownership, with the aim of ensuring that patients have control over their personal information. This data is then utilized in federated learning systems, where AI models are trained on decentralized data sets without compromising patient privacy. To further protect this data, various privacy-preserving techniques, such as differential privacy, introduce statistical noise to data sets to prevent individual data points from being identified. Anonymization and pseudonymization are also used to obscure patient identities; anonymization removes identifying details altogether and pseudonymization replaces private identifiers with fictitious labels.

At the core of this framework is a focus on data privacy and security, which is reinforced by AI with a data protection focus. This ensures that, as algorithms are developed and trained, the integrity and confidentiality of healthcare data are not compromised. The security measures are underpinned by homomorphic encryption, which allows for computations to be performed on encrypted data, and collaborative security, which emphasizes the collective effort that is required to safeguard sensitive information against potential threats like identity theft, data inversion, and manipulation.

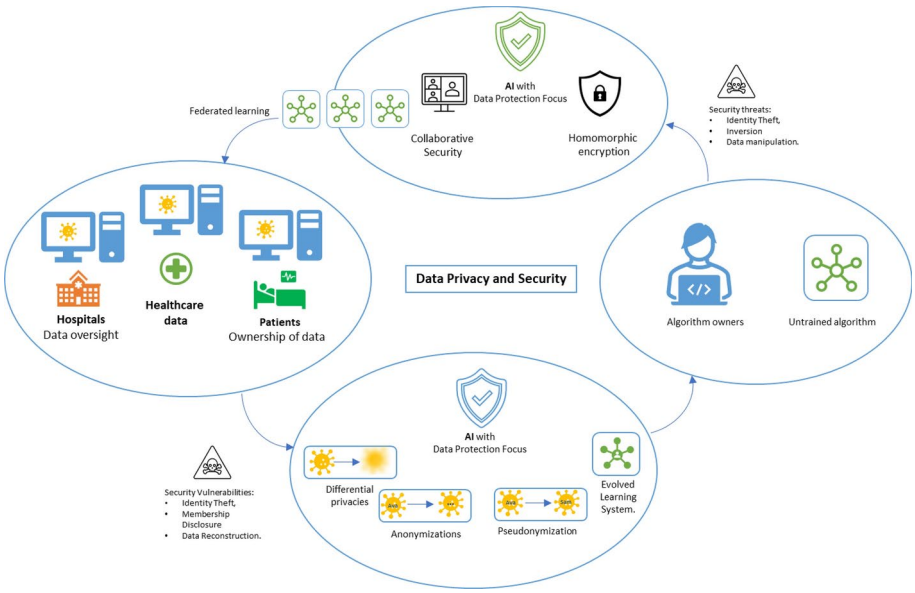

Fig. 6 AI model robustness: the privacy and security spectrum

The goal is to have algorithm owners develop robust, trained algorithms that are secure and reliable. This ensures compliance with regulatory requirements, maintains the ethical use of AI, and, most importantly, upholds the trust of patients, whose data play a critical role in achieving these advancements in healthcare technology.

Challenges in Data Privacy and Security:

1. *Data Sharing vs Privacy*: This trade-off can be quantified using Information Theory metrics such as mutual Information $(X;Y)$, which measures the amount of information one random variable X (in this case, the shared data) reveals about another random variable Y (the sensitive attributes).
2. *Regulatory Compliance*: Compliance with regulations like HIPAA can be considered a constraint in an optimization problem, where the objective function is to maximize data utility subject to these privacy constraints.
3. *Cybersecurity Threats*: Risk analysis often uses Bayesian models to predict the likelihood of different types of cyber threats. For example, if $P(A)$ is the prior probability of a breach, and $P(B|A)$ is the likelihood of observing some data given a breach, then Bayes' theorem gives us the posterior probability $P(A|B)$ of a breach, given the observed data.

Acknowledging the critical importance of data privacy in medical applications, many models that have been proposed in the literature ensure compliance with regulatory standards such as GDPR and HIPAA using a variety of techniques (Khalid et al. 2023b). One common approach involves incorporating differential privacy mechanisms, which add noise to the data in a way that preserves individual privacy while still allowing for accurate model predictions (Choudhury et al. 2019). Federated learning has also emerged as a popular method that enables models to be trained across multiple decentralized devices or servers holding local data samples while eliminating the need to exchange raw data. This technique aligns with GDPR's data minimization principle by ensuring that personal data remains on local devices (Ali et al. 2022). Other studies also emphasize the importance of data encryption during both storage and transmission, using advanced cryptographic techniques, such as homomorphic encryption, to process data without decrypting it, thereby maintaining compliance with HIPAA's security standards. Model architectures like privacy-preserving generative adversarial networks (PPGANs) are also being actively developed to generate synthetic data that closely resembles real data, thus allowing for model training without exposing sensitive information. These methods collectively enhance data security and privacy without compromising model performance, ultimately that AI systems can be effectively deployed in sensitive healthcare settings while adhering to stringent regulatory requirements (Liu et al. 2019).

Meanwhile, a study by Lin et al. (2022) studying DNNs found that they excel in complex tasks such as image recognition and malware identification but are susceptible to adversarial examples wherein inputs are subtly altered to deceive models. The existing defenses to such attacks are not sufficiently scalable. The defense they presented in that paper involves introducing a scalable iterative retraining method that combines Gaussian noise and adversarial techniques to enhance DNN resilience, particularly against established attacks like FGSA,

C&W, PGD, and DeepFool. Their method achieved near-perfect accuracy on standard tests and efficiently handled C&W and BIM attacks, with C&W proving quicker on GPU. It also offered a parallel version of the defense method for handling larger datasets and more intricate models. The limitations of this study include potential overfitting to specific adversarial attacks and the fact that it did not address all types of adaptive adversarial strategies. Future work could explore the adaptability of the defense techniques to new or unseen adversarial tactics and further optimize computational efficiency for real-world applicability.

Another study (Javaid et al. 2023) explored the multifaceted challenges of cybersecurity, with the results ultimately highlighting how healthcare institutions are particularly vulnerable due to a combination of valuable medical data and weaker cybersecurity infrastructures. That study provides a comprehensive overview of threat vectors such as data breaches and phishing attacks but stops short of offering healthcare-specific, actionable recommendations.

Meanwhile, the work in Cyran (2018) presents a different technological perspective for healthcare data management. That study proposes that blockchain technology be used to develop secure and efficient methods for healthcare data sharing. They argue that the decentralized and immutable characteristics of blockchain make it a robust solution to eliminate single points of failure and ensure data integrity. However, that paper focuses primarily on blockchain technology as a remedy for healthcare data challenges, and it ignores the complexities involved in integrating this advanced technology into existing healthcare infrastructure. The limitation here is the lack of consideration of the practical challenges and potential disadvantages of a blockchain-based healthcare data management system.

3.3 Clinical validation and regulatory compliance

Clinical validation and regulatory compliance in healthcare inherently involve taking security and privacy measures to ensure that machine learning models are safe and accurate, protect patient data, and adhere to relevant regulations. These are essential aspects of deploying machine learning models and algorithms as well as AI-driven solutions in healthcare while ensuring patient safety (Ghosh et al. 2021). This crucial topic is elucidated in the following:

Clinical Validation Clinical validation involves rigorous testing and evaluation of machine learning models and algorithms to ensure their accuracy, reliability, and clinical utility in real-world healthcare settings (Walonoski et al. 2018). It typically involves assessing a model's performance in diverse patient populations while considering various clinical scenarios and then comparing its results with those of human experts or established medical standards. Clinical validation helps confirm that the model's predictions and recommendations are clinically meaningful and beneficial to patient care (Misra et al. 2017). The area under the Receiver Operating Characteristic curve (AUC-ROC) is a frequently utilized metric for clinical validation. This curve graphically represents the True Positive Rate (sensitivity) versus the False Positive Rate (1-specificity) at different threshold levels, thereby providing a detailed assessment of the model's capacity to distinguish between classes. Mathematically, AUC-ROC can be represented as:

$$AUC - ROC = \int_0^1 TPR\left(FPR^{-1}(x)\right) dx$$

Another essential statistical measure is the F1 score, which considers both precision and recall when evaluating a model's performance:

$$F1score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Regulatory Compliance Regulatory compliance in healthcare directly mandates the implementation of robust security measures and stringent privacy protections for AI systems to meet legal and ethical standards. In healthcare, regulatory compliance refers to adherence to legal and industry-specific regulations and standards governing the use of AI and data in medical applications. For example, in the United States, the Health Insurance Portability and Accountability Act (HIPAA) establishes protocols for safeguarding the privacy of patient data, while the Food and Drug Administration (FDA) regulates the approval process for medical devices and software. Compliance ensures that healthcare organizations and AI developers meet legal requirements and maintain the highest ethical standards in healthcare AI deployment (Higgins and Johner 2023).

In a clinical setting, these measures are typically computed for diverse patient populations to ensure generalizability across various demographics and clinical scenarios.

For regulatory compliance, one possible approach is to apply frameworks like the Common Criteria for Information Technology Security Evaluation (ISO/IEC 15408), which involves formal mathematical modeling of security attributes and threat scenarios (Bates et al. 2020). Compliance can be statistically quantified through metrics like the Compliance Score (CS), which evaluates an organization's adherence to various regulations:

$$CS = \frac{TotalNumberofStandards}{NumberofCompliantStandards} \times 100$$

Compliance with HIPAA regulations in the United States might also involve statistical assessments of risk levels related to data breaches, which can be modeled using Bayesian networks or other probabilistic models.

$$RiskScore = Likelihoodof Breach \times Impactof Breach$$

Clinical validation and regulatory compliance are closely linked; a machine learning model that excels in clinical performance metrics must also meet stringent regulatory standards to be deployed in healthcare settings (Verma et al. 2020). For example, before obtaining FDA approval, the AUC-ROC or F1 scores of a medical machine learning model must be shown to be statistically significant and clinically relevant. This often involves extensive clinical trials and expert reviews.

One study (Weng 2020) emphasizes the promise of deep learning algorithms for diagnosis and prediction in healthcare settings. While the focus on benchmarking these AI models against clinical standards is certainly a strength, focusing solely on deep learning models may be a significant limitation. In particular, taking such a narrow focus could neglect other machine learning techniques that could be easier to interpret or better suited to specific clinical applications. To address this issue, future research should consider including a variety of

machine learning approaches, such as decision trees or ensemble methods, to ensure a more complete coverage of the tools available for healthcare analytics.

Nicholson (2017) provides invaluable insight into the regulatory landscape. However, focusing solely on regulatory aspects without linking them to real-world clinical validation outcomes may limit the model's applicability. Suggested improvements may include case studies that explicitly show how regulatory requirements have affected the development, validation, and deployment of AI models in healthcare. This would offer a more complete picture and guidance for professionals dealing with these challenges.

However, another study (Magrabi et al. 2019) focused on the difficulties associated with applying machine learning algorithms in healthcare settings. While it is crucial to identify challenges, such a paper might fall short of proposing specific solutions or guidelines for overcoming these hurdles. An appendix or section dedicated to best practices or recommended frameworks for achieving reliable validation drawn from successful case studies would be a helpful addition.

3.4 Model accountability

Accountability is a critical aspect affecting the robustness of machine learning models in healthcare, thereby ensuring the longevity of accurate, safe, and accountable AI systems post-deployment (Pronovost et al. 2018).

Continuous Monitoring This involves regular assessments of the machine learning model's performance and behavior in real-world healthcare settings. Tracking metrics such as accuracy, precision, recall, and others helps identify deviations or degradation in performance (DeVore and Champion 2011). Continuous monitoring can help detect issues such as concept drift (changes in the model's input data over time) or adversarial attacks that may impact model reliability (Meier et al. 2021).

In technical terms, continuous monitoring could involve defining a set of performance metrics $M = \{m1, m2, \ldots, mn\}$ to be evaluated at regular intervals T. Common metrics could include accuracy, precision, and recall. For example, accuracy could be defined mathematically as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Here, TP, TN, FP, and FN represent the counts of true positives, true negatives, false positives, and false negatives, respectively.

Concept drift can be measured using statistical tests. One approach is to perform a Chi-square test assessing independence on data features at different times. If the test indicates a significant change, that could serve as an alert for concept drift.

Accountability Accountability in AI ensures and traces fair decision-making, where measures like Demographic Parity aim to equitably distribute outcomes across sensitive demographic attributes. This involves mathematical representations of fairness. For instance, Demographic Parity might be defined as:

$$P\left(\hat{Y} = 1 | D = 0\right) = P\left(\hat{Y} = 1 | D = 1\right)$$

where $\hat{Y}$ is the prediction and D is a sensitive demographic attribute like race or gender, with the ultimate aim of ensuring equal distribution of positive outcomes across the demographic groups.*Challenges and Technical Solutions* In a healthcare setting, Kullback–Leibler divergence can be used to measure how much the data distribution Pt at time t diverges from a baseline distribution P_0:

$$D_{KL}\left(P_t||P_0\right) = \sum_i P_t\left(i\right) log\frac{P_t\left(i\right)}{P_0\left(i\right)}$$

In healthcare, KL divergence can be used to compare the distribution of data at a certain time P_t with a baseline distribution P_0. If D_{KL} exceeds a predefined threshold, it signals the need for model retraining or adjustment.

Kennedy et al. (2011) at Mayo Clinic Arizona (MCA) developed and tested a data-driven model aiming to achieve improved service quality and long-term value in anticipation of value-based purchasing, where reimbursement is influenced by patient experience and care quality. This model has seven elements, including using varied data to drive improvements, ensuring accountability for service quality, offering consulting tools, emphasizing service values, providing training, continuous supervision, and integrating recognition and rewards. When fully staffed, departments with patient satisfaction levels below the 90th percentile saw significant improvements in patient perceptions. This framework could guide other healthcare institutions transitioning to value-based purchasing systems.

Kass and Faden (2018) discussed how, although medical innovations offer immense potential health benefits, most medical data remains unanalyzed, with insights rarely applied to enhance care. The Learning Health Care System (LHCS) aims to fill this gap, being ethically based on mutual understanding in which patients allow their data to be used for continuous learning, and healthcare systems commit to improving care based on these insights. This study underscores the ethical obligations of LHCS, emphasizing patient rights and dignity and advocating three respect-promoting actions: patient engagement, transparency in learning processes, and accountability in integrating newly acquired knowledge.

3.5 XAI of responsible ML-based medical applications

Explainable AI (XAI) refers to AI systems that are structured to offer transparent and comprehensible justifications for their decisions and behaviors. This transparency is particularly vital in the medical field due to the high stakes involved in healthcare decisions (Lombardi et al. 2022).

Explainable AI (XAI) in Medicine The ability of healthcare professionals to understand and trust AI recommendations is crucial for effective implementation. XAI not only facilitates compliance with regulatory standards that demand transparency in decision-making processes, it also ensures that ethical considerations are met (Jahan et al. 2023a). In healthcare, where decisions can significantly impact patient outcomes, the ability to comprehend and verify AI-driven conclusions is essential for patient safety, ethical responsibility, and maintaining the integrity of medical practices (Khodabandehloo et al. 2021).

Interpretability challenges of AI models in medicine The complexity of AI models, and particularly of deep learning algorithms, poses significant interpretability challenges. These models often function as "black boxes," where the decision-making process is opaque and

not easily understood, even by experts. In medicine, where every decision can have profound implications, the inability to interpret AI decisions can represent a major obstacle (Jahan et al. 2023a).

Misinterpretation of AI outputs or unrecognized errors in AI systems can lead to incorrect diagnoses, inappropriate treatment recommendations, or missed patient nuances. The complexity of medical data, combined with the intricacies of AI algorithms, increases these challenges, ultimately underscoring the importance of explainability in medical AI applications (Band 2023).

Benefits of XAI in Healthcare XAI significantly enhances trust among clinicians and patients, a crucial aspect of any therapeutic relationship. When clinicians understand the rationale behind AI-driven recommendations, their confidence in utilizing these tools increases, thereby fostering a more harmonious integration of AI in clinical settings (Duamwan and Bird 2023). Improved decision-making is another critical benefit; with XAI, physicians can make more informed decisions by understanding the "why" behind AI suggestions. This insight allows them to more effectively compare AI input with their clinical judgment. XAI can also serve as an educational tool for medical professionals. By offering insight into AI decision-making processes, XAI can help clinicians deepen their understanding of certain conditions, thus facilitating continuous learning and improving their professional competencies (Jahan et al. 2023b).

XAI focuses primarily on understanding the decision-making processes of AI models. While XAI itself does not have specific 'mathematical equations' like those found in classical physics or mathematics, it uses various methods and algorithms to interpret complex models. Some of the common methods in XAI and the mathematical concepts on which they are based are described below.

XAI's role in facilitating enhanced learning and medical professional development cannot be overstated. By providing clear explanations for its recommendations, XAI allows healthcare professionals to gain new insights into patient care. This process helps validate the AI's suggestions and contributes to the medical practitioner's knowledge base, ultimately leading to improved patient care practices. For instance, understanding the factors that lead an AI system to identify a particular treatment plan can help clinicians recognize similar patterns in future cases. This ongoing learning process, which is supported by XAI, ensures that medical professionals remain at the forefront of technological advancements by improving their ability to deliver high-quality patient care (Amoroso et al. 2023).

Some researchers have focused on Explainable AI in healthcare, including a study by Essemlali et al. (2020). The authors of that study utilized a modified variant of the deep BrainNet convolutional neural network (CNN) (Ur Rehman et al. 2016) that was trained using diffusion-weighted MRI (DW-MRI) tractography connectomes from individuals with Mild Cognitive Impairment (MCI) and Alzheimer's Disease (AD). This work aimed to better understand structural connectomics in these conditions. They demonstrated that simple BrainNetCNN, when utilized in conjunction with explainable AI methods to classify brain images, could underscore particular brain areas and their links in AD. The results demonstrate that regions exhibiting significant structural differences between groups correspond with those that have been recognized in previous AD studies. This highlights the potential utility of deep learning in elucidating structural connectomes from MRI diffusion tractography to unpack the intricate structure within these connectomes.

However, there are certain limitations to this approach that should be acknowledged, such as potential model overfitting, bias in datasets, and the need for broader validation in diverse patient groups. Future improvements could focus on incorporating larger, more diverse datasets to enhance generalizability and refine AI models to increase accuracy and minimize biases, thus strengthening the reliability of findings in the broader context of degenerative disease research (ur Rehman et al. 2018, 2019).

Kamal et al. (2021) established a novel approach to diagnosing Alzheimer's Disease (AD) by combining machine learning techniques with both MRI images and gene expression data. They utilized SpinalNet and CNN for image analysis along with k-nearest neighbors (KNN), Xboost, and support vector classifier (SVC) for gene expression classification. Notably, the authors integrated an explainable AI method, Local Interpretable Model-Agnostic Explanations (LIME), to make the results more transparent and understandable. This method clarified how the classifiers predicted diseases and identified key genes in AD. While CNN achieved a high accuracy of 97.6%, SVC was found to be the most effective for gene expression data.

However, the complexity of integrating multiple data types and AI methods leads to certain challenges, including potential problems with overfitting and interpreting data. Future improvements could focus on refining the integration of these diverse datasets and further enhancing the interpretability of AI models to support more accurate and reliable diagnostic processes.

3.6 Ethical consideration in healthcare

Informed consent and patient-centered care are critical ethical issues when implementing machine learning models and AI-driven decision support systems in healthcare. These issues are focused on respecting patient autonomy, ensuring transparency, and delivering healthcare that is tailored to individual patient needs and preferences (Čartolovni et al. 2022). An explanation of these important aspects follows:

Informed Consent This is a fundamental ethical principle that requires healthcare providers to obtain free and informed consent from patients before any medical procedure or intervention, including the use of AI and machine learning in their care (Naveed 2023). When AI systems are involved in diagnosing, treating, or assisting in patient care, healthcare providers must communicate the purpose, benefits, risks, and potential alternatives to patients in a way that they can understand. Informed consent provides patients with the autonomy to make informed decisions about their healthcare and to determine whether or not they want AI-driven interventions to be included in their care (Ahmad et al. 2023a).

Several aspects come into play when considering informed consent and patient-centered care from a technical perspective. These include the metrics that are used to quantify the "effectiveness" of consent procedures, privacy-preserving methods in data analysis, and personalization algorithms for patient-centered care (Amann et al. 2020). Statistical measures can be used to quantify the effectiveness of obtaining informed consent. For example, a Bayesian model could be used to compare the "awareness" of one's choices before and after receiving an explanation about the AI-driven intervention. This model can use variables such as time spent reading consent forms, the number of questions asked by the patient, and even psychometric measures to assess understanding.

$$InformednessScore = \frac{PostExplanation - PreExplanationUnderstanding}{TotalPossibleUnderstanding}$$

In this context, "Understanding" can refer to a composite score based on patients' responses to a set of standardized questions designed to assess their grasp of given AI technologies (Jellouli et al. 2023).

Privacy-Preserving Algorithms Techniques like Differential Privacy can be used to address privacy concerns in AI-driven healthcare. This mathematical framework quantifies the extent to which the output of a function (e.g., an AI model's prediction) changes after a single database entry is modified, thus providing a measure of data privacy.

$$\epsilon = ln\left(\frac{Pr\left[Outputwithoutdata\right]}{Pr\left[Outputwithdata\right]}\right)$$

Here, ϵ represents the privacy loss parameter. Lower values of ϵ indicate better privacy guarantees.

Personalization in Patient-Centered Care Machine learning models used in patient-centered care might incorporate multi-criteria decision-making algorithms to weigh various aspects of a patient's medical history, current health status, and personal preferences. For example, the "Weighted Sum Model" could be employed as follows:

$$PersonalizationScore = w1 \times ClinicalOutcome + w2 \times PatientPreference + w3 \times Cost + \cdots$$

where $w1, w2, w3, \ldots$ are weights that can be adapted over time based on patient feedback and clinical effectiveness. However, various quantitative and algorithmic techniques can be applied to measure and enhance the effectiveness of informed consent and patient-centered care in healthcare settings employing machine learning and AI technologies. These methods can serve as the foundation for ongoing assessment and continual improvement in these critical ethical dimensions (Beil et al. 2019).Informed consent and patient-centered care are crucial ethical dimensions in the implementation of machine learning and AI in healthcare. In a relevant study, Naik et al. (2022) highlights that the complexity and opacity of machine learning models pose a significant challenge for healthcare providers in adequately informing patients. That paper suggests that providers often lack the technical expertise to explain the decision-making process used with machine learning models, ultimately leading to incomplete or insufficient informed consent.

Holtz et al. (2023) note that the AI-driven transformation of the medical field will lead to concerns regarding ethical considerations, data privacy, representation, and the potential redundancy of physicians. However, AI cannot replicate a doctor's nuanced understanding of a patient's personal life and surroundings. That paper emphasizes the diminishing focus on patient-centered communication and argues for its resurgence, positing that AI integration could free up more time for meaningful patient-doctor interactions. Such enhanced communication would foster trust, rapport, and empathy, ultimately facilitating better health outcomes.

Harrison and Sidey-Gibbons (2021) delves into the potential utility of unstructured text, such as medical records and patient feedback, as valuable data for clinical research through natural language processing (NLP). Using free software, the authors demonstrate the use of

three NLP techniques on public medicine review data: lexicon-based sentiment analysis of four drugs, clustering of similar drugs using unsupervised machine learning (LDA), and the prediction of drug review sentiments with three supervised machine learning models. The obtained results highlight the viability of these methods, with sentiment analysis revealing varying drug perceptions and varying accuracy of the drug ratings predicted by machine learning models. That study offers a hands-on guide for analyzing extensive text data, complemented by accessible, reproducible code. Future research could address the limitations of these methods and explore their applications in broader clinical contexts.

3.7 Robustness in medical applications

In this section, we present a summary of the various forms of robustness that have been investigated within the realm of medical systems. We also highlight some of the key approaches and techniques that have been suggested as ways to enhance the resilience of these systems potentially. Robustness is a crucial aspect of medical systems as it significantly aids physicians in diagnosing and treating various diseases with greater accuracy and confidence. In the following sections, we will discuss several specific types of robustness that are critical in this context, including Noise Robustness, Adversarial Robustness, Domain Robustness, Conceptual Robustness, Model Robustness, Interpretable Robustness, Scalable Robustness, Ethical Robustness, Resilient Robustness, and Transfer Robustness. Each of these types of robustness addresses different challenges in ensuring that medical systems perform reliably and effectively under various conditions.

Robustness applications: a big picture

Medical systems must effectively manage noisy, incomplete, or inconsistent data, while also being adaptable to changes in the patient population and evolving medical practices over time (Huang et al. 2020; Hamon et al. 2020). In recent years, researchers have focused on developing robust medical systems to improve the accuracy and reliability of medical diagnoses (Akkus et al. 2019). Many studies have focused on exploring the applications of robustness in medical systems to improve their accuracy and reliability (Md Nor et al. 2020; Masud et al. 2021). Various types of robustness, including noise, adversarial, domain, conceptual, and model robustness, have been explored in this context.

Noise Robustness This refers to the ability of a medical system to handle noisy data, such as data with missing values or errors (Najafi et al. 2019; Pandey and Jain 2022). Various factors, such as measurement errors, sensor malfunctions, or data transmission errors, can introduce noise into medical systems. Agarwal and Zhang (2022) focused on developing algorithms that are robust to noise to achieve improved accuracy and reliability in diagnosis systems. Noise robustness deals with the ability of a system to handle noisy data. It can be mathematically modeled using statistical techniques, such as signal-to-noise ratio, mean squared error, or Bayesian inference, to characterize the noise properties and then devise noise-robust algorithms.

To simulate noisy labels in melanoma and lymph node datasets, we can randomly choose a specific percentage γ of images from each class and then change their labels using a conventional method. This leads to the formation of a training data set containing a certain number of images with inaccurate labels. Specifically, the noisy label $y'i$ can be defined as being equal to the clean label yi with a probability of $1 - y$, and as a different label $yk\,(whereyk \neq yi)$ with a probability of $y\cdot$ By introducing this noise, we can obtain a

dataset in which a certain proportion of images has corrupted labels (Xue et al. 2022). Another way to represent noise robustness can be found in regression problems. In regression, a common approach to handle noise is to introduce regularization, such as L2 regularization. The general equation for L2 regularized linear regression, also known as Ridge regression, is as follows: $||y - xW||^2 + \lambda ||W||^2$, where y represents the target variable (dependent variable), x represents the feature matrix (independent variables), W represents the weight vector, $||^2$ denotes the squared Euclidean norm, and λ is the regularization parameter that controls the trade-off between fitting the training data and controlling the magnitude of the weights. The first term, $||y-xW||^2$, measures the squared difference between the predicted target values (XW) and the actual target values (Y). This term quantifies how well the model fits the training data. The second term, $\lambda ||W||^2$, is the regularization term. It penalizes the magnitude of the weight vector W, thus discouraging large weight values. This regularization term helps prevent overfitting and improves noise robustness by reducing the model's sensitivity to small variations in the input data. By adjusting the value of the regularization parameter λ, it is possible to control the strength of the regularization. A higher λ emphasizes the regularization term, leading to a more robust model that is less sensitive to noise in the training data.

Numerous researchers have dedicated their efforts to discussing the robustness of noise in various research domains. In a notable example, Dagni et al. (2018) proposed a training strategy that incorporates label noise directly into a network architecture focusing on the binary classification of breast mammography. Further investigation is needed to understand the types and characteristics of noise that are commonly found in medical image annotations. Such investigation will involve examining noise sources, such as inter-observer variability, imaging artifacts, or inherent ambiguity in image features, to understand the challenges associated with noisy annotations. In another study, Xue et al. (2019) introduced a robust training approach for CNN classifiers that employs online sample mining and re-weighting based on model uncertainty, while a robust training method for CNN classifiers was introduced in a different study by Zhu et al. (2019) that incorporated online sample mining and re-weighting according to model uncertainty while also introducing a technique including an automatic quality assessment module and an overfitting control module to adjust the network parameters. The development of realistic noise models and simulation techniques can also help create synthetic datasets with controlled noise levels and types. This ultimately enables a systematic evaluation of the robustness of deep learning models to different types and levels of annotation noise. Xue et al. (2020) suggested an innovative loss function that integrates noisy labels with local visual indicators to improve semantic segmentation. Xue et al. (2020) introduced a joint optimization scheme that includes a label correction module. In the broader field of deep learning, noise label learning has received considerable attention in natural image processing.

Certain techniques focus on calculating a hidden label transition matrix to adjust for losses associated with noisy samples. For instance, Patrini et al. (2017) estimated the transition matrix by introducing an additional SoftMax operation while implementing a

dual-phase approach that first computes the label transition matrix and then performs error correction on potentially noisy samples. The performance of these techniques depends significantly on estimating the posterior of the noisy class, which is often a difficult task due to the randomness of label noise and intricate data biases, which can potentially lead to errors in the estimation of the transition matrix. Existing evaluation metrics for medical image segmentation may not adequately reflect the impact of noisy annotations. Further research is needed to develop specific evaluation metrics that can quantify the effect of different types and levels of annotation noise on segmentation performance. Studies should focus on accurately modeling noise, designing robust algorithms to handle diverse noise types, assessing the data quality affected by noise, and generalizing noise robustness techniques to various fields, as noise robustness can make systems less vulnerable to attacks that involve exploiting noise injection or interference, thereby enhancing security against certain threats.

Adversarial Robustness Adversarial robustness refers to the ability of a medical system to resist deliberate attacks or manipulations of data, such as by malicious actors seeking to manipulate or subvert the system (Ghaffari Laleh et al. 2022; Wang 2021). Adversarial robustness focuses on designing resilient systems to deliberate attempts to manipulate or subvert their operation. It involves modeling adversarial attacks and developing defense mechanisms using game-theoretic or adversarial modeling frameworks. Adversarial attacks can take various forms, such as input perturbations, data poisoning, or model evasion (Madry et al. 2018). Developing algorithms that are robust to adversarial attacks can ensure the security and integrity of diagnosis systems, as discussed by Miyato et al. (2017). Consider a classification task where we have a dataset with input–output pairs represented as (x, y), where x is the input data, and y is its corresponding label. In this context, we aim to train a model $f(x;\theta)$ parameterized by θ that accurately predicts y for unseen inputs. However, it is also important for the model to withstand adversarial attacks. To frame adversarial robustness in mathematical terms, we can approach it as a minimax optimization challenge. The goal here is to determine the ideal model parameters, denoted as θ*, that simultaneously minimize the classification loss on clean examples and maximize the worst-case loss on adversarial examples. We can express this as:

$$\theta^* = argmin\theta * \max \delta L\left(f\left(x+\delta;\theta\right), y\right) + \alpha R\left(\delta\right)$$

where θ represents the model parameters while δ is the adversarial perturbation applied to the input x. Meanwhile, L is a loss function that measures the model's prediction error, R is a regularization term that encourages small perturbations, and α is a trade-off parameter that controls the importance of robustness relative to accuracy.The inner maximization problem represents the attacker's objective of finding the worst-case perturbation that maximizes the loss of the model; the outer minimization problem represents the defender's objective of finding model parameters that minimize the worst-case loss.

Finlayson et al. (2018) investigates adversarial attacks against deep learning models used in medical applications. The authors of that study explore the vulnerabilities of medical deep learning systems to various attack techniques, and they propose potential defense mechanisms to enhance adversarial robustness. In another study, Kaviani et al. (2022) present a comprehensive survey of adversarial attacks and defenses on artificial intelligence (AI) in medical imaging informatics. It explores different aspects of these methods, including attack generation strategies and their impact on the performance of DNNs. The paper also

highlights the challenges in and future directions for improving the robustness of neural networks in medical imaging informatics. By providing an overview of the current landscape of adversarial attacks and defenses in this domain, this survey enhances the understanding of potential security risks and informs the development of robust AI systems for medical imaging applications. In another study, Navarro et al. (2021) present an evaluation of the robustness of self-supervised learning in the context of medical imaging. That evaluation encompasses various medical imaging modalities and datasets, examining the robustness of self-supervised models against different types of perturbations and attacks. The findings provide insights into the strengths and limitations of self-supervised learning in the medical imaging domain, ultimately highlighting the need for further advancements and defense mechanisms to enhance the robustness of these models.

Since adversarial robustness is a critical consideration in cybersecurity, computer vision, and machine learning, it must address problems such as developing robust defense mechanisms against sophisticated attacks, understanding the limits of robustness, and balancing robustness with other performance criteria. The surveys analyzed also show that adversarial robustness must mitigate the risks associated with data poisoning, evasion attacks, or model manipulation, thus achieving enhanced system security against adversarial threats.

There is also a need for further research and analysis of specific adaptive adversarial attack strategies that can target medical systems (Liu et al. Aug. 2024; Abdukhamidov et al. 2024). Understanding potential attack vectors, such as input perturbations, data poisoning, or adversarial examples that are specific to the medical domain would provide insight into the vulnerabilities of these systems and ultimately enable the development of more robust defense mechanisms.

Further, it would be beneficial to investigate the generalizability of robust defense mechanisms across different medical subdomains. Medical datasets can vary significantly regarding data modalities, imaging techniques, and clinical applications. Therefore, exploring the transferability and performance of adversarial defense mechanisms across diverse medical subdomains would provide a more comprehensive understanding of their effectiveness. Investigating techniques that can simultaneously enhance both robustness and accuracy, such as robust optimization, ensemble methods, or regularization techniques, would also be valuable in developing more practical and effective defense mechanisms for medical systems.

Domain Robustness Domain Robustness in the medical domain that focuses on ensuring that a mathematical model performs consistently and accurately across different domains or distribution shifts. By employing mathematical techniques such as domain adaptation and transfer learning, models can effectively incorporate domain changes and enhance their robustness in the medical domain. Domain robustness refers to the ability of a medical system to generalize well to different domains, such as different patient populations or medical institutions (Banu and Amirtharajan 2020; Na and Park 2021; Li et al. 2020). In medical systems, domain shifts can occur due to various factors, such as changes in patient demographics, medical practices, or data collection methods. Di et al. (2021) discussed systems that are robust to domain shifts and which can, therefore, improve the generalizability and applicability of diagnosis systems.

Domain robustness focuses on the ability of a system to perform consistently across different domains or under distribution shifts. It involves statistical modeling, such as domain adaptation or transfer learning, to account for changes in distribution. Domain Robustness

refers to the ability of a system or model to perform consistently and accurately across different domains or with distribution shifts. We can mathematically express this concept by considering a set of domains or distributions, denoted as $D = \{D_1, D_2, \ldots, D_N\}$, where each domain D_i represents a distinct data distribution in the medical field. Suppose we have a mathematical model M that aims to solve a specific medical task, such as disease diagnosis or medical image analysis. The performance of an M model can be evaluated using performance metrics such as accuracy, precision, or recall. To achieve Domain Robustness, model M should exhibit consistent and reliable performance across all domains in D. In other words, for any given domain Di, the model performance should remain stable and accurate. Mathematically, we can express this as:

$$P(M, Di) \approx P(M, Dj), forall i, j \in \{1, 2, \ldots, N\},$$

where $P(M, Di)$ represents the performance of a model M in the D_i domain, and ≈ denotes approximate equality. This equation indicates that the model performance should be similar across different domains. Several mathematical techniques can be used to ensure Domain Robustness. One common approach is domain adaptation, where the model is trained on a source domain with abundant labeled data and then adapted to a target domain with limited labeled data. This adaptation process aligns the statistical properties and characteristics of the source and target domains, reducing the domain shift and enhancing the model's performance on the target domain.Several researchers have aimed to explain the concept of domain robustness within the medical field in particular. Li et al. (2020) focused on domain generalization in the pursuit of improving the robustness of medical image classification models across different domains. It addresses the challenge of limited labeled data by leveraging unlabeled data from multiple source domains. However, further investigation is needed into the proposed system's scalability and computational efficiency in handling large-scale medical datasets. Gadepally et al. (2022) explored the challenges involved in generalization in medical imaging, including domain shift and dataset bias. They discussed various techniques for improving generalization performance and provided insight into the importance of dataset diversity.

In another study, Gadepally et al. (2022) addressed some challenges involved in unsupervised domain adaptation by proposing a novel approach that leverages causal structure to enhance model selection. The authors recognized that existing methods for unsupervised domain adaptation often struggle to achieve satisfactory performance due to the lack of labeled target domain data. To overcome this limitation, the study introduced a framework that incorporates causal reasoning to guide the selection of a robust model for adaptation. By analyzing the causal relations between source and target domains, the proposed approach aims to identify and utilize the most relevant features that align with the target domain. However, there is a need for empirical evaluations and comparative studies to more deeply assess the impact of causal reasoning on adaptation performance across various domains and datasets.

Challenges may also arise from the nature of large-scale datasets and complex causal relationships, which pose issues in terms of computational resources and real-time applicability. While exploring strategies to improve scalability and efficiency, investigating adaptive learning mechanisms and self-organization strategies within the multi-agent system would ultimately enhance the system's flexibility and adaptability in dynamic medical envi-

ronments. In deep learning, optimizing hyperparameters is essential for enhancing model robustness, particularly in medical contexts where reliability is crucial. Methods like Bayesian optimization provide a more efficient way to select hyperparameters, thereby improving model performance and resistance to adversarial attacks. Techniques such as adaptive learning rates and robustness-focused tuning further refine models, thus enabling them to maintain high accuracy even under adverse conditions. These advanced approaches significantly strengthen models against adversarial threats while ensuring adaptability to various data perturbations, meaning they play vital roles in developing reliable AI systems in healthcare.

However, a combination of grid search and Bayesian optimization can be used to fine-tune hyperparameters, with a particular focus on enhancing model robustness against adversarial attacks. In this process, key parameters such as learning rate, regularization strength, and network depth are optimized, as these significantly impact a model's ability to resist adversarial perturbations (Ye et al. 2022). Unlike conventional hyperparameter optimization, this approach integrates adversarial training within the optimization process, while specifically targeting robustness as a performance metric. A comparative analysis demonstrates that this method outperforms other state-of-the-art approaches, such as random search and evolutionary algorithms, in bolstering the model's defense against adversarial attacks. The relationship between the optimized hyperparameters and the observed improvements in robustness can be further elucidated through detailed mathematical discussions and comparative evaluations (Di et al. 2021).

Ultimately, domain robustness is crucial in scenarios where the training and deployment domains differ, which is often the case in real-world applications. Domain robustness ensures generalization and performance stability across diverse environments. Substantial research is needed focused on accurately modeling domain shifts, developing domain-invariant representations, handling unlabeled target domain data, and addressing the bias introduced by the source domain. Domain robustness enhances system resilience when facing domain-specific attacks or changes, ensuring consistent performance and maintaining security across different deployment scenarios.

Conceptual Robustness Conceptual robustness is vital in natural language processing, information retrieval, or knowledge-based systems. It refers to the ability of a medical diagnosis system to handle variations in the definition and interpretation of medical concepts, such as disease symptoms or risk factors. Conceptual robustness ensures the effectiveness of a system despite variations in the expression of concepts or information. In medical diagnosis systems, conceptual variations can arise due to differences in medical terminology, cultural or regional disparities, or the evolving nature of medical knowledge. Developing algorithms that are robust to conceptual variations can improve the accuracy and validity of diagnosis systems (Alnajem et al. 2019). Conceptual robustness deals with maintaining system performance under conceptual or semantic changes in the input data. It involves modeling high-level concepts, semantic spaces, or ontologies to handle variations in input representation. Consider a conceptual model M that takes input data x and produces an output or prediction y. The goal is to characterize the conceptual robustness of the model by ensuring that it maintains its understanding and reasoning abilities across different variations in the input space.

We can mathematically express the conceptual robustness problem as an optimization problem with the objective of finding model parameters θ^* that minimize the discrepancy

or loss between the model's predictions in different domains or input variations. This can be represented as:

$$\theta^* = argmin_\theta \Sigma_D \left(L \left(M \left(x; \theta \right), M \left(x'; \theta \right) \right) \right)$$

where

- θ represents the model parameters.
- D denotes different domains or variations in the input space.
- L is a loss function that measures the discrepancy or difference between the model's predictions in different domains or input variations.

x and x′ are inputs from different domains or variations.Here, the optimization problem aims to find model parameters that minimize the overall discrepancy in the model's predictions across various domains or input variations. The choice of the loss function L depends on the specific problem as well as the desired notion of conceptual robustness.

Several researchers have put substantial effort into explaining the concept of Conceptual Robustness within the medical domain. For example, Shaikh et al. (2021a) highlight the potential of artificial intelligence (AI) in medicine, particularly in clinical decision support (CDS), while emphasizing the importance of AI in managing large volumes of patient data to reach optimal clinical decisions. That study further explores advancements in medical image analysis, such as Radiomics, and how machine learning has improved our understanding of diseases and their management. Another study by Natsiavas et al. (2019) revealed the increasing application of KE in DS, particularly for Adverse Drug Event (ADE) assessment, with potential improvements in Adverse Drug Reaction (ADR) predictions and drug interactions detection. However, most such studies remained at the proof-of-concept stage, signaling a need for more comprehensive research to integrate KE methods into routine DS practices, especially with the emergence of newer data sources like genetic databases and social media.

Sheehan et al. (2013) proposed a conceptual framework for the design of robust clinical decision support systems (CDSS). It emphasized the importance of ensuring conceptual robustness by aligning the system's knowledge representation with clinical practice guidelines and incorporating contextual information for accurate and reliable decision support. Common limitations of these studies include the challenge of moving from theoretical concepts to practical applications and ensuring the reliability of the proposed technologies under diverse real-world conditions. For future implementations, these technologies need to be refined to ensure they are adaptable, accurate, and aligned with continually evolving clinical practices.

Model Robustness Model robustness refers to the ability of a medical diagnosis system to maintain its performance and accuracy under different conditions, such as changes in training data, model parameters, or input data. Model robustness is a crucial aspect of ensuring that a model performs reliably and consistently in real-world scenarios (Lane et al. 2015). It helps mitigate the risks of overfitting, data biases, or model instability, thus leading to more reliable and generalizable models. Model robustness helps mitigate the risks of model manipulation or exploitation by ensuring system security against attacks that seek to undermine model performance. In medical diagnosis systems, model instability can stem

from factors such as overfitting, underfitting, or model drift. The development of algorithms that are robust to model instability can ensure the reliability and consistency of diagnosis systems, as discussed by Madry et al. (2018).

We can express this concept of model robustness, which includes the analysis of model stability, sensitivity, or robust optimization techniques, as an optimization problem, the goal of which is to find model parameters that minimize the expected loss in the distribution of perturbations (Beyer and Sendhoff 2007). This can be represented as:

$$\theta^* = argmin_\theta * E\left[L\left(f\left(x+\delta;\theta\right),y\right)\right]$$

where

- θ represents the model parameters.
- δ represents the perturbation applied to the input x.
- L is a loss function that measures the discrepancy or error between the model's predictions and the ground truth y.

E [·] denotes the expectation over a distribution of perturbations.The optimization problem aims to find model parameters that minimize the expected loss while considering the model's performance under different perturbations. The choice of the perturbation distribution and the loss function depend on the specific problem and the desired notion of model robustness. The perturbation δ can take various forms depending on the context and the type of robustness being considered (Beyer and Sendhoff 2007). For example, it can represent random noise, adversarial perturbations, or variations in the input domain. The model f can be any type of model, such as a neural network, a statistical model, or a decision tree-based model. The mathematical modeling of model robustness focuses on optimizing the model parameters to improve its robustness against perturbations.

Several researchers have attempted to explain the concept of Model Robustness within the medical domain. For example, Joel et al. (2022) investigated the robustness of deep learning models trained on diagnostic images in the field of oncology using adversarial images. The authors of that study recognized that deep learning models are susceptible to adversarial attacks, where imperceptible perturbations are added to input images to deceive the models into producing incorrect predictions. They evaluated the performance of the models on both the original and adversarial images, analyzing any significant changes in their predictions. The results of the study reveal that the deep learning models exhibit vulnerabilities to adversarial attacks. The performance of the models deteriorates significantly when presented with adversarial images, thus leading to misclassifications and erroneous predictions. The authors discuss the implications of these vulnerabilities and emphasize the need to develop robust deep-learning models in oncology to enhance their reliability and clinical applicability. However, future studies in this area should focus on further research and the development of robust techniques to ensure the reliability and trustworthiness of deep learning-based diagnostic systems in oncology.

Wang et al. (2022b) introduced SurvMaximin, a robust federated approach for transporting survival risk prediction models in healthcare. SurvMaximin leverages federated learning and the maximin principle to optimize model transportability while preserving patient privacy and handling variations in data distributions across different institutions.

The obtained experimental results demonstrated that SurvMaximin achieves competitive performance in model transportability and robustness, thus offering a promising solution for the collaborative development and deployment of survival risk prediction models in a manner that preserves privacy.

Cui et al. (2021) presents DEAttack, a differential evolution-based attack method for assessing the robustness of medical image segmentation algorithms. DEAttack generates adversarial examples that can deceive segmentation models by iteratively modifying pixel values in the input images. The proposed method effectively highlights vulnerabilities in segmentation models and provides valuable insights for improving the robustness of medical image segmentation techniques—lastly, Kajić et al. (2012) presented an automated method for segmenting the choroid in eye images using optical coherence tomography (OCT). Utilizing a statistical model, the method accurately differentiated the choroid from other structures in the eye. It showed promising results in accurately segmenting the choroid in both healthy and diseased eyes, and it can therefore potentially aid in the diagnosis of ocular conditions.

Interpretable Robustness Interpretable robustness refers to the ability of a medical diagnosis system to provide interpretable and understandable explanations for its decisions (Chen et al. 2018). In medical systems, interpretability is important for gaining the trust and acceptance of clinicians and patients (Wang et al. 2021). Developing algorithms that are robust to interpretability challenges, such as black box models or complex decision boundaries, can improve the interpretability and transparency of diagnosis systems, as presented in Molnar et al. (2020), Zeiler and Fergus 2014).

We can express the concept of interpretable robustness that prioritizes both accuracy and interpretability as the robustness of the model to perturbations, with the aim of ensuring that small changes in the input data do not lead to significant changes in the model's predictions or explanations. The goal of a robust interpretable ML system is to create models that not only make accurate predictions but also provide explanations for their decisions in a manner that can be understood by humans (Ali et al. 2023).

Consider a binary classification problem where we have a dataset consisting of input–output pairs: $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where x_i represents the input features and y_i represents the corresponding class labels. To build an interpretable and robust model, we can use a combination of techniques, such as linear models, rule-based models, and regularization.

- Linear Models: Linear models provide interpretability by assigning weights to the input features, which indicate their importance in determining the output. We can formulate the linear model as:
- $f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p,$

  where f(x) represents the predicted output; x_1, x_2, …, x_p are the input features; and β_0, β_1, β_2, …, β_p are the corresponding weights. Analyzing the magnitude and sign of the weights makes it possible to understand the influence of each feature on the model's decision.

- Rule-based Models: Rule-based models provide explicit decision rules that map input features to class labels. These rules are often expressed in the form of logical

statements or if–then rules. We can represent a rule-based model as a set of rules: $f(x) = \{Rule^1, Rule^2, \ldots, Rule_{\mathrm{m}}\}$, where each $\mathrm{Rule}_{\mathrm{i}}$ consists of conditions and corresponding class labels. Examining the rules allows us to gain insights into how the model makes decisions based on different combinations of input features.

- Regularization: Regularization techniques can be used to promote sparsity and stability in the model's parameters. L1 regularization (Lasso) encourages some weights to be exactly zero, effectively leading to the selection of a subset of features that are most relevant to the prediction task. This helps simplify the model and makes it more interpretable. L2 regularization (Ridge) can also be used to prevent large weight magnitudes, thus leading to a smoother decision boundary and reducing the model's sensitivity to small changes in the input data.

Several researchers have attempted to explain the concept of Interpretable Robustness in the medical domain. Linardatos et al. (2021) focused on developing methods for explaining and interpreting machine learning models. That study centered on machine learning interpretation methods, and it provided a literature review and taxonomy of these techniques. It also included links to their programming implementations, thus providing a valuable resource for both researchers and practitioners in the complex realm of machine learning interpretability. The authors discussed how AI systems make decisions, particularly in critical fields like healthcare.

Alvarez-Melis and Jaakkola (2018) presented an in-depth detailed analysis of interpretability, which states that, for explanations provided by models to be truly effective, they must consistently correlate with similar inputs. To evaluate interpretability, new criteria for measuring such consistency are presented. The findings suggest that many current interpretability strategies do not work well when assessed using these criteria. Nevertheless, that study offered techniques that can be used to enhance the consistency of prevalent interpretation methods.

The approach taken by the authors highlights a critical gap in the domain of AI interpretability: the need for explanations to be both accurate and stable across similar data points. While their metrics offer a fresh perspective, potential limitations include the universality of these metrics across diverse AI applications and the risk of oversimplifying complex interpretability aspects in the pursuit of robustness. In terms for future implementations, ensuring that explanations are robust can lead to more transparent and reliable AI systems, which is particularly critical in high-stakes areas like medicine or finance. Embedding these metrics into AI development tools could become standard practice that promotes the design of models that are both powerful and comprehensible.

Gilpin et al. (2018) sought to establish best practices and highlight unresolved challenges by defining the core explainability and interpretability concepts and categorizing the existing research in this area. The authors of that study also underscored the imperfect nature of current explanatory methods, particularly for deep neural networks, cumulating in a proposal for potential directions for future XAI research based on our findings.

The authors emphasize the growing importance of XAI in the face of increasingly complex algorithms. Their work underlines the need to standardize and evaluate XAI methodologies. Potential limitations of their study may include the breadth of XAI techniques they cover and the risk of oversimplifying the diversity of explainability challenges across

various domains. That study focused on the insufficiency of explanations for deep neural networks, which suggests a significant gap in current research. In terms of future implementations, their findings could help develop standardized XAI frameworks and evaluation metrics that would play an instrumental role in creating universally understood and trusted AI systems in a variety of sectors ranging from healthcare to finance.

Scalable Robustness Scalable robustness refers to the ability of a medical system to handle large-scale data and computational requirements (Verbraeken et al. 2020; Jayabalan and Jeyanthi 2022). In medical systems, scalability is important for processing and analyzing large amounts of medical data as well as providing timely and accurate diagnoses. Ngiam and Khor (2019) considered that systems that are robust to scalability challenges, such as distributed computing or parallel processing, can improve the scalability and efficiency of diagnosis systems. Scalable robustness focuses on achieving robustness at scale in large-scale systems or datasets. It involves designing algorithms and architectures that can efficiently handle massive amounts of data while maintaining robustness.

- Stochastic Gradient Descent (SGD) Equation: In many large-scale learning scenarios, stochastic gradient descent is employed to incrementally optimize model parameters using mini-batches of data (Oymak 2019). The update equation for SGD can be written as:

$$\theta_{\mathrm{t}}^{+1} = \theta_{\mathrm{t}} - \alpha \nabla L\left(\theta_{\mathrm{t}}, x, y\right),$$

- where θ_t represents the model parameters at time step t, α is the learning rate, and $\nabla L\left(\theta_t, x, y\right)$ is the gradient of the loss function L with respect to θ_t computed on a mini-batch of data (x, y). This equation updates the model parameters by subtracting a gradient fraction from the current parameter values.
- Regularization techniques are used to prevent overfitting and improve the generalizability of models in large-scale systems (Wu et al. 2019b). One common form of regularization is L2 regularization (Ridge), which adds a penalty term to the loss function. The regularization equation can be written as follows:

$$Loss\left(\theta\right) = Loss\left(\theta\right) + \lambda \left|\left|\theta\right|\right|^{22},$$

- where $Loss\left(\theta\right)$ represents the original loss function, θ is the model parameter vector, $\left|\left|\theta\right|\right|^{22}$ is the squared L2 norm of the parameter vector, and λ is the regularization parameter that controls the strength of regularization. This equation penalizes large values of the parameter vector, which encourages the model to distribute the weights more evenly and reduces over-reliance on a few features. These equations provide a glimpse into the mathematical modeling of scalable robustness. However, it is important to note that the specific equations used can vary depending on the algorithms and architectures employed, and that different techniques may require different equations or modifications to suit the specific requirements of robustness in large-scale systems or datasets.

However, several researchers have attempted to clarify the concept of Scalable Robustness. Lestas and Vinnicombe (2005) demonstrated a method that could be used to ensure the stability of consensus protocols with diverse dynamics without central control. This is achieved by each agent performing a self-assessment, which eliminates the need for complete network knowledge. Using examples involving random graphs, the research has revealed that the proposed solutions remain effective in extensive networks, even with dynamic heterogeneity. While the findings apply to symmetric protocols, there are more stringent stability guidelines for non-symmetric interconnections.

That paper introduces a decentralized approach to maintaining the stability of consensus protocols while emphasizing its effectiveness even in the presence of diverse dynamics. A potential limitation is that the primary application of that approach is to symmetric protocols, implying that non-symmetric interconnections may not provide benefits that are as effective, given the more stringent stability conditions. Future implementations could delve deeper into optimizing protocols for non-symmetric interconnections, and further research could enhance the nonlinear extensions within the IQC framework, ultimately expanding the applicability and efficiency of the approach to more complex network structures.

Zhang et al. (2015) introduce a resilient and scalable strategy. A unique segmentation technique is designed to label areas of interest (like cells) precisely, and hierarchical voting is employed in combination with a repulsive active contour mechanism. A hashing-centric retrieval system is also formulated to classify these regions using an extensive training dataset. When tested on the complex task of distinguishing two lung cancer variants using numerous histopathological images from several patients, the study solution records an impressive accuracy of 87.3%, and the search across half a million cells was completed in 1.68 s.

That study presented a novel approach to detailed medical image analysis that particularly excels in differentiating the two major types of lung cancer. However, potential limitations could arise from the generalizability of the method to other disease categorizations or variations in image quality and source. Although this method is highly accurate, it still has much room for improvement. The technique's adaptability to other types of medical imaging and its integration with emerging AI models could be further explored, which will likely enhance its diagnostic accuracy and speed, thus making it more valuable in real-world clinical applications.

Ethical Robustness Ethical robustness refers to a medical system's ability to uphold ethical principles and values, such as fairness, privacy, and autonomy (Antunes et al. 2018; Holzinger et al. 2022). Ethical considerations are particularly important in medical systems to ensure equal and respectful treatment of patients and to protect their rights and interests. Arrieta and Roy discussed the ethical robustness of diagnosis systems in Barredo Arrieta et al. (2020), Roy et al. 2023b), which must be robust to ethical challenges such as bias, discrimination, or unintended consequences to ensure the ethical integrity and accountability of diagnostic systems.

In the context of ethical robustness in medical systems, mathematical modeling can help address ethical challenges such as bias, discrimination, and unintended consequences. While the specific equations used may depend on the techniques and algorithms used, the commonly applied equations in this domain are as follows:

1- Fairness Equation (Fairness Metrics) Fairness indices quantify and measure the fairness of the medical system while ensuring equitable treatment across different demographic

groups. One popular fairness metric is the equalized odds or equal opportunity difference, which measures the difference in true positive rates (TPR) or false positive rates (FPR) among different groups. The fairness equation can be expressed as follows:

$$FairnessMetric = |TPR\,(group1) - TPR\,(group2)|$$

where $TPR\,(group1)$ represents the true positive rate for group 1, $TPR\,(group2)$ represents the true positive rate for group 2, and | | denotes the absolute difference. This equation quantifies the disparity in the system's performance between different groups, thus making it possible to identify and mitigate instances of unfair treatment.*2- Privacy Equation (Differential Privacy)* Differential privacy is a privacy-preserving technique that aims to protect individuals' sensitive information during data processing. It ensures that including or excluding an individual's data does not significantly affect the system's results or conclusions. The privacy equation can be expressed as follows:

$$\Pr\left[Output^1\right] \leq e^{\exp}\,(\epsilon) * \Pr\left[Output^2\right]$$

where $Pr\,[Output_1]$ and $Pr\,[Output_2]$ are the probabilities of obtaining different outputs for two datasets that differ only in terms of one individual's information, and ε is the privacy parameter. This equation forces the difference in probabilities to be bounded by the exponential of the privacy parameter, which provides a quantitative measure of privacy guarantees. It is important to note that these equations represent examples of ethical considerations and mathematical modeling techniques in medical systems specifically. Ethical robustness requires a comprehensive framework that considers multiple ethical dimensions and may include additional equations, methodologies, and techniques to address specific challenges related to fairness, privacy, and autonomy in the medical system.Several researchers have dedicated efforts to explaining the concept of Ethical Robustness. Amugongo et al. (2023) provided some recommendations for moving from generic guidelines to a case-specific ethical framework, with AI mHealth app used as an example. Drawing from AI4People (Automotive Committee), the EU High-Level Expert Group, and Human Rights, we integrate an "ethics by design" approach. These groups emphasized seven principles: fairness, agility, precision, safeguarding humanity, respect, trust and accountability, and reproducibility. Incorporating these principles into standard software development processes makes it possible to address ethical concerns proactively.

However, developers should be trained on ethical guidelines while ensuring a proactive "ethics by design" approach. Feedback from stakeholders should also be obtained regularly to maintain relevance and ensure compliance with ethical considerations. This case-specific approach, while more feasible, may not be universally applicable or exhaustive. Relying on existing frameworks may also overlook emerging ethical challenges that are unique to rapidly evolving AI technologies.

Pitas (2021) introduced some issues that are related to the ethical and regulatory robustness of Autonomous Systems (AS), which also include some drones and Autonomous Vehicles (AV); these face ethical, security and privacy challenges due to their embedded intelligence. Their versatile applications, which range both the civilian and military domains, when combined with potential cybersecurity vulnerabilities, pose a risk of misuse, both accidental and intentional. Current legislation addresses these issues inadequately, particularly in terms of

data collection, privacy, and use. While efforts are being made to establish privacy laws and technical measures to protect the use of AS and data, key concerns remain unsolved, such as the distinction between personal and non-personal data and the "re-identification" of data. Recommendations include improved cybersecurity training for AS developers and comprehensive data governance policies. A limitation is the emerging stage of data management and protection in AVs, which leads to unresolved data privacy issues.

Resilient Robustness Resilient robustness refers to a medical system's ability to recover and adapt to unexpected or disruptive events, such as system failures, data breaches, or natural disasters (Fang and Zio 2019; Moskalenko and Moskalenko 2022). In medical systems, resilience is an important aspect of maintaining the continuity and availability of diagnoses and minimizing the impact of disruptions on patient outcomes. Zamir et al. (2020), Nan and Sansavini 2017) also discussed systems that are robust to resilience challenges, such as fault tolerance, redundancy, or disaster recovery, which can improve the resilience and reliability of diagnosis systems. In the context of resilient robustness in medical systems, mathematical modeling can help address the challenges related to fault tolerance, redundancy, and disaster recovery. The commonly used equations for these purposes are presented below:

1- Redundancy Equation (N-Version Programming) Redundancy techniques such as N-version programming involve running multiple versions of the diagnostic system concurrently and comparing their outputs to detect and mitigate errors or faults. The redundancy equation can be represented as:

$$Output\left(system\right) = MajorityVote\left(\begin{array}{c} Output\left(version^{1}\right), Output\left(version^{2}\right), \\ \ldots, Output\left(version_{\mathrm{n}}\right) \end{array}\right),$$

where Output(system) represents the final output of the system, and Output(version_1), Output(version_2), …, Output(version_n) are the outputs of different versions of the diagnosis system. The majority vote operation selects the output that has been agreed upon by the majority of versions. This equation helps improve resilience by reducing the likelihood of misdiagnoses and enhancing error tolerance.*2- Mean Time to Recovery (MTTR) Equation* The Mean Time to Recovery (MTTR) is a metric that quantifies the average time required to recover from a system failure or disruption. It represents the system's ability to restore functionality and resume normal operations. The MTTR equation can be expressed as follows:

$$MTTR = \sum \frac{\left(Downtime^{1} + Downtime^{2} + \cdots + Downtime_{\mathrm{n}}\right)}{N}$$

where Downtime_1, Downtime_2, …, Downtime_n represent the durations of individual system downtime events, and N is the total number of downtime events considered. This equation calculates the average downtime duration over multiple incidents, with the calculated value providing a measure of the system's recovery efficiency and resilience. These equations represent examples of mathematical modeling techniques that are used in resilient robustness for medical systems. Resilient robustness requires a comprehensive approach that considers the system architecture, fault tolerance mechanisms, and disaster recovery strategies. The specific equations and methodologies employed may vary depending on the nature of the disruptions and the system's design goals. These four concepts are vital for understanding agriculture's adaptability, and the distinctions among these concepts lie in the system com-

ponents and perturbation type. Therefore, it has some limitations like potential variability in assessment methods under different scenarios. Recommendations to adapt to unpredictable changes include enhancing diversity and adaptive capacity in agriculture. Several researchers have attempted to explain the concept of resilient robustness. For example, Urruty et al. (2016) addressed some agricultural systems issues with a focus on the resiliency of the systems that face increasing uncertainty due to global warming and price volatility, thus necessitating sustainability under both average and extreme conditions. This review distinguishes four overlapping concepts: stability, robustness, vulnerability, and resilience, that can be used to evaluate agriculture's response to perturbations. Zhu and Başar (2015) discussed some critical infrastructure elements, like power grids and transportation elements, that are increasingly reliant on open networks, thus introducing challenges that may affect control systems that focus on resilient Robustness issues. Traditional control system designs are focused on physical disturbances and modeling uncertainties, and integrating them with modern information technologies exposes them to certain vulnerabilities stemming from cyber components. Software vulnerabilities in these systems open avenues for potential threats and attacks, as evidenced by the Stuxnet worm targeting Siemens SCADA systems. Therefore, this study has limitations for current control system designs that might not account for all potential cyber vulnerabilities, thus requiring constant vigilance and updates. However, it is necessary to enhance cybersecurity measures and regularly update software to combat evolving cyber threats.*Transfer Robustness* Transfer robustness refers to the ability of a medical system to transfer knowledge and skills acquired from one domain to another, such as from one disease or modality to another (Abbas 2022; Asif et al. 2022). In medical diagnosis systems, transfer learning is an important aspect for leveraging similarities and differences between different medical conditions and treatments as well as reducing the need for large amounts of labeled data**.** Abbas discussed in detail the systems used in the above study in Abbas (2022), which are robust to transferring learning challenges, such as domain adaptation, fine-tuning, or multi-task learning, and which can thus improve the transferability and efficiency of diagnosis systems.

In the context of transfer robustness in medical diagnostic systems, mathematical modeling can help address the challenges of domain adaptation, fine-tuning, and multi-task learning. The commonly used equations in this domain are as follows:

1- Domain Adaptation Equation (Adversarial Adaptation) Domain adaptation techniques aim to transfer knowledge and models from a source domain (where labeled data is abundant) to a target domain (where there is little labeled data). Adversarial adaptation is one such technique that aligns the feature distributions between the source and target domains. The domain adaptation equation is given by:

$$Loss = Loss\,(source) + \lambda * AdversarialLoss\,(target)$$

where Loss(source) is the loss on the source domain, Adversarial Loss(target) is the adversarial loss on the target domain, and λ is a regularization parameter that controls the balance between the two losses. This equation combines the loss on the source domain with the adversarial loss on the target domain, so the model is encouraged to learn domain-invariant representations and improve transferability.*2- Multi-Task Learning Equation* Multi-task learning is a technique where a model is trained to perform multiple related tasks simul-

taneously. This can be beneficial in medical diagnosis systems where different diseases or modalities share common features. The multi-task learning equation can be expressed as:

$$Loss = Loss\left(task^{1}\right) + Loss\left(task^{2}\right) + \ldots + Loss\left(task_{\mathrm{n}}\right)$$

where $Loss\left(task_{1}\right), Loss\left(task_{2}\right), ..., Loss\left(task_{\mathrm{n}}\right)$ are the losses associated with individual tasks, and n is the total number of tasks. This equation combines losses from multiple tasks and jointly optimizes model parameters, thus allowing the model to leverage shared knowledge and ultimately improving the performance of different tasks. Transfer robustness requires careful consideration of specific domains, tasks, and available data. The choice of equations and methodology may vary depending on the transfer learning challenges and the specific objectives of the diagnosis system.Some scholars have elucidated the idea of transfer robustness. For example, Bhardwaj et al. (2021) introduced a transfer learning-based CNN system for automatically detecting DR lesions and assessing their severity levels. Leveraging a pre-trained architecture, that paper proposes two models: the Image Feature-based Transfer Learning (IFTL) model, which extracts image features from CNN's fully connected layers, and the Prominent Feature-based Transfer Learning (PFTL) model, which employs statistical methods aiming to filter out irrelevant features. By combining CNNs with an SVM classifier, the study achieved significant performance even when used with smaller datasets, ultimately reducing computational complexity. The approach was then validated using the MESSIDOR dataset for retinal images, and its generalizability was confirmed using the IDRiD dataset. The results of comparative tests show that the method surpasses existing solutions. It should be noted that that study has some limitations like system performance, which may vary when used with different datasets or imaging techniques. However, it is recommended that future research focus on improving feature extraction techniques to achieve even greater accuracy.

Medical diagnosis datasets consist of specialized medical imaging like MRI and X-rays, requiring expert annotations and strict privacy controls due to their sensitive nature(Johann et al. 2023). These datasets are typically smaller and less accessible, with varying quality that depends on medical equipment and patient factors. The tasks associated with medical data are complex and directly impact clinical decisions, requiring models to meet rigorous regulatory standards for accuracy and reliability (Blagec et al. 2022). However, common computer vision datasets include natural images with simpler, often crowdsourced annotations. They are large, publicly available, and involve fewer privacy concerns. The data quality is more consistent, and tasks range from simple object recognition to complex scene understanding, generally without direct clinical implications. Validation standards are less stringent, focusing on usability rather than clinical compliance, reflecting the broader and less specialized nature of these datasets. Table 2 summarizes the main differences between medical diagnosis and common vision task datasets (Johann et al. 2023; Albahri et al. 2023; Drenkow et al. 2021).

In the field of medical diagnostics, the robustness of deep learning models is heavily influenced by the datasets used during training, and the deployment of these widely used datasets in medical diagnoses is essential for understanding the specific challenges and opportunities within this domain. Notable datasets such as MIMIC-III, ChestX-ray14, and the UK Biobank have established themselves as benchmarks for developing and evaluating deep learning models in medical diagnostics. These datasets often contain patient informa-

Table 2 Differences between datasets used in medical diagnoses and common computer vision tasks

Aspect	Medical diagnoses	Common computer vision tasks
Data format	Specialized medical imaging (MRI, CT, X-rays, etc.), often grayscale, 3D, or multi-channel with detailed anatomical structures	Natural images (photographs of objects, scenes), typically RGB with clear, well-defined objects
Annotations	Highly specialized annotations by domain experts, including segmentation masks or pixel-level annotations	Simpler annotations, often crowdsourced, such as bounding boxes or segmentation without needing expert knowledge
Size and Availability	Smaller datasets due to acquisition challenges, privacy concerns, and restricted access	Larger, publicly available datasets with open access and less restriction
Data Privacy & Ethical	Strict privacy laws, ethical standards, and controlled access are essential due to sensitive patient information	Fewer privacy concerns, mainly avoiding copyright and identifiable personal data exposure
Data Quality	Quality varies with imaging modality, equipment, and patient factors; preprocessing is often needed	Consistent quality; images collected with standardized settings and more varied in content
Task Complexity	Tasks involve complex, clinically relevant problems impacting patient care, requiring high accuracy	Tasks range from simple (object recognition) to complex (scene understanding), generally less critical
Regulatory and Standards	Models must meet strict validation and regulatory standards (e.g., FDA approval), emphasizing interpretability	Less stringent validation; focuses on benchmark scores and usability rather than clinical standards
Examples	ADNI, MIMIC, PadChest, Retinal OCT Images (Johann et al. 2023)	ImageNet, COCO, CIFAR, ADE20K (Blagec et al. 2022)

tion that is heterogeneous, noisy, and limited in volume compared to the expansive datasets that are typically utilized in general computer vision tasks like ImageNet. The variability in data quality, which is coupled with the inherent complexity of medical images, necessitates meticulous preprocessing and model design to mitigate biases that could compromise diagnostic accuracy and the robustness of AI systems in medical applications (Vaishnavi et al. 2022).

The differences between medical and general computer vision datasets also present significant security concerns (Kaviani et al. 2022). Medical datasets are inherently more sensitive and frequently contain fewer samples, thus making models more vulnerable to adversarial attacks. Unlike general vision tasks, where data is abundant, the limited avail-

Table 3 Evaluation and impact of robustness techniques in medical diagnostic systems across various studies

References	Robustness techniques	Importance of robustness	Application	Datasets used	Models used	Evaluation metrics for accuracy	Performance	Results
Lundqvist and Fabricio Oliveira (2023)	Adversarial	Assessing model vulnerability and ensuring reliable diagnostic outcomes	Image classification	Diabetic retinopathy dataset	DNN, imageNet pretrained ResNet-18	Adversarial accuracy, robust accuracy	Increased model complexity	The aRUB model was competitive in terms of both robustness and training times
Xing et al. (2023)	Noise	Handling noise and uncertainty in annotation for accurate segmentation	Neuroendocrine tumors (NETs)	Ga-DOTATATE PET image dataset	U-Net-like neural network	Accuracy under noisy input	Reduced classification time	Better lesion detection performances
Shaikh et al. (2021b)	Conceptual Robustness	Ensuring accurate and reliable decision support by addressing conceptual robustness	Medical image analysis of Radiomics	Various medical imaging datasets	CNN, ANN, DLNN models	Conceptual accuracy	Enhanced interpretability	Detection of concept drift
Chen (2021)	Domain Robustness	Abnormality segmentation (e.g., lesion segmentation, tumor)	Image classification and MR atrial segmentation shifts of short-axis MR images	The UK Biobank (UKBB)	MR segmentation frameworks based on CNN	Domain-specific accuracy	Improved training time	Enhance cardiac image segmentation across several unseen domains
Rueckert and Schnabel (2020)	Model Robustness	Assessing the vulnerability of medical ML models and improving their resilience	Image classification, Detection of model vulnerabilities	Various medical imaging datasets	CNN	Model accuracy, performance on various architectures	Efficient resource utilization	Detection of model vulnerabilities
Khakzar et al. (2019)	Interpretable Robustness, Adversarial Robustness	Feature interpretability evaluating the weak supervised localization	Pneumonia detection on NIH ChestX-ray14	NIH ChestX-ray14	CNN	Interpretability metrics (e.g., feature importance)	Reduced model complexity	Feature interpretability both quantitatively and visually

Table 3 (continued)

References	Robustness techniques	Importance of robustness	Application	Datasets used	Models used	Evaluation metrics for accuracy	Performance	Results
Xu et al. (2021b)	Scalable Robustness, Adversarial Robustness, Noise Robustness	Ensuring reliable and robust classification of deep diagnostic models against adversarial attacks	Scalability of security defenses	A new dataset (called Robust-Benchmark)	Deep diagnostic models such as Misclassification-Aware Adversarial Training (MAAdvT)	Scalability of accuracy measures	Reduced memory footprint	Scalability of security defenses
Coutellec (2020)	Ethical Robustness	Enhancing the robustness of nodule classification models for accurate detection	Detection of ethical issues or biases	General Study	General Study	Fairness metrics (e.g., demographic parity)	Reduced inference time	Detection of ethical issues or biases
Lakshminarayana et al. (2019)	Resilient Robustness	Ensuring reliable detection and resilience of a linear cyber-physical control system (CPCS) in power grids	Detection and mitigation of attacks	IEEE-39 bus system and IEEE-118 bus system	Markov decision process (MDP) framework	Overall accuracy, robustness under various attacks	Increased system robustness	To lower the impact of the attack detection errors
Tang et al. (2020)	Transfer Robustness, Adversarial Robustness	To improve the robustness of GNNs against the poisoning attacks	Resilience to distribution shifts and biases	Datasets from Yelp Review	GNN with penalized aggregation mechanism (PA-GNN)	Transfer accuracy	Improved adaptability and transferability	To enhance PA-GNN for the poisoned graph
Amini et al. (2021)	Noise Robustness	Ensuring reliable classification of Alzheimer's disease for accurate diagnosis	Alzheimer's disease classification, Increased resilience to noise	fMRI images from the ADNI dataset	KNN, LDA, RF, SVM, DT, and CNN	Accuracy under noisy conditions	Reduced false positives	Increased resilience to noise

Table 3 (continued)

References	Robustness techniques	Importance of robustness	Application	Datasets used	Models used	Evaluation metrics for accuracy	Performance	Results
Mok and Chung (2019)	Adversarial Robustness	Enhancing the robustness of brain tumor segmentation models for accurate tumor delineation	Brain tumor segmentation Improved defense against adversarial attacks	BRATS15 Challenge dataset	CNN	Adversarial accuracy, robust accuracy	Increased inference time	Improved defense against adversarial attacks
Chougrad et al. 2020)	Domain Robustness	Ensuring reliable classification of breast cancer for accurate diagnosis	Breast cancer classification, Enhanced generalization to new domains	CBIS-DDSM, INbreast	CNN	Domain-specific accuracy	Improved training convergence	Enhanced generalization to new domains
Chang and Ward (1995)	Conceptual Robustness	To define and discuss how to create the least possible modular designs	Improved detection of concept drift	General Study	General Study	Conceptual accuracy	Reduced model complexity	Defining modular design embodies conceptual complexity
Joel et al. (2023)	Model Robustness	Ensuring reliable classification of cancer against adversarial attacks	Cancer classification Enhanced detection of model vulnerabilities	LIDC-IDRI	VGG16 classification models	Model accuracy, performance on different architectures	Reduced memory consumption	Enhanced detection of model vulnerabilities
Ghosh et al. (2023)	Interpretable Robustness	Ensuring reliable classification of retinal diseases for Diabetic Retinopathy Detection	Diabetic retinopathy (DR) detection, Enhanced interpretability and fairness	DiaretDB1, IDRiD	CNN	Interpretability metrics (e.g., feature importance)	Improved prediction time	Enhanced interpretability and fairness

Table 3 (continued)

References	Robustness techniques	Importance of robustness	Application	Datasets used	Models used	Evaluation metrics for accuracy	Performance	Results
Zhong et al. (2019)	Scalable Robustness	Enhancing the robustness of prostate cancer classification models for accurate detection	Prostate cancer classification, Scalable deployment of security mechanisms	PANDA, PCam	CNN	Scalability of accuracy measures	Reduced computational overhead	Scalable deployment of security mechanisms
Maron et al. (2021)	Ethical Robustness	Enhancing the robustness of skin disease classification models for accurate diagnosis	Unique skin disease identification, Mitigation of fairness and ethical issues	Skin Archive Munich (SAM), and SAM-perturbed (SAM-P)	CNN	Fairness metrics (e.g., disparate impact)	Reduced bias in predictions	Mitigation of fairness and ethical issues and advancing stronger classifiers for skin cancer detection
Qiu et al. (2022)	Resilient Robustness	Ensuring reliable detection of cardiac abnormalities by improving the robustness of DL models	Cardiac abnormality detection, Improved detection and recovery from attacks	Various cardiac imaging datasets	CNN	Overall accuracy under attack scenarios	Increased system robustness	To improve deep learning performance in the ECG classification

ability of medical data may make it difficult to sufficiently capture the full variability of conditions, thereby leading to overfitting and heightened susceptibility to adversarial perturbations. Moreover, the specialized nature of medical data means that adversarial attacks could exploit unique characteristics of medical images, potentially resulting in harmful misdiagnoses. These factors underscore the need to develop robust and secure AI models that are specifically designed to address the complexities and security challenges posed by medical datasets (Krizhevsky et al. 2017). The unique characteristics of medical datasets also introduce specific security threats that directly impact the robustness of diagnostic models(Wang and Wang 2023). For example, the limited variability in medical datasets can make models more vulnerable to adversarial attacks, as small, imperceptible changes to the input data can lead to incorrect diagnoses (Silva and Najafirad 2020). Such vulnerabilities undermine the model's robustness, as the system fails to maintain high performance under adversarial conditions (Krizhevsky et al. 2017). Table 3 provides an overview of evaluations of the impact of robustness techniques in medical diagnostic systems, as has been documented across various studies. The table highlights the effectiveness of these techniques on different medical datasets, such as imaging and patient records, ultimately emphasizing their role in improving diagnostic accuracy and system reliability.

3.8 ML for robust diagnoses: applications

ML for Robust Diagnoses is a rapidly growing field that has the potential to revolutionize the way medical diagnoses are made. To elaborate, Machine learning (ML) algorithms can be used to process vast amounts of data from medical imaging, genetic sequencing, and electronic health records to help diagnose complex diseases (Alves et al. 2021). The robustness of ML models is crucial to their effectiveness in medical diagnoses. In this section, we will discuss how unsupervised learning, supervised learning, semi-supervised learning, and reinforcement learning can be used to improve the robustness of ML models in medical diagnosis systems (Abbas 2022).

1. *Unsupervised Learning* Unsupervised learning is a type of ML that involves training models on data that has no pre-existing labels or categories. Instead, the algorithm identifies patterns and relationships in the data to ultimately create its own classifications (Casolla et al. 2020). Unsupervised learning can be used to identify groups of patients with similar medical conditions or to detect anomalies in medical images. By identifying patterns and outliers in the data, unsupervised learning can help improve the robustness of medical systems (Shen et al. 2021).
2 *Supervised Learning* This is a type of ML that involves training models on labeled data to predict new, unlabeled data (Muhammad and Yan 2015). In medical systems, supervised learning can be used to predict a patients' likelihood of developing a certain medical condition based on their symptoms or medical history. By accurately classifying patients into different medical conditions, supervised learning can improve the robustness of medical systems (Żurański et al. 2021).
3 *Semi-Supervised Learning* Semi-supervised learning is a type of ML that involves training models on a combination of labeled and unlabeled data. This approach is particularly useful in medical systems where labeled data may be limited or difficult to obtain.

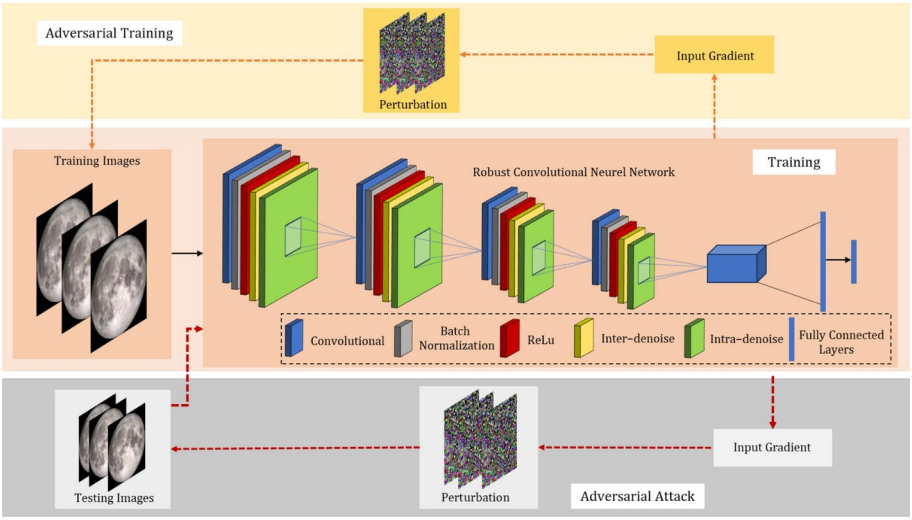

Fig. 7 A comparative look at robust CNNs: adversarial applications in medical image analysis

By using both labeled and unlabeled data, semi-supervised learning can improve the accuracy and robustness of ML models in medical systems (Reddy et al. 2018).

4 *Reinforcement Learning* Reinforcement learning is a type of ML that involves training models to make decisions based on rewards and punishments. In medical systems, reinforcement learning can be used to recommend treatment options based on patient outcomes. By learning from patient outcomes, reinforcement learning can help improve the robustness of medical systems (Kaelbling et al. 1996).

Figure 7 provides a visual introduction to the application of machine learning algorithms in improving the robustness of medical systems through adversarial training. The figure illustrates the process by which a convolutional neural network (CNN), a class of deep learning algorithms that is particularly adept at processing visual data, is trained to withstand adversarial examples. In the first phase, which is labeled as "Adversarial Training," training images are processed by adding perturbations—intentional distortions that are created by manipulating the input data based on the gradient of the loss with respect to the input image (input gradient). These perturbations are typically small and imperceptible, but they are designed to mislead the AI model. By training CNN on both original and perturbed images, the model learns to recognize and resist these adversarial patterns, thus improving its reliability.

The robust CNN architecture is detailed and includes layers such as convolutional layers for feature extraction, batch normalization layers for learning stabilization, ReLU (Rectified Linear Unit) for non-linear activations, and denoise both inter-layers and intra-layers, which serve to reduce the noise in the data. The final part of the network consists of fully connected layers that make the ultimate classification decision. During the testing phase, new images (testing images) are subjected to adversarial attacks—similar perturbations are used to evaluate the network's performance under adversarial conditions (Pansota et al. 2021). The application of ML algorithms in medical systems has great potential to improve

the accuracy and efficiency of diagnoses. However, to ensure the effectiveness of these models, their robustness must be carefully considered. By utilizing different types of ML approaches, such as unsupervised learning, supervised learning, semi-supervised learning, and reinforcement learning, we can improve the reliability of medical systems and ultimately achieve improved patient outcomes.

The integration of machine learning algorithms with medical imaging techniques has revolutionized the field by automating aspects of image analysis and providing radiologists with invaluable support in their diagnostic interpretations (Roland et al. 2022). By leveraging the power of machine learning, healthcare professionals can process vast amounts of imaging data to detect abnormalities with greater accuracy and efficiency. Machine learning models can learn from labeled datasets and identify complex patterns and subtle features in medical images that may be difficult to detect through human observation alone (Cen et al. 2022).

This synergy between machine learning and medical imaging holds immense potential to expedite the identification of critical conditions, thus enabling prompt interventions and ultimately improving patient outcomes. By providing the ability to assist in disease identification, predict disease progression, and guide personalized treatment plans, machine learning in medical image diagnostics has become an invaluable tool in modern healthcare practices.

ML algorithms can also play an essential role in genomics research, where they can help analyze genetic sequencing data to identify potential disease markers and develop targeted therapies (Javed et al. 2023). By examining massive sets of genomic data, machine learning models can discover intricate patterns and relationships that may not be easily discernible through human analysis alone. This capability opens new possibilities in precision medicine, where treatment can be tailored to an individual's genetic makeup, thus optimizing efficacy and minimizing adverse effects. Incorporating machine learning into electronic health records (EHRs) is also a very promising approach. ML models can leverage the wealth of information contained in EHRs to identify patterns, risk factors, and correlations across diverse patient populations.

However, applying machine learning techniques to the diagnosis and treatment of diseases has the potential to revolutionize healthcare practices (Liu et al. 2022). Advancements in machine learning algorithms have revolutionized medical systems, thus offering unprecedented precision and dependability. Adversarial training, which is a technique in which models are exposed to manipulated inputs, equips these systems with the ability to resist subtle perturbations that could lead to diagnostic errors.

The robust CNN's ability to correctly classify these manipulated images, despite the presence of adversarial perturbations, indicates its robustness and reliability for medical diagnostic purposes. A few of the numerous ways that the machine learning improves the diagnosis are mentioned below:

1. *Medical Imaging Analysis* Machine learning algorithms can learn from large datasets of annotated medical images to accurately detect and classify abnormalities. Deep learning models, such as convolutional neural networks (CNNs), can analyze image features and patterns to help radiologists make more precise diagnoses (ur Rehman et al. 2018; Niyirora et al. 2022). Transfer learning techniques enable pre-trained models

to be adapted to specific medical imaging tasks, thus reducing the need for extensive labeled data (ur Rehman et al. 2019).

2. *Disease Prediction and Progression* Machine learning models can analyze diverse patient data, including demographic, genetic, and lifestyle factors, to predict the risk of developing specific diseases (Anter and Abualigah 2023). Longitudinal data analysis and predictive modeling techniques can help identify the patterns of disease progression, thus facilitating more timely interventions. Ensemble models and interpretable machine learning approaches can increase prediction accuracy and provide insight into the factors influencing disease progression (Koçak et al. 2023).
3. *Personalized Treatment Approaches* By integrating genetic data, clinical records, and treatment outcomes, machine learning algorithms can identify the biomarkers that are associated with drug response and tailor treatment plans accordingly. Reinforcement learning techniques can optimize treatment strategies by learning from patient feedback and adapting treatment plans over time. Collaborative filtering and recommendation systems can aid in selecting optimal drug combinations based on patient characteristics and previous treatment responses (Abd-Ellah et al. 2023).
4. *Genomic Analysis* Machine learning algorithms can analyze large-scale genomic datasets to identify disease-associated genetic variants, detect gene expression patterns, and predict disease risk. Feature selection techniques help identify the most relevant genetic features for disease diagnosis and treatment. Deep learning models can unravel complex genetic interactions and discover new disease mechanisms (Elseddik et al. 2023).
5. *Decision Support Systems* Machine learning algorithms can analyze patient data, clinical guidelines, and medical literature to provide evidence-based recommendations to healthcare professionals. Natural language processing (NLP) techniques make it possible to extract valuable information from unstructured clinical text, such as electronic health records and medical literature. Explainable AI techniques help generate transparent and interpretable decision support systems, which provides insight into the reasoning behind recommendations (Ahmad et al. 2023b).
6. *Electronic Health Records (EHR) Analysis* Machine learning algorithms can mine large-scale EHR data to identify patterns, risk factors, and correlations that are associated with specific diseases. Clustering and anomaly detection algorithms help identify patient subgroups and detect unusual patterns in EHR data. Time-series analysis techniques can identify patterns of disease progression and predict patient outcomes (Albayati et al. 2023).
7. *Drug Discovery and Development* Machine learning models can assist in the virtual screening of large chemical libraries to identify potential drug candidates with desired properties (Sugimoto et al. 2023). Generative models and reinforcement learning can aid in de novo drug design and optimization. Predictive models can assess drug safety profiles and identify potential adverse effects during the early stages of drug development (Wang et al. 2022c).
8. *Precision in Oncology* Machine learning algorithms can integrate genomic data, clinical records, and treatment outcomes to predict optimal treatment options for cancer patients (El-Ghany et al. 2023). Survival analysis techniques can be used to select patient subgroups with different prognoses and guide personalized treatment decisions. Radiomics, which combines medical imaging and machine learning, can extract

quantitative features from images to predict tumor behavior and response to treatment (Bordoloi et al. 2023).

9. *Remote Patient Monitoring* Machine learning algorithms can analyze sensor data from wearable devices to remotely monitor vital signs, activity levels, and disease-specific indicators. Anomaly detection algorithms can flag deviations from normal physiological patterns, thus alerting healthcare providers to potential health risks. Predictive models can anticipate deterioration in a patient's condition, thus allowing for timely interventions and reduced hospital readmissions (Liu et al. 2023).
10. *Medical Research and Clinical Trials* Machine learning models can analyze large-scale clinical trial data to identify treatment efficacy, detect adverse effects, and identify subgroups of patients who respond best to specific interventions. Feature engineering and selection techniques can extract important predictors from clinical trial data, thus increasing the accuracy of treatment outcome predictions. Machine learning algorithms can assist in patient selection for clinical trials based on specific eligibility criteria, ultimately achieving a more targeted and efficient recruitment process (Shim et al. 2023; Yadav et al. 2022).
11. *Early Detection and Diagnosis* Machine learning models allow for the early detection of diseases by identifying subtle patterns or biomarkers in patient data. Fusions of multi-modal data, such as genetic, imaging, and clinical data, help achieve more accurate and comprehensive diagnoses. Ensemble learning techniques combine the predictions of multiple models to improve diagnostic accuracy.
12. *Prognostic Modeling* Machine learning algorithms predict disease prognosis and estimate patient outcomes based on various clinical and genetic factors. Longitudinal data analysis techniques can model disease progression over time and predict future outcomes. Survival analysis methods estimate survival probabilities and identify prognostic factors that influence patient outcomes (Elseddik et al. 2023).
13. *Treatment Response Prediction* Machine learning models are used to predict individual patient responses to specific treatments or interventions. Real-time patient monitoring

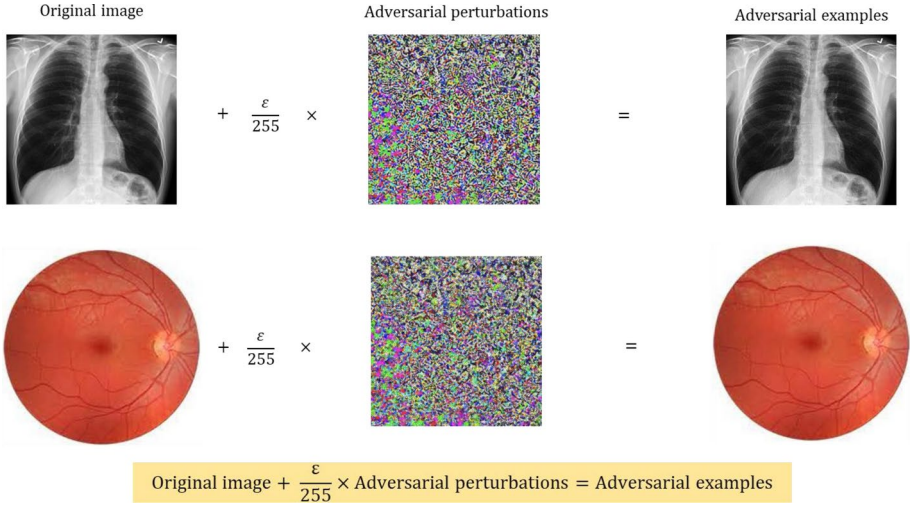

Fig. 8 Adversarial perturbations in medical diagnostics: a visual explanatio

data is integrated with machine learning algorithms to personalize treatment decisions and optimize response rates. Interpretable machine learning methods are used to identify key factors influencing treatment response and thereby guide treatment selection (Anooj et al. 2023; Mahoto et al. 2023).

14. *Clinical Risk Assessment* Risk prediction models are developed using machine learning to identify patients who are at high risk of developing specific diseases or complications. Electronic health records, genetic data, and lifestyle factors are included to assess individual risk profiles. Machine learning can also identify risk factors and optimize risk stratification strategies in clinical practice (Zhang et al. 2023b). However, Fig. 8 depicts the process of generating adversarial examples from original medical images, which is a key concept for studying the robustness of AI in medical diagnosis. Generating adversarial examples, as depicted in the diagram, are essentially an application of machine learning that is intended to test and enhance the precision and reliability of diagnostic models (Dai et al. 2023). In machine learning, particularly in the context of medical diagnostics, it is crucial that models are not only accurate but also resilient to various types of data perturbations. Adversarial examples serve as a tool that can be used to challenge and evaluate the robustness of these models (Priya and Dinesh 2023).

By intentionally adding noise or perturbations to medical images and observing whether the AI model can still correctly diagnose the condition, researchers can identify vulnerabilities in a model. This process can be used to refine models, thus making them more resistant to potential real-world variations and adversarial attacks. Therefore, this practice is considered to be vital to the development of machine learning applications that maintain high precision in medical diagnoses under a wide range of conditions (Bhandari et al. 2023).

The original chest X-ray image is shown in the first row. Adversarial perturbations (visual noise that is often imperceptible to the human eye) are added to this image. These perturbations are calculated by an algorithm and are designed to be small (scaled by a factor of epsilon above 255, where epsilon is a small value) but sufficiently effective to mislead an AI system. The result is an adversarial example, which is an image that looks almost identical to the original to a human but contains carefully crafted distortions that can cause the image to be misclassified by an AI model (Tsai et al. 2023). The second row follows the same process with a fundus photograph, which is used to diagnose conditions that are related to the retina. Again, adversarial perturbations are added to the original image to obtain an adversarial example. The purpose of creating such examples is to test and improve the AI systems' resistance to such subtle attacks that can potentially lead to incorrect diagnoses (Muoka et al. 2023).

In each of these applications, machine learning techniques provide the basis for harnessing the power of extensive and complex data sets. They efficiently navigate through multiple layers of information, whether the information consists of clinical trials, imaging genomics, or electronic health records, to discern patterns that may elude traditional methodologies. Their prowess extends beyond mere diagnosis; they empower medical professionals by predicting treatment outcomes, facilitating early disease detection, and personalizing medical interventions based on individual patient data. This is reinforced by combining multi-modal data sources, from genetics to medical imaging, thus ensuring a comprehensive analysis that leads to more accurate and timely medical decisions.

Table 4 Comparative analysis of machine learning innovations in healthcare delivery

Application	Data used	ML techniques	Purpose	Benefit for healthcare	Challenges	Potential impact	Required expertise
Medical Imaging Analysis	Annotated medical images	CNNs, Transfer Learning	Detect and classify abnormalities	Assists in precise diagnoses	Requires extensive labeled data	Improved diagnostic accuracy	High (Data Science, Radiology)
Disease Prediction and Progression	Patient demographics, genetic, lifestyle factors	Predictive Modeling, Ensemble Models	Predict disease risk, identify progression patterns	Timely interventions, improved outcomes	Interpretability of models	Preventative healthcare strategies	Moderate to High (Epidemiology, Data Science)
Personalized Treatment Approaches	Genetic data, clinical records, treatment outcomes	Reinforcement Learning, CollaborativeFiltering	Tailor treatment plans, identify biomarkers	Personalized patient care	Integrating diverse data types	Enhanced treatment efficacy	High (Genetics, Data Science)
Genomic Analysis	Genomic datasets	Feature Selection, Deep Learning	Identify genetic variants, predict disease risk	Novel disease mechanisms discovery	Handling large-scale data	Advances in genomics and personalized medicine	High (Genetics, Bioinformatics)
Decision Support Systems	Patient data, clinical guidelines, medical literature	NLP, Explainable AI	Provide evidence-based recommendations	Enhance clinical decision-making	Requires high-quality data sources	More informed clinical decisions	Moderate to High (Health Informatics, AI)
Electronic Health Records (EHR) Analysis	EHR data	Clustering, Time-Series Analysis	Identify patterns, predict patient outcomes	Insight into disease correlations	Privacy and security concerns	Better population health management	Moderate (Health Informatics, Data Science)
Drug Discovery and Development	Chemical libraries	Generative Models, Predictive Modeling	Screen for drug candidates, predict safety profiles	Faster drug discovery	Computational costs	Accelerated drug market entry	High (Pharmacology, Data Science)

Table 4 (continued)

Application	Data used	ML techniques	Purpose	Benefit for healthcare	Challenges	Potential impact	Required expertise
Precision Oncology	Genomic data, clinical records, treatment outcomes	Survival Analysis, Radiomics	Predict treatment options, understand tumor behavior	Personalized cancer treatment	Data heterogeneity and integration	Improved cancer treatment outcomes	High (Oncology, Data Science)
Remote Patient Monitoring	Sensor data from wearables	Anomaly Detection, Predictive Modeling	Monitor patients remotely, alert health risks	Reduced hospital readmissions	Data interpretation and accuracy	Enhanced chronic disease management	Moderate (Health Informatics, Data Analysis)
Medical Research and Clinical Trials	Clinical trial data	Feature Engineering, Machine Learning Algorithms	Identify treatment efficacy, optimize trials	Improved clinical trial outcomes	Complex trial designs	More efficient clinical research	High (Clinical Research, Data Science)
Early Detection and Diagnosis	Multimodal data (genetic, imaging, etc.)	Ensemble Learning, Fusion Techniques	Early detection of diseases	More comprehensive diagnoses	Requires advanced data analysis	Early interventions and better outcomes	High (Data Science, Clinical Medicine)
Prognostic Modeling	Clinical and genetic factors	Longitudinal Data Analysis, Survival Analysis	Predict prognosis and patient outcomes	More informed clinical decisions	Long-term data collection	Improved life expectancy and quality of life	Moderate to High (Epidemiology, Data Science)
Treatment Response Prediction	Patient monitoring data, treatment data	Interpretable Machine Learning	Predict treatment responses	Optimized treatment efficacy	Dynamic treatment adaptation	Personalized and adaptive therapies	Moderate to High (Clinical Medicine, Data Science)

As the amounts of medical data grow exponentially, these machine learning models also evolve, thus showcasing greater accuracy in addition to the ability to discover new information, such as disease subtypes or drug interactions. The emphasis on explainable AI also ensures that diagnostic and prognostic decisions remain transparent and interpretable, which strengthens trust among both clinicians and patients. Therefore, machine learning not only augments the medicine's capacity for diagnosis and treatment but also ensures that a more holistic and patient-centric approach to healthcare is taken.

Table 4 summarizes the vital roles that machine learning (ML) plays in healthcare, highlighting key applications from image analysis to drug discovery. It compares data types, ML techniques, goals, and benefits while also considering implementation challenges and potential impacts on healthcare, ultimately providing a snapshot of ML's transformative influence on medical diagnostics and patient care. The "Potential Impact" column in this table aims to provide insights into how these applications might change healthcare outcomes, patient care, and medical research in the long term. The "Required Expertise" column indicates the interdisciplinary knowledge necessary to develop and implement machine learning applications, which often require a combination of domain expertise in healthcare and technical knowledge in data science and machine learning.

4 Assessing and improving robustness: tools and techniques

In the emerging field of deep learning, as models become increasingly sophisticated, it is paramount to ensure their robustness against adversarial attacks. Adversarial attacks introduce subtle and often unnoticeable changes to the input data to mislead the model, which makes them a serious problem, particularly in critical applications like medical diagnosis. As summarized in Fig. 9, several packages and libraries have been developed to evaluate and increase the reliability of machine learning models in the face of these challenges:

Fig. 9 Overview of existing packages for AI model robustness and security

4.1 Adversarial robustness toolbox (ART)

The Adversarial Robustness Toolbox is a comprehensive library that provides a wide range of tools and techniques that can be used to analyze and enhance the resilience of machine learning models to adversarial attacks (Woldeyohannes 2021). In particular, it offers various attack methods, defense mechanisms, and evaluation metrics to assess the robustness of models. ART supports multiple deep learning frameworks, including TensorFlow and PyTorch, which makes it universally applicable to a variety of medical systems. ART is widely used and offers a rich set of functionalities. It provides a high level of flexibility in terms of attack and defense methods, thus allowing users to compare different techniques and assess their effectiveness in improving model robustness. ART also has an active community and receives continuous updates, which ensures that new advancements in robustness research are incorporated into the library (Hu 2021).

Nicolae et al. (2018) introduced the Adversarial Robustness Toolbox (ART), which is a complete Python library intended to assist developers and researchers in protecting machine learning models from adversarial threats. ART offers resources for creating and implementing defenses, assessing them through adversarial attacks, and measuring model robustness. The toolbox is compatible with a range of machine learning models, including deep neural networks, decision trees, and support vector machines. ART also accommodates comparisons of model robustness by allowing users to select a machine learning model and a dataset and then assess its baseline accuracy. Adversarial examples are generated using ART-supported attacks, such as the Fast Gradient Sign Method (FGSM) (Ruiz et al. 2022). The model's accuracy in these adversarial examples provides an indication of its vulnerability. Defenses that are available for ART, like adversarial training or feature squeezing, can then be applied. The defended model's performance on new adversarial examples can be evaluated and compared to the baseline defense or other defenses, ultimately providing a quantitative measure of its robustness (Qayyum et al. 2021).

Goodfellow et al. (2015) introduces the concept of adversarial examples and demonstrates their effectiveness in fooling machine learning models. The authors of that study utilized the CleverHans library to generate adversarial examples and evaluate their impact on various neural network architectures. That study discussed different attack methods that are implemented in CleverHans, such as the Basic Iterative Method (BIM) and Fast Gradient Sign Method (FGSM) methods, while presenting strategies to improve model robustness. Meanwhile, Kurakin et al. (2017) explored the vulnerability of machine learning models to adversarial attacks at scale. To this end, they used the CleverHans library to generate adversarial examples and evaluated their impact on multiple classifiers, including deep neural networks. That study investigated the transferability of adversarial examples across different models along with the effectiveness of various defense mechanisms. The use of the CleverHans library allowed the authors to experiment with different attack and defense strategies on a large scale.

Papernot et al. (2016) also introduced the CleverHans library itself, providing an overview of its features, functionality, and usage. The authors highlighted the importance of benchmarking and evaluating the robustness of machine learning models against adversarial attacks and ultimately presented CleverHans as a comprehensive toolbox for conducting such evaluations. They discussed the attack methods implemented, including FGSM, BIM, and Carlini-Wagner L2 attack. That study also demonstrated the usage of CleverHans

for adversarial training and showcased its compatibility with popular machine learning frameworks.

4.2 Foolbox

Foolbox is another popular Python library for generating and evaluating adversarial attacks against machine learning models. It supports various attack methods, including gradient-based attacks, decision-based attacks, and score-based attacks. Foolbox is compatible with multiple deep learning frameworks, such as PyTorch, TensorFlow, and Keras, thus making it suitable for assessing the resilience of medical systems (Rauber et al. 2020). It also provides a user-friendly interface for generating adversarial attacks and evaluating their effectiveness. It offers a wide range of attack algorithms, which facilitates more comprehensive comparisons. However, compared to ART, Foolbox may have a smaller set of pre-implemented defense mechanisms (Rauber et al. 2020).

Rauber et al. (2017) introduced the Foolbox library and its features for use in comparing the reliability of machine learning models. The authors discussed various attack methods implemented in Foolbox, such as the Projected Gradient Descent (PGD), Fast Gradient Sign Method (FGSM), and DeepFool. They also demonstrated the usage of Foolbox to evaluate the robustness of various models, including deep neural networks, against these attacks. The paper also drew attention to the extensibility and ease of use of the Foolbox package. In another study, Dong et al. (2020b) presented a comprehensive evaluation of different adversarial attacks and defenses for image classification models. They utilized the Foolbox library to implement various attack methods, including FGSM, PGD, and CW (Carlini and Wagner) attacks, to generate adversarial examples. The researchers then evaluated the efficiency of various defense strategies in countering these attacks while using Foolbox as a comparison framework (Xie et al. 2020).

4.3 IBM adversarial robustness toolbox (ART)

The IBM Adversarial Robustness Toolbox (Nicolae et al. 2018) is an extensive open-source library that is dedicated to machine learning security. It offers a range of tools and techniques to evaluate and improve the resilience of models against adversarial attacks. ART supports different attack techniques, defense mechanisms, and model certification methods. It also includes tools that can be used for data preprocessing, feature extraction, and visualization (Nicolae et al. 2018).

IBM ART is similar to the Adversarial Robustness Toolbox mentioned earlier, but it also has additional features and enhancements. It offers a comprehensive set of robustness analysis features and provides state-of-the-art model certification techniques. IBM ART is actively maintained and continuously updated, thus providing access to the latest advancements in resilience research (Nicolae et al. 2018). Xie et al. (2020) introduced a novel approach called Robust Representation Matching (RRM), which transfers the robustness of a model that has been trained against adversarial attacks to another model designed for the same task, regardless of architectural differences. Under the influence of the student–teacher learning paradigm, RRM introduces a distinctive training loss that motivates the student model to adopt the robust characteristics of the teacher model. Compared to existing methods, RRM has excellent model performance and significantly reduces the training time required for

adversarial training. On the CIFAR-10 dataset, RRM trains a robust model approximately 1.8 times faster than the current state-of-the-art model. Notably, RRM remains effective even when applied to higher-dimensional datasets. On Restricted-ImageNet, RRM achieves an impressive improvement in training speed of approximately 18 times for a ResNet50 model compared to standard adversarial training techniques.

4.4 AI Fairness 360

AI Fairness 360 (Bellamy et al. 2019) is an open-source toolkit created by IBM Research that offers a suite of metrics, algorithms, and methods for reducing bias, with the ultimate aim of evaluating and improving the fairness and resilience of machine learning models. It encompasses features for identifying bias, assessing fairness, and training models with fairness considerations (Trewin 2018).

Usage: Researchers and practitioners who are interested in addressing bias and fairness issues in machine learning models can apply AI Fairness 360, which has been used in research papers focusing on algorithmic fairness, bias mitigation, and fairness-aware learning (Bellamy et al. 2019). Bellamy et al. (2019) introduced AI Fairness 360 (AIF360), a novel open-source Python toolkit that is specifically designed for algorithmic fairness. AIF360 is released under the Apache v2.0 license and serves two main objectives: to facilitate the seamless integration of fairness research algorithms into industrial applications and to establish a standardized framework to help fairness researchers exchange and assess algorithms.

The AI Fairness 360 toolkit consists of numerous fairness metrics that can be tailored to both datasets and models. These metrics are accompanied by comprehensive explanations, thus allowing users to thoroughly understand their interpretability and implications. AIF360 also incorporates advanced algorithms to mitigate bias in both datasets and models. To increase accessibility, it also provides an interactive web interface, thereby offering an intuitive introduction to the toolkit's concepts and capabilities. This user-friendly interface helps line-of-business users, researchers, and developers enrich the toolkit with new algorithms and enhancements while utilizing it for performance benchmarking purposes. AIF360 integrates a robust testing infrastructure to maintain code quality and reliability. By including these features, AIF360 aims to support the application of fairness principles in real-world scenarios and increase collaboration among fairness researchers.

4.5 Certified robustness to adversarial examples (Cleverhans v3)

Cleverhans v3 (Papernot et al. 2016) is an updated version of the CleverHans library that provides robustness certifications for machine learning models against adversarial attacks. It offers methods for computing certified lower bounds on adversarial accuracy and also provides certified defense techniques.

Usage: Researchers and practitioners who are interested in formally certifying the robustness of machine learning models against adversarial attacks can use Cleverhans v3. It has been used in research papers addressing certified robustness, provable defenses, and robust training (Goodfellow et al. 2016). Asha and Vinod (2022) presented a comprehensive analysis and comparison of three prominent adversarial machine learning (AML) tools: ART, CleverHans, and Foolbox, across various dimensions. The effectiveness of these tools was

examined by evaluating multiple adversarial samples. That study encompassed six different adversarial attacks conducted on diverse datasets while considering a range of machine learning models, namely InceptionV3, ResNet, and Madry. The Cifar-10 dataset was utilized for ResNet, ImageNet was used for InceptionV3, and MNIST was used for Madry. Specifically, CleverHans with the InceptionV3 model achieved an impressive efficiency of 0.95. The total attack accuracy and perturbation rate achieved by CleverHans were reported as 0.74 and 0.126, respectively. Overall, this analysis sheds light on the strengths and capabilities of ART, CleverHans, and Foolbox in terms of the attack effectiveness, defense mechanisms, and subjective perception of perturbed outputs.

4.6 Robustness gym

Robustness Gym is a library that provides a unified environment that can be used in evaluating the robustness of machine learning models. It offers a set of standardized tasks, benchmarks, and metrics for assessing model resilience to various types of attacks, including adversarial examples and data corruption (Goel et al. 2021).

Usage: Researchers and practitioners interested in benchmarking and comparing the reliability of different machine learning models can utilize Robustness Gym. It has been used in a range of research papers focusing on robustness evaluation, benchmarking, and adversarial defense. Goel et al. (2021) introduced Robustness Gym (RG) as an evaluation toolkit for assessing the robustness of natural language processing (NLP) systems. RG unifies four evaluation paradigms and provides a common platform for comparing and developing evaluation methods. One case study revealed the existence of performance degradation in sentiment modeling, while an analysis of named entity linking (NEL) systems and summarization models highlights both challenges and performance differences. RG is ultimately useful for both practitioners and researchers in evaluating and analyzing NLP systems.

Ovaisi et al. (2022) proposed a comprehensive approach to reliability evaluation by considering multiple dimensions such as transformations, sub-populations, distributional disparity, attacks, and data sparsity. Although there are existing libraries that can be used to compare recommender system models, there is no dedicated software library for full assessments of resilience in various scenarios. As a significant contribution, we introduce Robustness Gym for RecSys (RGRecSys), which is a toolkit that is specifically designed to uniformly evaluate the robustness of recommender system models.

4.7 IBM federated learning toolkit

The IBM Federated Learning Toolkit (IBM 2022) is a software library designed to support secure and privacy-conscious machine learning within the federated learning framework. It offers resources and methods for training models on decentralized data without exchanging raw data, thus maintaining privacy and increasing robustness.

Usage: Researchers and developers interested in federated learning and secure collaborative training can utilize the IBM Federated Learning Toolkit, which has been used in research papers focusing on privacy-preserving machine learning, distributed learning, and secure aggregation (IBM). Lo et al. (2022) explored a number of architectural patterns that were specifically developed to address the design challenges that are specific to federated learning systems. These patterns present adaptable solutions to frequently occurring issues

in the software architecture design which have been derived from an extensive review of the relevant literature. Covering a variety of elements such as client and model management, model training and aggregation, as well as configuration, these patterns are synchronized with the various state transitions observed throughout the federated learning model's lifecycle. This offers a crucial direction for effectively integrating these patterns into the architecture of federated learning systems.

Choudhury et al. (2019) presented a novel federated learning framework that was specifically designed to address these challenges. This framework makes it possible to train a global model from locally stored health data across multiple sites while ensuring two levels of privacy protection. First, this has the advantage of avoiding the need to transfer or share raw data during the model training process, thus preserving the confidentiality of sensitive information. Secondly, it incorporates a differential privacy mechanism to provide additional safeguards against potential privacy attacks. A comprehensive evaluation using electronic health data from one million patients in two healthcare applications demonstrates the effectiveness and viability of the proposed federated learning framework. The results highlight its ability to enhance privacy protection while maintaining the utility and effectiveness of the global model.

4.8 MIST (ML-driven intrusion detection system)

MIST is a machine learning-driven Intrusion Detection System (IDS) library. It provides algorithms and tools that can be used to detect and mitigate security threats and intrusions in computer networks. MIST combines machine learning techniques with network traffic analysis to identify anomalous patterns and potential attacks (Otoum 2019).

Usage: Researchers and practitioners in the field of network security and intrusion detection can use MIST to develop robust IDS solutions. It has been used in research papers focusing on network security, anomaly detection, and intrusion prevention (Otoum 2019). However, Alsarhan et al. (2021) introduced a novel approach by utilizing a support vector machine (SVM) for MIST (Machine learning-driven Intrusion Detection System) in VANET. SVM offers a number of computational advantages due to its structural characteristics, such as efficient handling of finite samples and independence between algorithm complexity and sample dimension. Intrusion detection in VANET poses a non-convex and combinatorial problem that requires the use of intelligent optimization algorithms to increase the accuracy of the SVM classifier. Particle swarm optimization (PSO), Genetic algorithm (GA), and ant colony optimization (ACO) are used as optimization techniques in that study. The experimental findings highlight the superiority of GA over those other two optimization algorithms in terms of performance.

Ge et al. (2021) presented a new approach to intrusion detection in the IoT by employing a customized deep learning technique. That approach utilizes a sophisticated IoT dataset, which includes authentic IoT traces and realistic attack scenarios such as denial of service, distributed denial of service, data gathering, and data theft. A multi-class classification model—i.e., a feed-forward neural network with embedding layers—is designed specifically to handle high-dimensional categorical features. The study also uses transfer learning to process these features, ultimately facilitating the development of a binary classifier using another feed-forward neural network model. After thorough testing, the proposed method

exhibits high accuracy in classification tasks for both binary and multi-class models, thus confirming its efficiency in detecting intrusions in IoT environments.

4.9 DeepSecure

DeepSecure is a Python library that focuses on deep learning-based security applications. It offers tools and models for tasks such as malware detection, intrusion detection, and network traffic analysis. DeepSecure provides pre-trained models and utilities for building robust and secure deep learning models in security domains (Rouhani et al. 2018).

Usage: Researchers and developers interested in applying deep learning techniques to security applications can utilize DeepSecure. It has been used in research papers focusing on deep learning for cybersecurity, malware detection, and network security (Kuadey et al. 2021). Pandey et al. (2016) developed a secure identification framework using deep neural networks that was specifically tailored to template protection in face password systems. They utilized deep convolutional neural networks (CNNs) to learn how to convert face images into maximum entropy binary (MEB) codes. The efficacy of this method has been validated through experiments on the Extended Yale B, CMU-PIE, and Multi-PIE face databases, where it achieved a high genuine accept rate (GAR) of 95% with a zero false accept rate (FAR), all while ensuring solid template security.

Pandey et al. (2015) introduced an innovative Deep Secure Encoding framework aiming to achieve secure classification with deep neural networks while emphasizing biometric template protection for facial recognition. This approach utilizes deep convolutional neural networks (CNNs) to link face classes and high entropy secure codes strongly. These codes are then hashed using standard hash functions like SHA-256, ultimately creating secure facial templates. The efficacy of this method was proven through tests on two well-known face databases, CMU-PIE and Extended Yale B, with the results exhibiting top-tier matching performance. The key features of this approach include cancelability, high security, and the avoidance of impractical assumptions. The proposed scheme is also adaptable for use in both identification and verification modes, which enhances its versatility and applicability.

4.10 Microsoft SEAL (simple encrypted arithmetic library)

Microsoft SEAL is a library for performing homomorphic encryption operations, which facilitate secure computation on encrypted data. It offers tools and utilities for encrypting data, performing operations on encrypted data, and decrypting the results, ultimately enabling secure computation while preserving data privacy (Chen et al. 2017).

Usage: Researchers and developers interested in secure and privacy-preserving machine learning can utilize Microsoft SEAL in building robust and secure applications. It has been used in a number of research papers focusing on homomorphic encryption, secure multi-party computation, and privacy-preserving analytics (Laine et al. 2017).

Natarajan and Dai (2021) developed SEAL-Embedded, the inaugural homomorphic encryption (HE) library tailored for embedded systems, while emphasizing the CKKS approximate homomorphic encryption scheme. This library integrates numerous computational and algorithmic enhancements along with a sophisticated memory re-use strategy. These features ultimately enable efficient memory use and high-performance CKKS encoding and encryption on embedded devices, thus maintaining robust security. This

makes SEAL-Embedded a complete solution for creating privacy-focused applications. As an example of its utility, it can perform asymmetric encryption of 2048 single-precision numbers in just 77 ms on the Azure Sphere Cortex-A7 platform and 737 ms on the Nordic nRF52840 Cortex-M4 platform, all within a modest 136 KB RAM footprint.

Laine and Player (2016) introduced the Simple Encrypted Arithmetic Library (SEAL) as a comprehensive and meticulously crafted homomorphic encryption library. Our primary objective was to develop a solution that not only exhibited robust engineering and clear documentation but also operated independently without relying on external dependencies. We aimed to make SEAL accessible to a wide range of users, including experts and individuals with limited or no cryptographic expertise. This library can be accessed at http://sealcrypto.codeplex.com and is made available under the MSR License Agreement, while ensuring compliance with licensing requirements and intellectual property rights.

4.11 Seldon Alibi

Seldon Alibi is a Python library that provides tools and algorithms for machine learning model inspection, explainability, and debugging. It offers techniques for model interpretation, adversarial detection, and outlier detection, ultimately enabling the evaluation and improvement of model robustness and security (Klaise et al. 2021).

Usage: Researchers and practitioners interested in model explainability and robustness can utilize Seldon Alibi to analyze and enhance the trustworthiness of machine learning models. It has been used in research papers focusing on model interpretability, adversarial detection, and model debugging.

Klaise et al. (2021) presented Alibi, a Python library (available at https://github.com/SeldonIO/alibi) designed to provide transparent insights into machine learning model predictions. This open-source library provides sophisticated algorithms for explainability in both classification and regression models which are suitable for both model-agnostic (black-box) and model-specific (white-box) situations. Alibi Explain supports various data types (tabular, text, images) while also offering both local and global explanation capabilities. With its unified API, users can easily work with explanations in a consistent manner. To ensure reliability, Alibi Explain follows the best development practices, as it undergoes extensive testing for code correctness and algorithm convergence in a continuous integration environment. The library provides comprehensive documentation on methods, usage, and theoretical backgrounds, which are accompanied by a collection of practical use cases. As a production-ready toolkit, Alibi Explain integrates with machine learning deployment platforms like Seldon Core and KFServing, and it also offers distributed explanation capabilities using Ray.

Rosa et al. (2022) focused on explainable artificial intelligence (XAI), which has recently come to be used across various scientific disciplines, thus shedding light on the rationale behind machine learning predictions. By offering transparency and interpretability, XAI techniques enhance trust and improve outcomes in decision-making processes driven by artificial intelligence. The use of XAI can be particularly valuable in the realm of human mobility research, as it promotes confidence in AI-driven analyses. That study presents a comparative analysis of XAI methods in the context of a regression problem related to smart human mobility. Decision Tree, LIME, SHAP, and Seldon Alibi are employed as explainable approaches to characterize human mobility patterns using a dataset derived from New

York Services. The obtained results demonstrate that all of these approaches yield insights and indicators that are meaningful in addressing our research problem.

4.12 IBM AI Explainability 360

IBM AI Explainability 360 is an open-source library that focuses on providing interpretability and explainability techniques for machine learning models. It offers a comprehensive set of algorithms and visualizations to help understand and explain the decisions made by models, thus enhancing their trustworthiness and robustness (Arya et al. 2021).

Usage: Researchers and practitioners interested in model interpretability and transparency can utilize IBM AI Explainability 360. It has been used in research papers focusing on explainable AI, model interpretability, and model debugging (Ganapavarapu et al. 2023).

Arya et al. (2022) noted the increasing demand for explanations in the context of artificial intelligence and machine learning algorithms, which come from a range of stakeholders such as citizens, regulators, experts, and developers. In response to these diverse needs, they introduced AI Explainability 360, an open-source software toolkit, in 2019. This toolkit encompasses ten advanced methods for explainability and two evaluation metrics. This paper presents an evaluation of the toolkit's impact through multiple case studies, statistical analyses, and feedback from the community. The results underscore the varied benefits and enhancements experienced by different user groups who have adopted this toolkit, including the independent LF AI & Data Foundation. The paper also highlights the toolkit's adaptable design, showcases instances of its usage, and emphasizes the availability of comprehensive educational resources and documentation for users.

Bellamy et al. (2019) discussed the growing significance of fairness in machine learning models, particularly in critical domains like mortgage lending, hiring practices, and prison sentencing. To address this issue, they introduced AI Fairness 360 (AIF360), an open-source Python toolkit released under the Apache v2.0 license which is accessible at https://github.com/ibm/aif360. The purpose of this toolkit is to promote the integration of fairness research algorithms with real-world applications and establish a common platform for fairness researchers to exchange and assess algorithms. AIF360 offers a wide array of fairness metrics for both datasets and models along with comprehensive explanations. It also includes algorithms to mitigate bias in both data and models. The toolkit also features an interactive web interface that serves as a user-friendly introduction for individuals in various roles, including line-of-business users, researchers, and developers. This helps users extend the toolkit by incorporating their own algorithms and enhancements while facilitating performance benchmarking.

4.13 OpenDP

OpenDP is a library that provides differential privacy features to analyze and share sensitive data while maintaining privacy. It offers mechanisms for adding noise to data, perform statistical analysis on differentially private data, and enable privacy-aware data collaborations.

Usage: Researchers and practitioners who are interested in privacy-preserving data analysis and secure data sharing can utilize OpenDP. It has been used in research papers focusing on differential privacy, secure data collaboration, and privacy-aware analytics (Tian 2023). Tian (2023) introduced an extension to the OpenDP library software framework that

enables it to handle differential privacy (DP) composition with adaptive privacy budgets using Renyi Differential Privacy (RDP). The authors achieved that by developing a Renyi filter and odometer, and they then established their privacy guarantees through a generalization of RDP Adaptive Composition. They also implemented a constructor that converts any odometer into a filter, thereby extending the applicability of their results. These advancements facilitate the practical deployment of machine learning algorithms and interactive query interfaces in real-world scenarios, ultimately allowing for adaptive updates to the privacy budget.

4.14 DeepArmor

DeepArmor is a comprehensive endpoint protection platform that leverages machine learning to detect and prevent advanced cyber threats, including malware, ransomware, and zero-day attacks. It uses deep learning models and behavioral analysis to identify and mitigate security risks in real time, thus providing robust security for endpoints.

Usage: Organizations and individuals concerned with securing their endpoints from advanced cyber threats can deploy DeepArmor as part of their security infrastructure. It has been used in research papers and industry applications focusing on endpoint security, machine learning-based threat detection, and malware prevention.

Ji et al. (2019) focused on the cognitive systems and machine learning techniques that have had great benefits in various applications. However, the recent surge in adversarial attacks, encompassing data poisoning, evasion attacks, and exploratory attacks, poses a threat to machine learning methods and can expose sensitive model parameters. To address these challenges, the authors of that study introduced DeepArmour, a cognitive system for malware classification and defense against adversarial attacks. Their approach relies on a voting system comprising three distinct machine-learning malware classifiers: random forest, multi-layer perceptron, and structure2vec. DeepArmour also incorporates adversarial countermeasures such as feature reconstruction and adversarial retraining to bolster robustness. In evaluating DeepArmour on a dataset featuring 12,536 malware samples spanning five categories, they found that it achieved an accuracy of 0.989 through tenfold cross-validation. They also conducted a white-box evasion attack on the dataset to assess the system's resilience, with DeepArmour demonstrating an accuracy of 0.675 for the generated unknown adversarial attacks. After retraining with just 10% adversarial samples, DeepArmour's accuracy improved to 0.839.

Table 5 serves as an analytical guide that can be used to compare and contrast two specialized machine learning packages that are designed with an emphasis on security and robustness in AI applications: secML and TensorFlow Privacy. Meanwhile, Table 6 provides an overview and comparative analysis of various security-oriented machine learning packages that are designed to enhance robustness in AI applications. While these packages address general AI and machine learning security needs, their relevance extends to medical diagnostics, where the use of specialized medical datasets like ChestX-ray14 or BraTS necessitates the development of robust defenses against adversarial and privacy threats to ensure reliable and secure medical diagnoses.

Table 5 Comparative analysis of AI security packages

Package name	Adversarial attacks	Data privacy	Model inversion	Poisoning attacks	Evasion attacks	Model stealing	Backdoor attacks	Membership inference	Other attacks	Robustness evaluation
secML (Pintor et al. 2022)	✓	✗	✗	✓	✓	✗	✗	✗	✗	✓
TensorFlow Privacy (Ramage and McMahan 2017)	✓	✓	✗	✓	✓	✗	✓	✓	✗	✓
IBM Differential Privacy Library (Bassily and Smith 2015)	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
OpenMined (Ayre 2023)	✗	✓	✗	✓	✗	✗	✓	✗	✗	✓
OWASP (Burato et al. 2017)	✓	✓	✗	✓	✗	✗	✗	✓	✗	✓
Artificial Intelligence Robustness Toolbox (Fawzi et al. 2016)	✓	✗	✗	✓	✓	✗	✗	✗	✗	✓
ART (adversarial robustness toolbox) (Nicolae et al. 2018)	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓
AI Fairness 360 (Bellamy et al. 2019)	✓	✓	✗	✓	✓	✗	✓	✓	✗	✓
Cleverhans v3 (Papernot et al. 2016)	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓
Robustness Gym (Rauber et al. 2017)	✗	✗	✗	✓	✗	✗	✗	✗	✗	✓
IBM Federated Learning Toolkit (Ludwig et al. 2007)	✗	✓	✗	✓	✗	✗	✗	✓	✗	✓
DeepArmor (deeparmor.com)	✓		✗	✗	✓	✗	✗	✗	✗	✓
Microsoft SEAL (Fawaz et al. 2021)		✓	✓	✗	✗	✓	✗	✗	✗	✓
Seldon Alibi (Klaise et al. 2021)	✗	✗	✓	✗	✗	✗	✗	✗	✗	✓
IBM AI Explainability 360 (Ganapavarapu et al. 2023)	✗	✓	✓	✗	✗	✗	✗	✗	✗	✓
OpenDP (Gaboardi et al. 2020)	✗	✓	✗	✗	✗	✗	✗	✗	✗	✓

5 Measuring robustness: quantitative approaches and evaluation techniques

In machine learning, robustness refers to a model's ability to perform well and make accurate predictions across various conditions, inputs, and scenarios. Evaluating the robustness of a model is a crucial aspect of ensuring its reliability and generalization beyond the training data. Here are some common metrics and approaches used to evaluate the robustness of machine learning models that are also summarized in Fig. 10:

5.1 Adversarial testing

Adversarial testing entails deliberately adding minor alterations to the input data to evaluate the model's susceptibility to attacks (Gaur et al. 2022). Adversarial examples are crafted to cause the model to make incorrect predictions while remaining imperceptible to human observers (Huang and Li 2023). Metrics like accuracy on adversarial examples and robustness against various attack methods (e.g., Projected Gradient Descent, Fast Gradient Sign Method) can be used in these cases. In the context of medical diagnosis, adversarial testing can simulate different lighting conditions, image qualities, or even variations in patient demographics. The model's performance should ideally remain stable and accurate despite these perturbations (Hendrycks and Gimpel 2017).

The concept of adversarial examples involves perturbing input data to cause a neural network to misclassify it. The Fast Gradient Sign Method (FGSM) is an efficient approach that is used to generate of adversarial examples. The process can be mathematically summarized as follows:

Given an input data point x , its true label y_true , a neural network with a loss function L , and a small perturbation ε:

Compute the gradient of the loss with respect to the input data:

$$\nabla\_xL\left(x, y\_true\right).$$

This step involves calculating the gradient of the loss function (L) with respect to the input data (x). y_true represents the true label of the input data. The gradient (∇_x) represents the direction in which the loss would increase the most if one were to tweak the input data. This is calculated using backpropagation.

Normalize the gradient:

$$\nabla\_xL\left(x, y\_true\right) / \left|\left|\nabla\_xL\left(x, y\_true\right)\right|\right|.$$

After the gradient has been obtained, it is normalized to ensure that the direction of the gradient is maintained but that the magnitude is scaled to 1. This is done by dividing the gradient by its norm ($\|\cdot\|$), which is the length of the gradient vector. Normalizing the gradient ensures that the modifications to the input data are made in the direction that affects the loss the most while keeping the scale of the changes consistent.

Create an adversarial example by adding the normalized gradient scaled by

$$\varepsilon: \text{ x\_adversarial } = \text{ x } + \epsilon * \text{ sign } \left(\nabla\_\text{x L } \left(\text{x, y\_true}\right)\right).$$

Finally, an adversarial example is created by adding a small perturbation to the original input data (x). The perturbation is the sign of the gradient multiplied by a small factor (ϵ), which controls the magnitude of the change. The function sign $(\cdot\cdot)$ returns the sign of each element in the gradient vector (i.e., $+1$ or -1), which is used to create a small change in the input that is likely to change the model's prediction. Therefore, the adversarial example (*xadversarial*) is a slightly modified version of the original input that is intended to mislead the model into making an incorrect prediction.

Madry et al. (2018) delve into the creation of adversarial examples and the vulnerabilities they reveal in neural networks. In particular, they discussed different techniques for generating adversarial examples, such as the Fast Gradient Sign Method, and they highlighted the importance of adversarial training to enhance model robustness against these attacks. Meanwhile, Goodfellow et al. (2015) proposed the concept of adversarial examples—inputs that differ slightly from the original data and cause misclassification of deep neural networks. The authors investigated how these adversarial perturbations affect neural networks and introduced the Fast Gradient Sign Method (FGSM) as an efficient method for generating adversarial examples. That study also focused on the observation that neural networks, despite achieving impressive performance on clean data, are susceptible to small changes in the input that humans can barely perceive. The results show that adversarial examples exploit the linearity of the neural networks' decision boundaries and that even imperceptible perturbations can lead to incorrect predictions. The paper also suggests adversarial training as a potential solution to enhance model robustness. By augmenting the training dataset with adversarial examples and training the model to classify them correctly, the model becomes more resistant to adversarial attacks. This approach introduces a form of regularization that encourages the model to behave consistently, even in the presence of perturbations.

However, that study focused primarily on understanding the existence of adversarial examples and their creation while offering limited guidance on practical defense strategies. The proposed adversarial training method might require substantial computational resources and can be vulnerable to stronger attack methods. The trade-off between accuracy and robustness is yet to be thoroughly explored, and the effectiveness of adversarial training on complex tasks is yet to be fully understood. Further studies could investigate practical adversarial defense mechanisms beyond adversarial training, such as input preprocessing, regularization techniques, and model ensembles. Such studies could also explore ways to improve the efficiency and scalability of adversarial training methods, ultimately making them more accessible for real-world applications. Some of the adversarial defenses are mentioned in Table 7.

5.2 Domain generalization and transfer learning

Evaluating a model's performance across different domains or datasets is important to ensure its robustness. If a model trained on one dataset performs well on another previously unseen dataset, that indicates that the model has learned to generalize and is less sensitive to data distribution shifts. Healthcare datasets can come from various hospitals, equipment, and demographics. Domain generalization techniques assess a model's generalization ability across diverse data sources. In this case, evaluation involves testing the model's performance on data from hospitals or regions that were not part of its training set.

Table 6 Overview and comparative analysis of security-oriented machine learning packages for robust AI applications

Package name	Description	Main features	Application/usage	Effective approaches to robustness (advantages)	Limitations/ disadvantages	Implemented defenses
secML (Pintor et al. 2022)	A Python library for secure and explainable machine learning	Secure, robust AI & ML support It can implement Evasion and Poisoning attacks	Researchers and practitioners interested in secure and explainable machine learning	Offers a comprehensive set of features for secure and explainable machine learning Supports multiple machine learning tasks and models	May have a learning curve for users unfamiliar with the library Limited community support compared to more widely used libraries	1. Countermeasures against evasion 2. Countermeasures against poisoning
TensorFlow Privacy (Ramage and McMahan 2017)	An extension of TensorFlow for privacy-preserving machine learning	Implements differential privacy techniques in TensorFlow and Enables privacy-aware model training Not explicitly designed for adversarial attacks but focuses on preventing data leakage	Researchers and practitioners concerned with privacy in machine learning	Seamlessly integrates with TensorFlow, a widely used machine learning framework Offers a range of differential privacy techniques	Limited support for models implemented using frameworks other than TensorFlow Requires familiarity with differential privacy concepts and techniques	1. Differential privacy mechanisms 2. Noise addition techniques

Table 6 (continued)

Package name	Description	Main features	Application/usage	Effective approaches to robustness (advantages)	Limitations/ disadvantages	Implemented defenses
IBM Differential Privacy Library (Bassily and Smith 2015)	A collection of differential privacy algorithms and mechanisms developed by IBM Research	Offers Private data analysis, synthetic data generation Not designed for adversarial attacks	Researchers and developers interested in incorporating differential privacy techniques	Provides a range of differential privacy algorithms and mechanisms Supports privacy-preserving data analysis and sharing	May require additional implementation effort for integrating with existing systems Limited documentation and community support compared to more widely used libraries	Differential privacy mechanisms, like the Laplace and Gaussian mechanisms
OpenMined (Ayre 2023)	An open-source community and platform for privacy-preserving machine learning and secure data collaboration	Provides tools, libraries, and frameworks for privacy-enhancing technologies	Researchers and developers interested in building privacy-preserving and secure ML models	Offers a wide range of privacy-enhancing tools and libraries Supports various privacy preserving techniques, including federated learning and secure multi-party computation	May have a steeper learning curve due to the complexity of privacy-preserving techniques Limited support for certain machine learning frameworks or architectures	1. Federated learning mechanisms 2. Encrypted computation methods

Table 6 (continued)

Package name	Description	Main features	Application/usage	Effective approaches to robustness (advantages)	Limitations/ disadvantages	Implemented defenses
OWASP (Burato et al. 2017)	An organization focused on improving software security, offering resources, guidelines, and tools for web application security	Comprehensive set of resources and tools for web application security It can implement many attacks, like Base Class Evasion Attacks, Base Class Poisoning Attacks, etc	Developers and security professionals aiming to enhance web application security	Offers a wealth of resources and tools for web application security Provides vulnerability scanning and secure coding guidelines	Focuses primarily on web application security and may not cover other domains The rapidly evolving threat landscape may necessitate frequent updates and adjustments to security practices	Countermeasures against evasion Countermeasures against poisoning
Artificial Intelligence Robustness Toolbox (Fawzi et al. 2016)	A Python library for evaluating and enhancing the robustness of machine learning models against adversarial attacks	Provides tools and algorithms for generating adversarial samples It can implement—Fast Gradient Sign Method (FGSM) Jacobian-based Saliency Map Approach (JSMA)	Researchers and practitioners interested in assessing and improving model robustness	Enables evaluation and analysis of model robustness against adversarial attacks Provides techniques for generating adversarial samples and hardening models	Primarily focuses on robustness against adversarial attacks and may not cover other types of vulnerabilities May require careful configuration and parameter tuning for effective use	ART. defences. detector. evasion ART. defences. detector. poison

Table 6 (continued)

Package name	Description	Main features	Application/usage	Effective approaches to robustness (advantages)	Limitations/ disadvantages	Implemented defenses
ART (adversarial robustness toolbox) (Nicolae et al. 2018)	An open-source library for adversarial robustness and generalization of machine learning models	Offers tools and algorithms for adversarial attacks and defenses It can also implement many attacks, like Base Class Evasion Attacks, Base Class Inference Attacks, etc	Researchers and practitioners interested in evaluating and improving model robustness	Provides a wide range of tools and algorithms for adversarial robustness and generalization Supports adversarial training and robust model evaluation	May require additional computational resources and time for certain adversarial attack methods Some methods may impact model performance or require careful parameter tuning	ART. defences. detector. evasion Subset Scanning Detector Binary Activation Detector Binary Input Detector, Data Provenance Defense ART. defences. detector. poison Activation Defence Spectral Signature Defense

Table 6 (continued)

Package name	Description	Main features	Application/usage	Effective approaches to robustness (advantages)	Limitations/ disadvantages	Implemented defenses
AI Fairness 360 (Bellamy et al. 2019)	A publicly available toolkit for evaluating the susceptibility of machine learning models to adversarial examples	Provides metrics and algorithms for measuring and mitigating bias and fairness issues in ML models	Researchers and practitioners interested in addressing bias and fairness in ML models	Provides a comprehensive set of tools and metrics for bias and fairness evaluation Offers algorithms for bias mitigation and fairness-aware training	Requires careful consideration of fairness metrics and trade-offs in model development May require additional data preprocessing and feature engineering for bias mitigation	Model explanation and interpretability techniques
Cleverhans v3 (Papernot et al. 2016)	An open-source library for benchmarking the vulnerability of machine learning models to adversarial examples	Supports adversarial training and robustness evaluation	Researchers and practitioners interested in assessing model vulnerability to adversarial attacks	Offers a wide range of adversarial attacks and defenses for model vulnerability assessment Supports adversarial training and robustness evaluation	Some adversarial attacks may require additional computational resources and time Some defenses may impact model performance or require careful parameter tuning	1. Adversarial training techniques 2. Input transformation methods
Robustness Gym (Rauber et al. 2017)	A toolkit for evaluating and benchmarking the robustness of machine learning models through various attacks and defenses	Provides a unified interface for implementing and evaluating adversarial attacks and defenses Perturbation techniques for NLP	Researchers and practitioners interested in evaluating model robustness using standardized benchmarks	Offers a unified interface for implementing and evaluating adversarial attacks and defenses Supports multiple domains and machine learning models	May require additional computational resources and time for certain attack methods Some defenses may impact model performance or require careful parameter tuning	Offers evaluation metrics for robustness

Table 6 (continued)

Package name	Description	Main features	Application/usage	Effective approaches to robustness (advantages)	Limitations/ disadvantages	Implemented defenses
IBM Federated Learning Toolkit (Ludwig et al. 2007)	An open-source toolkit for privacy-preserving federated learning	Offers tools and frameworks for developing privacy-preserving federated learning systems	Researchers and developers interested in privacy-preserving federated learning	Provides tools and frameworks for privacy-preserving federated learning Supports distributed training and model aggregation Ensures privacy guarantees for data sharing	Requires careful consideration of privacy risks and legal requirements in federated learning setups May have a learning curve for users unfamiliar with federated learning concepts and frameworks Limited support for certain machine learning frameworks or architectures	Differential privacy mechanisms
DeepArmor (deeparmor.com)	An endpoint protection platform that uses AI to detect and prevent malware attacks	Provides techniques for quantifying and mitigating information leakage in ML models	Security professionals and organizations looking for AI-based malware detection and prevention solutions	Provides real-time threat intelligence and proactive defense capabilities Supports endpoint protection across various platforms and operating systems Actively updated with evolving threat intelligence	Primarily focuses on malware detection and prevention and may not cover other aspects of model security Requires a subscription or other licensing for commercial use Limited customization options compared to open-source solutions	Real-time threat intelligence and proactive defense capabilities
Microsoft SEAL (Fawaz et al. 2021)	An open-source homomorphic encryption library for secure computation on encrypted data	Utilizes AI and machine learning techniques for advanced malware detection and prevention	Researchers and practitioners interested in secure computation and privacy-preserving machine learning	Provides a comprehensive set of homomorphic encryption techniques and operations Supports various machine learning tasks on encrypted data	Requires familiarity with homomorphic encryption concepts and techniques May have a learning curve for users unfamiliar with encryption-based computation Some operations may require additional computational resources and time	Homomorphic encryption techniques

Table 6 (continued)

Package name	Description	Main features	Application/usage	Effective approaches to robustness (advantages)	Limitations/ disadvantages	Implemented defenses
Seldon Alibi (Klaise et al. 2021)	An open-source library for machine learning model explainability and monitoring	Enables secure computation on encrypted data using homomorphic encryption techniques	Researchers and practitioners interested in model explainability and monitoring	Offers a range of tools and techniques for model explainability and evaluation Supports various explainability methods and metrics	May require careful selection and interpretation of explainability methods and metrics Limited support for certain machine learning frameworks or architectures	Model explanation and interpretability techniques to understand model vulnerabilities
IBM AI Explainability 360 (Ganapavarapu et al. 2023)	An open-source toolkit for explaining the decisions and behaviors of machine learning models	Supports various explainability methods and evaluation metrics Enables monitoring and drift detection in ML models	Researchers and practitioners interested in model explainability and interpretability	Offers a comprehensive set of algorithms and tools for model explainability and interpretability Provides visualization and post-hoc explanation techniques	May require careful selection and interpretation of explainability methods and metrics Limited support for certain machine learning frameworks or architectures	Various model explainability and transparency techniques to understand and potentially mitigate biases and vulnerabilities
OpenDP (Gaboardi et al. 2020)	An open-source library for differential private data analysis and machine learning	Provides a suite of algorithms and tools for model explainability and interpretability	Researchers and practitioners interested in privacy-preserving data analysis and machine learning	Offers a wide range of tools and algorithms for differential private data analysis Supports various differentially private mechanisms and models	Requires familiarity with differential privacy concepts and techniques May have limitations in handling complex privacy scenarios	Differential privacy mechanisms for various data analyses

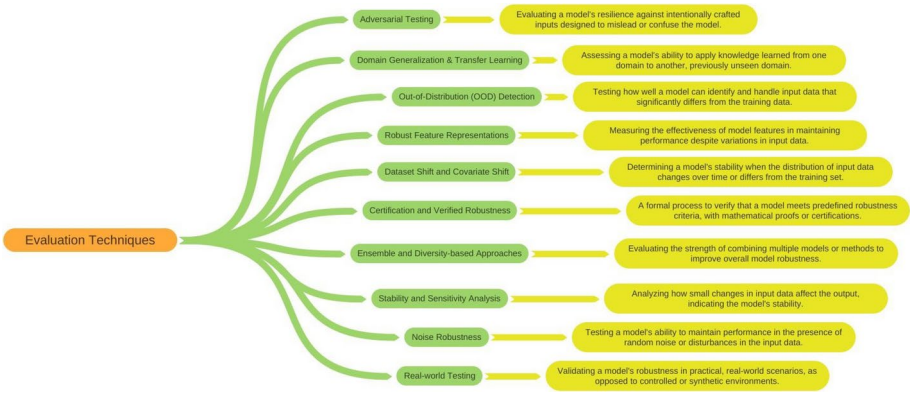

Fig. 10 Overview of robustness evaluation techniques for machine learning models

Mathematical Representation In the context of machine learning, Domain Generalization, and Transfer Learning are techniques used to ensure that a model M trained on one or more source domains D1, D2…. Dn can perform well on the target domain Dt, which may not have been seen during training. Domain Generalization involves training a model M on multiple source domains (M=Train (D1, D2…, Dn)) and then evaluating its performance on another target domain Performance (M, Dj)=Test (M, Dj)), where Dj is a domain that was not part of the training set. Meanwhile, Transfer Learning involves first training the Mpretrain model on the source domain Ds Mpretrain=Train (Ds)), then adapting it to the target domain Dt through a process denoted as Madapt=Adapt (Mpretrain, Dt), and finally evaluating the adapted model's performance on the target domain Performance (Madapt, Dt)=Test (Madapt, Dt)). These methodologies are fundamental to creating robust models that can handle various types of data distributions and apply the knowledge that has been acquired in one or several domains to solve problems that are different yet related.

$$Mpretrain = Train\,(Ds)$$

$$Madapt = Adapt\,(Mpretrain, Dt)$$

$$Performance\,(Madapt, Dt) = Test\,(Madapt, Dt)$$

Research by Ganin et al. (2016) focuses on training models that can generalize well across different domains. They discuss methods to reduce the distribution gap between the source and target domains, thus allowing the model to perform effectively in unseen scenarios. Another study by Ganin and Lempitsky (2015) addressed the domain adaptation challenge, where a model trained on one dataset (source domain) is expected to generalize well to another dataset (target domain). The authors introduce an unsupervised domain adaptation technique that minimizes the difference between source and target domain feature representations while retaining task-specific information. By minimizing the domain classifier's accuracy and maximizing the task classifier's accuracy, the model learns to disentangle domain-specific characteristics from task-related features. The paper demonstrates the efficacy of their approach on various image classification tasks across domains. However, the

Table 7 Comparative analysis of adversarial defense techniques in medical applications

Method	Effectiveness	Computational efficiency	Scalability with large datasets	Real-world application considerations
Multi-Perturbation Adversarial Training	High robustness against diverse attack vectors	High computational cost due to multiple perturbations	Limited scalability; resource-intensive	Effective but may not be practical for real-time applications
Misclassification-Aware Adversarial Training	Strong defense by focusing on high-risk misclassifications	Moderate computational demands	More scalable but still requires careful tuning	Suitable for critical medical scenarios with fewer resource constraints
Global Attention Noise (GATN) Injection	Effective in disrupting adversarial patterns	Computationally efficient	Scales well with large datasets	Practical for deployment in resource-limited environments, but varied effectiveness
Adversarial Logit Pairing (ALP)	Good improvement in robustness with pairwise regularization	Moderate computational overhead	Scalable with proper resource allocation	Balances robustness and efficiency, but requires careful hyperparameter tuning
Randomized Smoothing	Provides certified robustness with probabilistic guarantees	Low computational cost	Highly scalable; simple implementation	Practical for large-scale applications but may have lower robustness in certain scenarios
Input Gradient Regularization	Enhances robustness by minimizing gradient-based vulnerabilities	High computational cost due to regularization complexity	Limited scalability; resource-demanding	Effective but may slow down training, making it less suitable for urgent applications
Defensive Distillation	Moderately effective by reducing model sensitivity to perturbations	Low to moderate computational demand	Scales well with moderate computational resources	Useful in scenarios where model interpretability is key, but may be less effective against stronger attacks
Feature Squeezing	Reduces attack surface by limiting input space	Low computational cost	Highly scalable; lightweight and easy to implement	Practical for real-time applications with resource constraints, but may not defend against more sophisticated attacks
PixelDefend	Effective against specific types of adversarial attacks by reconstructing inputs	High computational demands due to complex input reconstruction	Limited scalability; not ideal for very large datasets	Suitable for targeted defense, but may be impractical for general use due to computational requirements
Spatial Smoothing	Provides robustness against spatially localized perturbations	Low to moderate computational cost	Scalable with careful implementation	Practical for certain types of medical images, but may not generalize well to all attack types

domain-adversarial training method used in the research assumes that domain shift occurs at the feature level, which is not always applicable in real-world scenarios. The domain adversarial framework may not be sufficient to handle either extreme domain shifts or scenarios in which domain-specific information is crucial to the target task. That paper does not deeply address the impact of dataset bias or its potential transfer across domains. Further research could investigate more sophisticated domain adaptation techniques for feature- and task-level domain shifts. Another direction could involve addressing the challenge of extreme domain shifts and ensuring effective adaptation in scenarios with limited labeled target data.

5.3 Out-of-distribution (OOD) detection

Models should be able to detect when the input data presented to them differs significantly from their training data. Metrics such as the Area Under the Receiver Operating Characteristic Curve (AUROC) can be used to evaluate a model's ability to distinguish between in-distribution and out-of-distribution samples. The ability to detect OOD samples is critical in a medical system, as it prevents the model from making confident but incorrect diagnoses for conditions in which it has yet to be trained. An effective OOD detection mechanism ensures that physicians are alerted when the model encounters unknown cases, thus prompting them to review and confirm the diagnosis before taking further actions.

Mathematical Representation Out-of-distribution (OOD) detection often relies on the premise that the model should output different confidence scores for in-distribution (ID) and OOD samples. In a deep learning context, models trained with SoftMax as their final layer produce probability distributions over the classes. The confidence of a model on a particular input can be measured as the maximum SoftMax probability:

$$confidence\,(x) = max_i P\,(y = i|x)$$

where $P(y=i|x)$ is the SoftMax probability of class i given input x. This confidence is typically high for well-trained models on ID data. However, for OOD samples, the confidence should ideally be lower, indicating the model's uncertainty.One standard method for OOD detection using deep neural networks involves the use of the temperature-scaled SoftMax. The equation then becomes:

$$confidence\,(x) = max_i \frac{e^{zi/T}}{\sum_j e^{zi/T}}$$

where zi is the logit for class i (i.e., the output of the model before the SoftMax), and T is the temperature parameter. By scaling with a higher temperature, the SoftMax outputs become more uniform (i.e., closer to a uniform distribution), which can make it easier to detect OOD samples based on their confidence scores. To assess the model's capability to differentiate between ID and OOD samples, the Receiver Operating Characteristic (ROC) curve can be used. This curve illustrates the true positive rate (sensitivity) versus the false positive rate (1-specificity) for various confidence thresholds. The Area Under the ROC Curve (AUROC) provides a single metric for evaluating the model's performance in detecting OOD samples.

Specifically, an AUROC of 0.5 implies no discrimination, similar to random guessing, while an AUROC of 1.0 signifies error-free discrimination.In the context of a medical system:

1. High confidence in ID samples means that the model is confident in its training-based diagnosis.
2. Low confidence (or ideally, recognized as OOD) for OOD samples means the model assumes that it hasn't seen such a case before and is uncertain about its prediction.

This mechanism can be a safety net to ensure that physicians intervene when the model encounters unknown conditions.

Out-of-distribution (OOD) detection is an active area of research, especially given the importance of robust and trustworthy machine learning models. Here are some research studies on OOD detection and related techniques in which several authors discussed OOD detection:

OOD detection is an important domain in deep learning that focuses on a model's ability to recognize input that diverges substantially from its training distribution. Hendrycks and Gimpel (2017) introduced a foundational approach advocating the utilization of max SoftMax probability from neural network classifiers to differentiate misclassified and OOD instances. However, its simplicity represents a potential drawback; relying solely on SoftMax probabilities may miss more subtle cases of OOD. Subsequent works like Lee et al. (2018) offered unified solutions for OOD and adversarial attack detection, such as by employing temperature-scaled SoftMax variants. Nevertheless, this approach can be more computationally intensive, particularly for large models. Ren et al. (2019) proposed a likelihood-ratio-based method that combines generative and discriminative models. The need to use a generative model, which is not always available or feasible, may represent a limitation. Choi et al. (2018) explored ensembles of generative models for robust OOD detection while employing the Watanabe-Akaike Information Criterion (WAIC) as a measure. However, using different ensemble methods may increase computational and memory requirements. While these contributions have significantly advanced the OOD detection field, a recurring theme is the trade-off between complexity, computational demands, and detection efficacy. Future works may benefit from exploring lightweight yet effective solutions, potentially through the development of hybrid approaches or the use of self-supervised learning paradigms for different data distributions.

5.4 Robust feature representations

Evaluating the robustness of learned features is crucial for model performance. Robust features withstand perturbations and enhance generalization across different scenarios. This evaluation involves stability metrics or comparing feature representations across different perturbed versions of the input data. Stability metrics assess how small changes in input data affect model predictions, particularly in medical diagnosis, where they are used to assess the impact of alterations in input data (such as lighting conditions in medical images) on the diagnoses made by the model. Greater stability in feature representations minimizes the likelihood of erratic diagnoses arising as a result of minor input variations.

Robust feature representations aim to identify functions (typically nonlinear) that transform raw input data into a space, highlighting and preserving intrinsic properties, even in the presence of noise or variations.

Definition: Let $f : Rn \rightarrow Rm$ be a function (often a neural network) that maps raw input $x \in Rn$ to a feature representation $z \in Rm$.

Robustness: A representation *z* is considered to be robust if small perturbations in the input *x* only result in small perturbations in *z*. Formally, for a perturbed input '*x*' with $\|x - x'\| 2 \leq \epsilon$, the corresponding feature representations should be close, i.e., $\|f(x) - f(x')\| 2 \leq \delta$, where δ is a small positive value.

Stability Metrics: Stability is often assessed by computing the Lipschitz constant of f. If *L* is the Lipschitz constant *f*, then for any *x* and '*x*':

$$\|f(x) - f(x')\| 2 \leq L \cdot \|x - x'\| 2$$

A smaller Lipschitz constant indicates more stable feature representations, meaning the representation changes slowly as the input changes.

Disentanglement: Disentangled representations aim to separate out the independent factors of variation in the data. Let $z = [z1, z2, \ldots zm]$ be the components of the feature representation. The representations are said to be disentangled if a change in one factor of variation in *x* predominantly changes only one *zi* while leaving others approximately unchanged.

Invariant Representations: Invariance implies that the feature representation remains unchanged under specific input transformations. Formally, if *T* is a transformation such that $x' = T(x)$, then an invariant representation ensures that $f(x) \approx f(x')$ for all such transformations within a set.

In a nutshell, robust feature representations mathematically encapsulate the idea that intrinsic data properties are captured in a manner that is resistant to noise, irrelevant variations, and specific transformations, ensuring stability, disentanglement, and invariance.

However, it is pivotal to include robust feature representations in the realm of deep learning, as this ensures that models extract salient and consistent patterns from data that remain invariant to various perturbations. One groundbreaking work in this area is by Goodfellow et al. (2015), which scrutinized the susceptibility of neural networks to adversarial examples and proposed adversarial training as a regularization method. While effective, adversarial training can be computationally expensive. Bengio et al. (2013) emphasized learning representations that are factorized over underlying explanatory factors, with their results suggesting that disentangling the learned features can improve robustness. However, explicitly enforcing disentanglement can sometimes lead to compromised performance. Moreover, Zhang et al. (2021) pointed out that deep neural networks can memorize noise and questioned the robustness of their learned representations. This brings forth the challenge of ensuring feature robustness without excessive memorization. A landmark work by Kornblith et al. (2019) used similarity metrics to assess the transferability and robustness of features, but their findings raised questions on the universality of these metrics across domains. Lastly, Sabour et al. (2017) introduced Capsule Networks that aim to learn spatial hierarchies between features, thus ensuring more robust feature representations against adversarial attacks. However, scalability and computational efficiency remain challenges

for Capsule Networks. While strides have been made in learning robust features, the balance between computational efficiency, interpretability, and true robustness is a frontier yet to be fully navigated, suggesting that research should continue exploring hybrid methods and domain-specific adaptations.

5.5 Dataset shift and covariate shift

Changes in data distribution can impact a model's performance. Metrics like the Wasserstein distance or Maximum Mean Discrepancy (MMD) can help quantify the divergence between training and test data distributions, thus providing insights into the model's potential sensitivity to dataset shifts. Medical datasets often come from different populations, demographics, or even time periods. In medical diagnosis, evaluation could involve measuring how well the model adapts to new patient groups without a significant loss of accuracy.

Dataset Shift At its core, a dataset shift occurs when the probability distribution that generated the training data $Ptrain\,(X, Y)$ differs from the one that generated the test data

$Ptest\,(X, Y)$. Formally, we have a dataset shift if: $Ptrain\,(X, Y) \neq Ptest\,(X, Y)$.

There are different types of dataset shifts, but one of the most common forms is *Covariate Shift*.

Covariate Shift Covariate shift occurs when the input distributions (or covariate distributions) differ between training and test sets, while the conditional distributions of the outputs given the inputs remain the same. Mathematically:

$$Ptrain\,(X) \neq Ptest\,(X)$$

$$Ptrain\,(Y|X) = Ptest\,(Y|X)$$

Several techniques and metrics can be employed to handle and measure covariate shift:

Wasserstein Distance Used in the context of optimal transport, the Wasserstein distance (or Earth Mover's Distance) between two distributions $Ptrain$ and $Ptest$ is the minimum mass transport cost that is needed to transform $Ptrain$ into $Ptest$. 1-Wasserstein distance is defined as: $W1\,(Ptrain, Ptest) = inf\gamma \in \Gamma\,(Ptrain, Ptest)\,E\,(x, x') \sim \gamma\,[\|x - x'\|]$, where $\Gamma\,(Ptrain, Ptest)$ denotes the set of all joint distributions on $X \times X$ with margins $Ptrain$ and $Ptest$. Understanding and measuring datasets and covariate shifts is crucial, particularly in domains like healthcare, where data distributions can vary across patient groups, demographics, and regions. Metrics like MMD and Wasserstein distance provide a solid basis for quantifying and subsequently accounting for these shifts, thus ensuring model robustness and reliability across varied distributions.

Dataset shift, particularly covariate shift, where the input distributions change while conditional distributions remain constant, poses significant challenges in machine learning applications. A seminal work by Shimodaira (2000) introduced weighting for re-weight training instances in an attempt to counteract covariate shifts. However, a limitation is that these weights can be noisy or even infeasible for high-dimensional data. Bengio et al. (2013) applied optimal transport, particularly the Wasserstein distance, to align distributions from different domains. Despite its mathematical elegance, its practical application can be computationally intensive for large datasets. In another study, "Invariant Representations for

Causal Inference", Gretton et al. (2012) explored invariant representations to ensure robustness across different environments, pointing to the complex relationship between invariant features and covariate shift. However, extracting invariant features is not always a straightforward task. In the paper "Optimal kernel choice for large-scale two-sample tests", Shi et al. (2021) provided a nonparametric approach that involved using the Maximum Mean Discrepancy (MMD) to measure distribution disparity, although it could have high computational costs when used with large-scale datasets. In summary, while these works present a strong foundation for addressing dataset shift, there is a recurring need to balance between robustness, computational efficiency, and adaptability to diverse, real-world shifts, suggesting that there is a path for hybrid methods or domain-adaptive architectures in future research (Sabour et al. 2017).

5.6 Certification and verified robustness

Some methods aim to guarantee a model's robustness. These methods involve computing a certification bound, which measures how far the model's predictions can deviate from certain perturbations. If the model's predictions remain within the certified bound, it is considered to be robust. A certification of robustness ensures that the diagnostic model's predictions will remain within specified limits, even under adversarial conditions. In healthcare, this may mean providing confidence intervals for the predicted probability of diagnosis. Verified robustness guarantees that model diagnoses are accurate even when faced with perturbations. This can be achieved by incorporating medical knowledge into the model architecture.

Certification Bound Given a model f and an input x with the prediction $y = f(x)$, the certification process involves identifying a bound ϵ such that, for all perturbed inputs, 'x' satisfies $\|x' - x\| \leq \epsilon$ (typically within some standard, e.g., L_2 norm), and the model's prediction $f(x')$ does not deviate beyond certain predefined limits. In other words, for the classifier: $f(x') = y$ for all 'x' in the ϵ-ball around x.

Confidence Intervals in Healthcare In the context of healthcare, the certified bound can be interpreted as a confidence interval. Suppose the model predicts the probability p of a particular diagnosis. A robustness certification can guarantee that, for perturbations in the ϵ, the predicted probability remains within the limits $[p - \delta, p + \delta]$, where δ is a small deviation.

Verified Robustness While certification limits deviations, verified robustness often leverages formal methods to provide more robust guarantees. This involves mathematically proving that the model's predictions remain accurate under all possible perturbations in a defined set. Given a model f, an input x, and its perturbed versions 'x', the verified robustness ensures: $\forall x' : \|x' - x\| \leq \epsilon \Rightarrow f(x') = y$ This implies that, for all threshold perturbations, the model's predictions remain consistent with the original predictions.

To achieve verified robustness, particularly in domain-specific contexts like healthcare, domain knowledge can be incorporated into the constraints. For instance, in medical imaging, perturbations that alter key image features (e.g., tumor size in a scan) cannot be specifi-

cally considered, which would help ensure that the model's predictions are consistent even under these medically relevant perturbations.

As an emerging field, the Certification and Verified Robustness has been the subject of extensive exploration to ensure model reliability under adverse conditions. A primary paper by Cohen et al. (2019) introduced a method to certify the robustness of deep neural networks using randomized smoothing, but it can have prohibitive computational costs in certain applications. In another work, Gehr et al. (2018) Raghunathan et al. presented a dual approach that could be used to provide verifiable robustness guarantees, although scalability for deeper networks remains a challenge. Raghunathan et al. (2018) used semi-definite programming to offer robustness certification; the drawback is that it may not be sufficiently tight for use with all datasets. In a domain-specific context, Gehr et al. (2018) presented an approach where they integrated abstract interpretation with neural networks to guarantee robustness specifically for vision-based tasks, but the method's generalizability in other fields is still under exploration. Finally, Singh et al. (2019) proposed an analysis framework using an abstract interpretation of deep neural network robustness, suggesting that domain-specific layers could improve certification boundaries. In summary, while these works present promising approaches for certification and verified robustness, challenges like computational cost, scalability, and domain specificity still require further exploration and refinement.

5.7 Ensemble and diversity-based approaches

Training multiple diverse models and combining their predictions can enhance robustness. Disagreement among ensemble members can serve as an indicator of uncertainty and robustness. It is possible to use metrics like ensemble diversity and accuracy improvement compared to individual models. In medical diagnostics, using a set of diverse models can provide more accurate and robust diagnoses by aggregating multiple perspectives on the same patient case. For instance, an ensemble can include models that have been trained on different patient demographics, imaging modalities, or disease manifestations.

Ensemble Prediction Consider a set of N models $\{M_1, M_2, ..., M_N\}$. Given an input x, the prediction of each model Mi is $yi = Mi(x)$. The ensemble prediction $\hat{y}$ can be an average for a majority vote: $\hat{y} = mode\{y1, y2, ..., yN\}$.

Diversity Metric Diversity can be quantified using various metrics. A simple metric is pairwise disagreement: $D = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} d(yi, yj)$, where $d(yi, yj)$ is a disagreement function (1 if $yi \neq yj$ and 0 otherwise for classification).

Ensemble Accuracy Improvement The accuracy improvement of the ensemble over individual models can be quantified as: $\Delta = A_{ensemble} - \frac{1}{N} \sum_{i=1}^{N} A_{M_i}$, where $A_{ensemble}$ is the accuracy of the ensemble and A_{M_i} is the accuracy of the i^{th} model.

Uncertainty Estimation Disagreement among ensemble members can be used to estimate prediction uncertainty. For a classification task with classes $C1, C2, ..., CK$, the uncertainty

U can be derived from the entropy of the class probabilities: $U = -\sum_{k=1}^{K} p\,(Ck)\,logp\,(Ck)$, where $p\,(Ck)$ is the proportion of ensemble members predicting class Ck .

In the context of medical diagnosis, the ensemble approach's power lies in aggregating predictions from models that might have been trained on diverse data sources or with different architectures. This aggregation reduces the risk associated with any single model's biases or shortcomings and provides a broader, more reliable perspective on a patient's case.

Ensemble methods have long been recognized for their potential to improve the prediction accuracy and robustness in machine learning. Breiman's seminal 1996 paper "Bagging Predictors" (Breiman 1996) was one of the first to describe how bootstrapped ensembles can reduce variance and increase generalizability. However, as Lakshminarayanan et al. claim, improved accuracy alone is not the only benefit. Zhou et al. (2002) in a study highlighting that deep ensembles can also serve as a tool for uncertainty estimation, which is crucial in healthcare applications. Meanwhile, Huang et al. (2017) introduced an efficient method for creating ensemble networks during a single training process, albeit sometimes at the cost of diversity. In terms of medical scenarios, Rajpurkar et al. (2017) pointed to the potential of ensembles to diagnose diseases using imaging data while also calling for the use of more diverse data sources to train individual models in the ensemble. Lastly, Zhou et al. (2002) stressed the significance of model diversity in the ensemble for achieving optimal performance, suggesting a focus on incorporating heterogeneity in the ensemble's constitution. In summary, while ensemble methods promise enhanced robustness and accuracy, the challenges of maintaining diversity, computational efficiency, and adaptability to new data sources continue to encourage the search for innovative solutions.

5.8 Stability and sensitivity analysis

Evaluating how small changes in input data affect a model's predictions can provide insight into its robustness. Sensitivity analysis involves perturbing inputs and measuring the resulting changes in predictions. The magnitude of these changes can be used to assess the model's stability. Stability metrics evaluate how small changes in patient data affect the model's diagnosis. Sensitivity analysis can help identify critical features for an accurate diagnosis. For example, analyzing the sensitivity of a medical image diagnostic system to variations in lighting conditions can provide insights into its robustness.

Stability and Sensitivity Analysis is fundamentally based on understanding how perturbations to the input of a function (in this case, a model) can influence its output. Mathematically speaking, sensitivity analysis is a general term for a series of techniques that can be used to analyze systems by studying the relationships between input and output variables. A mathematical approach to Stability and Sensitivity Analysis can be elaborated upon in the following:

Sensitivity Analysis For a given model *f*, $y = f\,(x)$, where *x* is our input data and y is our prediction. The sensitivity *S* of the model to the input variable xi can be described as:

$$Sxi = \frac{\delta y}{\delta xi}$$

This is the partial derivative of y with respect to xi, capturing how much y changes when xi is perturbed. A higher value of S indicates greater sensitivity of the model to changes in that specific input.

Stability Analysis Stability can be thought of as the inverse of sensitivity. If a small perturbation in x causes a large change in y, the system is sensitive but unstable. Stability can be quantified by looking at the norm of the perturbation in the input space versus the output space:

$$\delta y = f(x + \delta x) - f(x)$$

If $\| \delta y \|$ is small for a given $\| \delta x \|$, the model can be considered stable with respect to this perturbation.

Application in Medical Imaging Considering a medical image diagnostic system, suppose that I is the original image whereas 'I' is the same image but with altered lighting conditions. The sensitivity of the system to lighting variations can then be assessed using:

$$S_{light} = \frac{\delta_{diagnosis}}{\delta I'}$$

where the "diagnostics" function represents the model's output. If *Slight* is large, it means the diagnosis is highly sensitive to changes in lighting, which is a concern in robustness. Generally, both stability and sensitivity analysis provide tools for assessing the robustness and reliability of models, particularly when working with critical applications like medical diagnoses.Recently, Moussa et al. (2022) discussed the integration of deep learning models with traditional sensitivity analysis, emphasizing that the interpretability of deep neural networks can be improved, although at the cost of increased computational demands. Meanwhile, in a large-scale study in Caigny et al. (2020), they applied sensitivity analysis to various medical imaging tasks, finding that although the models showed increased robustness, there was a slight decrease in model accuracy in some edge cases. From a more theoretical point of view, Xiong et al. (2022) presented a comprehensive mathematical framework for stability analysis, although the proposed model requires further validation in various domains.

In an interesting approach by Campello et al. (2021), sensitivity analysis was applied to the area of natural language processing models. Their results highlight the susceptibility of NLP models to adversarial attacks, ultimately suggesting the need for more research in this area. Lastly, while these studies have substantially advanced our understanding in various ways, several limitations persist. As Jones noted, the computational intensity of integrating deep learning with sensitivity analysis may limit its scalability in real-world applications, while the observation of decreased accuracy in specific scenarios suggests the need for more detailed data preprocessing. The caution that must be taken when relying on a single measure of stability underscores the importance of adopting a multifaceted evaluation approach.

Moreover, to enhance the stability and sensitivity analyses, high-resolution sampling should be considered over the entire potential range of variations, and a variety of perturbations, such as Gaussian noise and domain-specific disturbances, should be applied. Global sensitivity methods can offer insight into the entire input space, whereas model assem-

bly can stabilize predictions. Leveraging regularization during training can prevent over-reliance on specific features, and domain-informed data augmentations, like rotations in medical imaging, can mirror real-world variations. Continuously updating the sensitivity analysis following model alterations ensures that model behavior is understood in a timely manner.

5.9 Noise robustness

Introducing noise to the input data and evaluating how well the model maintains its performance can indicate its robustness to noisy inputs. Metrics like Signal-to-Noise Ratio (SNR) or Mean Squared Error (MSE) can be used to quantify the model's performance under noise. In medical diagnosis, noise can affect various types of data, such as medical images, patient records, or sensor readings. A noise-robust diagnosis system should be able to provide accurate results even when the input data is corrupted by noise. For instance, a medical imaging system must be able to accurately diagnose a disease from an X-ray image, even if the image quality is affected by noise from the imaging device.

Noise robustness assesses a model's resilience to variations or disturbances in its inputs.

Signal-to-Noise Ratio (SNR) This is a measure that quantifies the extent to which a signal has been corrupted by noise. Mathematically, SNR is defined as: $SNR = \frac{Psignal}{Pnoise}$, where $Psignal$ is the power of the signal and $Pnoise$ is the power of the noise. In logarithmic decibel (dB) scale, it is given by: $SNR_{dB} = 10 \times log_{10}\left(\frac{Psignal}{Pnoise}\right)$

Mean Squared Error (MSE) This measures the average squared difference between the estimated values and the actual value. Given *n* true values $y1, y2, ...yn$ and their corresponding predicted values $\hat{y}1, \hat{y}2, ...\hat{y}n$, MSE is defined as: $MSE = \frac{1}{n}\sum_{i=1}^{n}(yi - \hat{y}i)^2$

In the context of medical diagnosis, noise can originate from various sources, such as a shaky hand causing blur in an imaging scan, electrical interference in an ECG trace, or even inconsistencies in manual data entry in patient records. The goal of noise robustness is to ensure that such perturbations do not significantly alter the model's predictions, ultimately achieving consistent and accurate medical diagnoses. Ensuring high SNR and minimizing the MSE are key indicators showing that the system is effectively handling noise and providing reliable outputs.

Over the years, several researchers in various domains have aimed to address the challenge of noise robustness. In the realm of speech recognition, Zhu et al. (2023) delved into techniques that could be used to enhance the robustness of DNNs against noisy data. However, a drawback is that the approach focuses primarily on speech data and might not have the ability to be generalized for other forms of data. They suggested further exploration into multimodal data noise robustness. In the domain of computer vision, Krizhevsky et al. (2017), in a study primarily introducing the AlexNet architecture, also highlighted concerns related to noisy visual data. A drawback was the limited focus on real-world noisy

scenarios, suggesting the need for real-world dataset experimentation. For medical imaging, a study by Manogaran et al. (2018) explored the Machine Learning Based Big Data Processing Framework for Cancer Diagnosis Using Hidden Markov Model and with GM Clustering. However, that study primarily focused on the framework's application within controlled environments, leaving room for further research to improve upon it by evaluating its effectiveness in more variable and real-world clinical settings. In the broader area of machine learning, Li et al. (2013) provided foundational insights on the impact of noisy data on learning algorithms. A potential drawback is its broader scope, which might miss out on specific nuances of noisy data in specialized domains, thus suggesting a need for domain-specific exploration. Lastly, "Generalization in Deep Learning" by Kawaguchi et al. (2022) offers a comprehensive view of the challenges and solutions related to noise robustness in deep learning models. Its main limitation is the primary focus on theoretical aspects, suggesting the need for more empirical studies for validation.

5.10 Real-world testing

Deploying a model in real-world scenarios and collecting performance metrics in practical settings is one of the most conclusive ways to evaluate a model's robustness. Deploying a medical system in healthcare environments and collecting performance metrics is crucial. However, the complexity of healthcare settings introduces challenges, such as patient conditions and equipment variations. Therefore, controlled evaluation should complement real-world testing to understand the model's performance under diverse conditions.

A general mathematical framing of Real-world Testing is as follows: When evaluating a model using real-world testing, we typically consider a set S of real-world scenarios. For each scenario $si \in S$, the model makes a prediction *pi* based on real-world input data xi.

The true label or output for each scenario is denoted as yi.

The performance metric, which is often denoted as $Performance\,(pi, yi)$, evaluates

the difference or similarity between pi and yi. Common metrics include accuracy, precision, recall, or even Mean Squared Error (MSE) for regression tasks.

The overall real-world performance P can be defined as:

$$P = |S|\,1 \sum i = 1\,|S|\,Performance\,(pi, yi)$$

P is the overall performance. |*S*| represents the cardinality of the set *S*, which is the number of samples in the set. The summation $\sum$ indicates the summation of all samples indexed by *i*. *i* is the index of a sample in the set *S*. Performance(*pi*, *yi*) is a function or measure of performance for each sample, comparing the prediction *pi* with the true value or label *yi*.

In a landmark study, Pavlitska et al. (2023) underscored the vulnerabilities and challenges faced when deploying models outside the safety of labs. Similarly, "Deploying AI in Clinical Settings" by Fihn et al. (2019) delves into the intricacies of transitioning from simulation to real-world scenarios, particularly emphasizing the variance in patient demographics and equipment standards.

Traditional non-AI diagnostic systems, which rely on conventional image processing and statistical techniques, are inherently more resistant to adversarial attacks due to their use of predefined rules and algorithms rather than complex data-driven models. This makes them more robust in certain scenarios, particularly when data integrity is compromised or when strict adherence to validated clinical protocols is required. However, this robustness often comes at the expense of lower overall performance compared to that of AI-based systems, particularly in handling large, complex datasets. While traditional methods excel in stable, well-defined conditions, they lack the scalability, precision, and adaptability that AI-enhanced systems can provide in more dynamic and data-intensive environments (Roy et al. 2022). That said, several studies have highlighted specific scenarios in which traditional non-AI diagnostic methods may outperform AI systems regarding robustness and reliability. For instance, traditional methods often exhibit superior performance in environments with poor data quality, such as those featuring low-resolution imaging or significant noise. Moreover, the image processing techniques used in traditional systems have been shown to maintain better diagnostic accuracy under such conditions compared to AI-based methods, which can be more sensitive to variations in data quality (Marulli et al. 2022). Another study emphasized the reliability of traditional clinical methods requiring strict adherence to validated protocols. In these cases, AI systems may struggle due to their reliance on large and diverse datasets, which may not always align with the specific requirements of established clinical protocols. Traditional systems, being rule-based and designed with specific clinical processes in mind, offer consistent and reliable outputs, which is crucial in highly controlled environments (Hu et al. 2022).

Table 8 provides a comprehensive overview of the robustness evaluation techniques and metrics used to assess the reliability of machine learning models, which serve as a critical resource for researchers and practitioners aiming to strengthen AI systems for various challenges. It can be used to determine distinct robustness types—adversarial, domain, noise, and more—with each having its own mathematical formulation, such as adversarial examples ($Xadv$) and domain features (XDi). The table further highlights the benefits of each approach, like improved model adaptability and safety guarantees, while also accounting for limitations, such as vulnerability to domain bias or information loss. By incorporating evaluation techniques like AUROC and calibration metrics, the table underscores the importance of rigorous validation methods to ensure model reliability in the face of adversarial attacks,

domain shifts, and other perturbations, which can help guide the development of more resilient AI applications.

6 Future directions and implementations

The robustness of deep learning stands as a bastion against the unpredictable nature of real-world data. It is a measure of a model's ability to remain calm and accurate, even when faced with input data that is noisy, incomplete, or deliberately perturbed. As deep learning models become an increasingly integral part of various applications, ensuring their robustness is not just a technical challenge—it is an imperative. Future research must enhance these models' ability to withstand and adapt to the various types of data irregularities common in the dynamic environments in which these models operate.

Table 8 Robustness evaluation techniques and metrics for machine learning models

References	Research aspect	Explanation	Robustness type	Mathematical notations and explanations	Benefits	Limitations	Evaluation techniques and metrics
Wang et al. (2023b), Chen and Hsu (2023), Ige et al. (2023)	Adversarial Testing	Evaluates model robustness against carefully crafted adversarial perturbations. Models should resist misclassification	Adversarial Robustness	X—Input data Y—Predicted labels X_{adv}—Adversarial input Y_{adv}—Adversarial prediction	Identifies vulnerabilities	Requires knowledge of potential attacks	Success rate of attacks Perturbation magnitude Robustness metrics (PGD accuracy, etc.)
Ding et al. (2023), Yao et al. (2023), Himeur et al. (2023)	Domain Generalization and Transfer Learning	Measures how well a model generalizes across diverse data sources. Ensures stability across different domains	Domain Robustness	Ds—Source domains Dt—Target domain	Enhances model adaptability	May suffer from domain bias	Model performance on unseen domains Domain discrepancy metrics
Wilson et al. (2023)	Out-of-Distribution (OOD) Detection	Focuses on detecting inputs significantly different from training data. Helps prevent overconfident incorrect predictions	Domain Robustness	$Pin\,(X)$—In-distribution probability $Pout\,(X)$—Out-of-distribution probability	Improves reliability of predictions	Performance degradation for in-distribution samples	AUROC (Area Under the Receiver Operating Characteristic Curve) Calibration metrics for uncertainty estimation
Singla et al. (2021)	Robust Feature Representations	Develops feature representations that capture essential information and are less sensitive to variations or noise	Noise Robustness	F—Feature representation	Enhances feature interpretability	May lead to information loss	Comparing feature distributions under perturbations Feature-level metrics like mutual information
Gretton et al. (2013), Sakai and Shimizu (2019)	Dataset Shift and Covariate Shift	Addresses shifts in data distributions. Models should generalize well across different data sources or covariate changes	Domain Robustness	X_i—Input features from domain Di	Increases model adaptability	May require additional data collection	Distribution distance metrics (Wasserstein, KL divergence) Improved performance under covariate shifts
Singh et al. (2018)	Certification and Verified Robustness	Aims to provide mathematical proofs of a model's robustness within certain bounds. Offers formal guarantees of safety	Adversarial Robustness	M_{cert}—Certified model	Provides formal safety guarantees	Limited to specific properties	Verification of model behavior under perturbations Proofs of correctness and robustness for specific properties

Table 8 (continued)

References	Research aspect	Explanation	Robustness type	Mathematical notations and explanations	Benefits	Limitations	Evaluation techniques and metrics
Partalas et al. (2008)	Ensemble and Diversity-based Approaches	Uses ensemble models to improve robustness by aggregating diverse predictions. Reduces agreement on incorrect predictions	Adversarial Robustness	M —Ensemble model	Enhances prediction diversity	Increased computational complexity	Ensemble performance compared to individual models Provides diverse metrics for ensemble predictions
Ngamkhanong et al. (2022)	Stability and Sensitivity Analysis	Measures how sensitive model predictions are to small changes in inputs. Helps understand the model's stability	Scalable Robustness	S —Sensitivity metric	Provides insights into model behavior	Limited to specific input changes	Sensitivity analysis on input perturbations Robustness under small changes
Hu et al. (2022), Shi et al. (2022), Chen et al. (2022)	Noise Robustness	Evaluates model performance in the presence of noisy input data. Models should handle noise without significant degradation	Noise Robustness	N —Noise	Ensures reliable performance	Performance degradation with increased noise	Accuracy under varying levels of noise Noise robustness metrics like Signal-to-Noise Ratio (SNR)
Deffayet et al. (2022)	Real-world Testing	Involves evaluating models in real-world scenarios to ensure practical robustness. Addresses unforeseen challenges	Model Robustness	E—Environment	Tests model performance in context	Challenges in replicating real-world scenarios	Improved performance in real-world environments Handling unexpected situations and edge cases

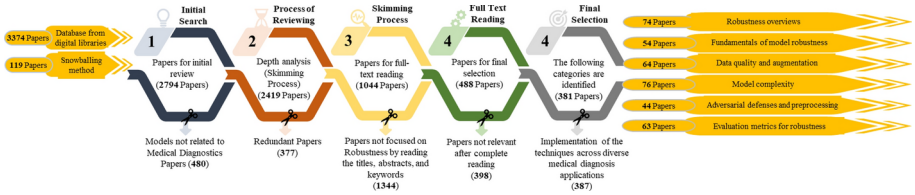

Fig. 11 Search strategy for categorizing the selected robustness research papers

One significant gap lies in the models' current ability to handle noisy and incomplete data. The next wave of research should aim to create algorithms that are inherently designed to be noise-resistant and that enable intelligent data attribution. Such models would be valuable in scenarios where the data collection process is particularly prone to interference or error. Developing sophisticated models to recognize underlying patterns in corrupted datasets will also be essential. This involves improving data preprocessing stages and embedding robustness into the very architecture of neural networks.

The threat of adversarial attacks represents the next frontier in robustness research. While current models have begun incorporating defense against such attacks, there is a continuous arms race between attack mechanisms and defensive strategies. Future directions should focus on developing more complex, adaptive algorithms that can dynamically respond to new and evolving adversarial tactics. This requires a dual approach of both advancing the theoretical understanding of adversarial machine learning and translating this knowledge into practical defensive techniques.

Alongside adversarial resistance, another aspect that is often overlooked is the robustness of models to changes in their operating environment, known as domain robustness. Future research should explore how models can maintain performance when deployed in environments other than those in which they were trained. This is particularly relevant for models to be used on global platforms, where data distribution can vary significantly. Research in this area would fill a critical gap by ensuring the versatility and reliability of deep learning models, regardless of geographic or system differences.

To advance the robustness of large language models (LLMs), future research should prioritize the development of inherently noise-resistant algorithms and intelligent imputation techniques to eliminate incomplete data. Particular emphasis should be placed on embedding robustness in neural architectures, enhancing adaptive defense against adversarial attacks, and enriching theoretical knowledge to provide practical solutions. Fostering domain adaptability will also be crucial for the global application of LLMs, requiring new robustness benchmarks that capture the model's resilience under varying conditions. This progress must be aligned with regulatory and ethical standards to ensure responsible deployment of the LLMs in real-world scenarios.

Finally, it is extremely important to introduce the concept of robustness in the realm of medical systems. The robustness of deep learning models in medical diagnostics ensures the reliability of automated analyses, which is a critical factor in patient care, where the cost of failure is immeasurable. Future research must be directed not only at strengthening these models in the face of the challenges mentioned above but also at ensuring that their deployment in healthcare settings does not compromise patient safety or privacy. This includes rigorously testing medical datasets, continuously monitoring model performance in clinical

settings, and developing models that clinicians can trust to assist them in making accurate diagnoses. Therefore, the pursuit of robustness in medical systems becomes a confluence of technological innovation, clinical insight, and ethical responsibility, leading to safer and more reliable healthcare solutions.

7 Conclusion

In summary, this paper has focused on strengthening the integrity of automated medical diagnostic systems through the robustness of the deep learning model, which is a quality that ensures that consistent performance is achieved even in the face of various types of disruptions. We carefully assessed how model architecture, data integrity, and algorithmic tuning influence robustness. Our investigation found that, although current literature presents numerous defense strategies against security threats like adversarial and privacy attacks, they are often ineffective against more sophisticated incursions. Our recommendations advocate incorporating techniques such as data augmentation, transfer learning, and uncertainty quantification to increase model robustness. We also explored and evaluated several tools and frameworks to enhance deep learning robustness. Our findings also highlight the need for improved metrics to assess model robustness, beyond the traditional accuracy measures accurately. To this end, our work underscores the need for an interdisciplinary strategy in developing medical systems that can withstand a broad spectrum of challenges, thereby ensuring their reliability for critical healthcare applications.

Appendix A: Literature research methodology

In constructing our study on the robustness of deep learning models in medical diagnoses, we adopted a rigorous literature review methodology. Given the interdisciplinary nature of the field, organizing and filtering relevant research was a challenge, and it led us to enforce specific inclusion criteria. We only considered studies focused on AI, Computer Science, Mathematics, Philosophy, and Psychology. We excluded research that primarily aimed to improve model transparency without directly focusing on explanations related to model robustness.

- Our interest was in Supervised ML models, so studies emphasizing different concepts were omitted. • Non-English literature was not included in our review.
- We only considered papers published after January 2010.

In our Google Scholar search, we used keywords like “deep learning robustness”, “medical model reliability”, “adversarial attacks on medical models”, and “medical model stability” during our search on Google Scholar. The time frame for our research spanned publications from January 2010 to August 2023. As depicted in Fig. 11, there has been a noticeable increase with time in the number of studies focusing on the robustness of the deep learning model. Beyond Google Scholar, we delved into PubMed, ScienceDirect, Web of Science, SpringerLink, Nature, Scopus, and IEEE Xplore to achieve a comprehensive literature review. We also included peer-reviewed papers on arXiv. Our selection of these

digital repositories was influenced by their reputation for providing seminal and up-to-date peer-reviewed publications that are particularly important to the model's robustness.

In our study of the robustness of deep learning models in medical diagnostics, we encompassed supervised ML, unsupervised ML, and reinforcement ML techniques to gain a holistic view. The main research direction in model robustness relates to supervised learning, so our investigation draws heavily on this domain. To ensure a comprehensive overview, we adopted the snowballing technique [516]. The 'Related Work' section of each referenced article was briefly reviewed to identify more pertinent studies. This method allowed us to discover further valuable contributions from platforms like the European Conference on Computer Vision (ECCV), ACM Computing Surveys, Computational Visual Media (CVM), the Workshop on Human-In-the-Loop Data Analytics (HILDA), and IEEE Transactions on Big Data.

As illustrated in Fig. 11, this search strategy yielded approximately 3374 peer-reviewed papers. Each of these papers underwent a meticulous evaluation of their title and abstract, in line with the intent of our study, which focused on the robustness of medical diagnostic models. Applying our stringent inclusion criteria, we filtered out studies that were not aligned with our objectives. Following this, we performed a deep dive into the content of the shortlisted papers, ensuring that we obtained the information that was most relevant to our study. Moreover, to ensure completeness, we manually reviewed the reference lists of these papers, which helped us identify other significant works. During our research focusing on the robustness of deep learning models in medical diagnostics, we were able to distinguish six fundamental categories from the vast collection of reviewed literature:

- Comprehensive Robustness Reviews—This category pertains to thorough reviews of model robustness techniques over the mentioned time frame. Table 1 provides a comprehensive summary of these reviews, shedding light on persistent challenges.
- Fundamental Robustness Concepts—Papers in this category focus on establishing the basic principles related to model robustness, with the aim of outlining its quintessential attributes.
- Data-driven Robustness—This case focuses on literature suggesting novel methodologies to augment robustness through critical analysis of training datasets.
- Inner Mechanism Robustness—This category consists of studies proposing cutting-edge techniques that aim to increase model robustness by deciphering the mechanisms underlying AI models.
- Adversarial Defense Robustness—The most important examples in this category are papers introducing novel methods of enhancing the model's robustness by elucidating human-comprehensible defense mechanisms against adversarial attacks.
- Robustness Evaluation—This category collects research analyzing the effectiveness of various techniques deployed to assess the model's robustness.

Following the compilation of relevant articles based on relevant keywords, we sequentially employed our exclusion criteria to select a targeted collection of papers consistent with the objectives of our study. Figure 11 illustrates the entire filtering process, marking the number of articles retained at each juncture. This categorization allowed us to construct a structured overview of the literature on model robustness. We paid special attention to certain specific facets: (i) papers highlighting unsupervised learning and RL models, (ii)

redundancies such as duplications or off-topic content, (iii) papers inconsistent with our main topic, and (iv) discrepancies or misclassifications in various evaluations. It is important to note that, given the multi-faceted nature of some papers, it is possible that some could potentially fit into multiple categories.

Acknowledgements This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(No. 2021R1A2C1011198), (Institute for Information & communications Technology Planning & Evaluation) (IITP) grant funded by the Korea government (MSIT) under the ICT Creative Consilience Program (IITP-2021-2020-0-01821), and AI Platform to Fully Adapt and Reflect Privacy-Policy Changes (RS-2022-II220688).

Author contributions H.J., S.E., and T.A. wrote the main manuscript text; H.J. and S.E., prepared figures and tables; T.A. provided the funding. All authors reviewed and edited the manuscript.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

Abbas Q (2022) A hybrid transfer learning-based architecture for recognition of medical imaging modalities for healthcare experts. J Intell Fuzzy Syst 43(5):5471–5486. https://doi.org/10.3233/JIFS-212171

Abd-Ellah MK, Khalaf AAM, Gharieb RR, Hassanin DA (2023) Automatic diagnosis of common carotid artery disease using different machine learning techniques. J Ambient Intell Humaniz Comput 14(1):113–129. https://doi.org/10.1007/s12652-021-03295-6

Abdukhamidov E, Abuhamad M, Woo SS, Chan-Tin E, Abuhmed T (2024) Hardening interpretable deep learning systems: investigating adversarial threats and defenses. IEEE Trans Depend Secure Comput 21(4):3963–3976. https://doi.org/10.1109/TDSC.2023.3341090

Agarwal A, Zhang T (2022) Minimax regret optimization for robust machine learning under distribution shift. In: Proceedings of machine learning research, PMLR, pp 2704–2729

Ahmad MA, Patel A, Eckert C, Kumar V, Teredesai A (2020) Fairness in machine learning for healthcare. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining, pp 3529–3530. https://doi.org/10.1145/3394486.3406461

Ahmad K, Maabreh M, Ghaly M, Khan K, Qadir J, Al-Fuqaha A (2022) Developing future human-centered smart cities: critical analysis of smart city security, Data management, and Ethical challenges. Comput Sci Rev 43:100452. https://doi.org/10.1016/j.cosrev.2021.100452

Ahmad A, Tariq A, Hussain HK, Gill AY (2023a) Equity and artificial intelligence in surgical care: a comprehensive review of current challenges and promising solutions. BULLET 2(2):443–455

Ahmad A, Saraswat D, El Gamal A (2023b) A survey on using deep learning techniques for plant disease diagnosis and recommendations for development of appropriate tools. Smart Agric Technol 3:100083. https://doi.org/10.1016/j.atech.2022.100083

Akkus Z et al (2019) A survey of deep-learning applications in ultrasound: artificial intelligence-powered ultrasound for improving clinical workflow. J Am Coll Radiol 16(9):1318–1328. https://doi.org/10.1016/j.jacr.2019.06.004

Akter S et al (2021) Algorithmic bias in data-driven innovation in the age of AI. Elsevier, Amsterdam

Albahri AS et al (2023) A systematic review of trustworthy and explainable artificial intelligence in healthcare: assessment of quality, bias risk, and data fusion. Inf Fusion 96:156–191. https://doi.org/10.1016/j.inffus.2023.03.008

Albayati MG, Faraj J, Thompson A, Patil P, Gorthala R, Rajasekaran S (2023) Semi-supervised machine learning for fault detection and diagnosis of a rooftop unit. Big Data Mining Anal 6(2):170–184. https://doi.org/10.26599/BDMA.2022.9020015

Ali M, Naeem F, Tariq M, Kaddoum G (2022) Federated learning for privacy preservation in smart healthcare systems: a comprehensive survey. IEEE J Biomed Health Inform 27(2):778–789

Ali S et al (2023) Explainable Artificial Intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. Inf Fusion 99:101805. https://doi.org/10.1016/j.inffus.2023.101805

Alnajem M, Garza-Reyes JA, Antony J (2019) Lean readiness within emergency departments: a conceptual framework. Benchmarking 26(6):1874–1904. https://doi.org/10.1108/BIJ-10-2018-0337

Alsarhan A, Alauthman M, Alshdaifat E, Al-Ghuwairi A-R, Al-Dubai A (2021) Machine Learning-driven optimization for SVM-based intrusion detection system in vehicular ad hoc networks. J Ambient Intell Humaniz Comput 1–10

Alvarez-Melis D, Jaakkola TS (2018) On the robustness of interpretability methods. arXiv preprint arXiv:1806.08049

Alves MA et al (2021) Explaining machine learning based diagnosis of COVID-19 from routine blood tests with decision trees and criteria graphs. Comput Biol Med 132:104335. https://doi.org/10.1016/j.compbiomed.2021.104335

Amann J, Blasimme A, Vayena E, Frey D, Madai VI (2020) Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 20(1):1–9. https://doi.org/10.1186/s12911-020-01332-6

Amini M, Pedram M, Moradi A, Ouchani M (2021) Diagnosis of Alzheimer's disease severity with fmri images using robust multitask feature extraction method and Convolutional Neural Network (CNN). Comput Math Methods Med 2021:1–15. https://doi.org/10.1155/2021/5514839

Amoroso N, Quarto S, La Rocca M, Tangaro S, Monaco A, Bellotti R (2023) An eXplainability Artificial Intelligence approach to brain connectivity in Alzheimer's disease. Front Aging Neurosci 15:1238065. https://doi.org/10.3389/fnagi.2023.1238065

Amugongo LM, Kriebitz A, Boch A, Lütge C (2023) Operationalising AI ethics through the agile software development lifecycle: a case study of AI-enabled mobile health applications. AI Ethics. https://doi.org/10.1007/s43681-023-00331-3

Anooj GVS, Marri GK, Balaji C (2023) A machine learning methodology for the diagnosis of phase change material-based thermal management systems. Appl Therm Eng 222:119864. https://doi.org/10.1016/j.applthermaleng.2022.119864

Anter AM, Abualigah L (2023) Deep federated machine learning-based optimization methods for liver tumor diagnosis: a review. Arch Comput Methods Eng 30(5):3359–3378. https://doi.org/10.1007/s11831-023-09901-4

Antunes N, Balby L, Figueiredo F, Lourenco N, Meira W, Santos W (2018) Fairness and transparency of machine learning for trustworthy cloud services. In: Proceedings—48th annual IEEE/IFIP international conference on dependable systems and networks workshops, DSN-W 2018, pp 188–193. IEEE. https://doi.org/10.1109/DSN-W.2018.00063

Apostolidis KD, Papakostas GA (2021) A survey on adversarial deep learning robustness in medical image analysis. Electronics 10(17):2132. https://doi.org/10.3390/electronics10172132

Argyroudis SA (2021) Resilience metrics for transport networks: a review and practical examples for bridges. In: Proceedings of the institution of civil engineers: bridge engineering, Thomas Telford Ltd, pp 179–192. https://doi.org/10.1680/jbren.21.00075

Arnold C, Biedebach L, Küpfer A, Neunhoeffer M (2024) The role of hyperparameters in machine learning models and how to tune them. Polit Sci Res Methods. https://doi.org/10.1017/psrm.2023.61

Arya V et al (2021) AI explainability 360 toolkit. In: Proceedings of the 3rd ACM India joint international conference on data science & management of data (8th ACM IKDD CODS & 26th COMAD), pp 376–379

Arya V et al (2022) AI explainability 360: impact and design. In: Proceedings of the AAAI conference on artificial intelligence, pp 12651–12657

Asha S, Vinod P (2022) Evaluation of adversarial machine learning tools for securing AI systems. Cluster Comput 1–20

Asif S, Yi W, Ain QU, Hou J, Yi T, Si J (2022) Improving effectiveness of different deep transfer learning-based models for detecting brain tumors from MR images. IEEE Access 10:34716–34730. https://doi.org/10.1109/ACCESS.2022.3153306
Ayre L (2023) OpenMined: an ecosystem for privacy-preserving machine learning. Accessed 8 Nov 2023. https://www.openmined.org/
Bai T, Luo J, Zhao J, Wen B, Wang Q (2021) Recent advances in adversarial training for adversarial robustness. In: IJCAI international joint conference on artificial intelligence, pp 4312–4321. https://doi.org/10.24963/ijcai.2021/591
Band SS et al (2023) Application of explainable artificial intelligence in medical health: a systematic review of interpretability methods. Inform Med Unlocked 40:101286. https://doi.org/10.1016/j.imu.2023.101286
Banu A, Amirtharajan R (2020) A robust medical image encryption in dual domain: chaos-DNA-IWT combined approach. Med Biol Eng Comput 58(7):1445–1458. https://doi.org/10.1007/s11517-020-02178-w
Barredo Arrieta A et al (2020) Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 58:82–115. https://doi.org/10.1016/j.inffus.2019.12.012
Bassily R, Smith A (2015) Local, private, efficient protocols for succinct histograms. In: Proceedings of the forty-seventh annual ACM symposium on theory of computing, pp 127–135
Bates DW, Auerbach A, Schulam P, Wright A, Saria S (2020) Reporting and implementing interventions involving machine learning and artificial intelligence. Ann Intern Med 172(11):S137–S144. https://doi.org/10.7326/M19-0872
Beil M, Proft I, van Heerden D, Sviri S, van Heerden PV (2019) Ethical considerations about artificial intelligence for prognostication in intensive care. Intensive Care Med Exp 7(1):1–13. https://doi.org/10.1186/s40635-019-0286-6
Bellamy RKE et al (2019) AI Fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. IBM J Res Dev 63(4–5):1–4. https://doi.org/10.1147/JRD.2019.2942287
Bengio Y, Courville A, Vincent P (2013) Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35(8):1798–1828. https://doi.org/10.1109/TPAMI.2013.50
Beyer H-G, Sendhoff B (2007) Robust optimization—a comprehensive survey. Comput Methods Appl Mech Eng 196(33–34):3190–3218
Bhandari M, Shahi TB, Neupane A (2023) Evaluating retinal disease diagnosis with an interpretable lightweight CNN model resistant to adversarial attacks. J Imaging 9(10):219. https://doi.org/10.3390/jimaging9100219
Bhardwaj C, Jain S, Sood M (2021) Transfer learning based robust automatic detection system for diabetic retinopathy grading. Neural Comput Appl 33(20):13999–14019. https://doi.org/10.1007/s00521-021-06042-2
Bin L et al (2022) Scheduling and sizing of campus microgrid considering demand response and economic analysis. Sensors 22(16):6150
Blagec K, Kraiger J, Frühwirt W, Samwald M (2023) Benchmark datasets driving artificial intelligence development fail to capture the needs of medical professionals. J Biomed Inform 137(2022):104274. https://doi.org/10.1016/j.jbi.2022.104274
Bordoloi D et al (2023) Classification and detection of skin disease based on machine learning and image processing evolutionary models. Comput Assist Methods Eng Sci 30(2):247–256. https://doi.org/10.24423/cames.479
Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140. https://doi.org/10.1007/bf00058655
Burato E, Ferrara P, Spoto F (2017) Security analysis of the OWASP benchmark with Julia. In: CEUR Workshop Proceedings, pp 242–247
Campello VM et al (2021) Multi-centre, multi-vendor and multi-disease cardiac segmentation: the M&Ms challenge. IEEE Trans Med Imaging 40(12):3543–3554
Čartolovni A, Tomičić A, Lazić Mosler E (2022) Ethical, legal, and social considerations of AI-based medical decision-support tools: a scoping review. Int J Med Inform 161:104738. https://doi.org/10.1016/j.ijmedinf.2022.104738
Casolla G, Cuomo S, Di Cola VS, Piccialli F (2020) Exploring unsupervised learning techniques for the Internet of Things. IEEE Trans Ind Inform 16(4):2621–2628. https://doi.org/10.1109/TII.2019.2941142
Cen J, Yang Z, Liu X, Xiong J, Chen H (2022) A review of data-driven machinery fault diagnosis using machine learning algorithms. J Vib Eng Technol 10(7):2481–2507. https://doi.org/10.1007/s42417-022-00498-9
Chang TS, Ward AC (1995) Design-in-modularity with conceptual robustness. In: American Society of Mechanical Engineers, Design Engineering Division (Publication) DE, American Society of Mechanical Engineers, pp 493–500
Chen C (2021) Improving the domain generalization and robustness of neural networks for medical imaging. BioMed Central

Chen GL, Hsu CC (2023) Jointly defending DeepFake manipulation and adversarial attack using decoy mechanism. IEEE Trans Pattern Anal Mach Intell 45(8):9922–9931. https://doi.org/10.1109/TPAMI.2023.3253390

Chen PY, Liu S (2023) Holistic adversarial robustness of deep learning models. In: Proceedings of the 37th AAAI conference on artificial intelligence, AAAI 2023, vol 37, pp 15411–15420. https://doi.org/10.1609/aaai.v37i13.26797

Chen H, Laine K, Player R (2017) Simple encrypted arithmetic library-SEAL v2. 1. In: Financial cryptography and data security: FC 2017 international workshops, WAHC, BITCOIN, VOTING, WTSC, and TA, Sliema, Malta, April 7, 2017, Revised Selected Papers 21. Springer, New York, pp 3–18

Chen J, Song L, Wainwright MJ, Jordan MI (2018) Learning to explain: an information-theoretic perspective on model interpretation. In: 35th international conference on machine learning, ICML 2018, pp 1386–1418. PMLR

Chen D, Tachella J, Davies ME (2022) Robust Equivariant Imaging: a fully unsupervised framework for learning to image from noisy and partial measurements. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 5637–5646. https://doi.org/10.1109/CVPR52688.2022.00556

Chen IY, Szolovits P, Ghassemi M (2019) Can AI help reduce disparities in general medical and mental health care? AMA J Ethics 21(2):167–179. https://doi.org/10.1001/amajethics.2019.167

Chivukula AS, Yang X, Liu B, Liu W, Zhou W (2023) Adversarial machine learning: attack surfaces, defence mechanisms, learning theories in artificial intelligence. Springer, New York. https://doi.org/10.1007/978-3-030-99772-4

Choi H, Jang E, Alemi AA (2018) WAIC, but Why? Generative ensembles for robust anomaly detection. arXiv preprint arXiv:1810.01392

Choudhury et al (2019) Differential privacy-enabled federated learning for sensitive health data. arXiv preprint arXiv:1910.02578

Chougrad H, Zouaki H, Alheyane O (2020) Multi-label transfer learning for the early diagnosis of breast cancer. Neurocomputing 392:168–180. https://doi.org/10.1016/j.neucom.2019.01.112

Cohen J, Rosenfeld E, Kolter JZ (2019) Certified adversarial robustness via randomized smoothing. In: 36th international conference on machine learning, ICML 2019, PMLR, 2019, pp 2323–2356

Coutellec L (2020) Ethics and scientific integrity in biomedical research. Handbook of research ethics and scientific integrity, pp 1–14. https://doi.org/10.1007/978-3-319-76040-7_36-1

Cuadra L, Salcedo-Sanz S, Del Ser J, Jiménez-Fernández S, Geem ZW (2015) A critical review of robustness in power grids using complex networks concepts. Energies 8(9):9211–9265. https://doi.org/10.3390/en8099211

Cui X et al (2021) DEAttack: a differential evolution based attack method for the robustness evaluation of medical image segmentation. Neurocomputing 465:38–52. https://doi.org/10.1016/j.neucom.2021.08.118

Cyran MA (2018) Blockchain as a foundation for sharing healthcare data. Blockchain Healthc Today. https://doi.org/10.30953/bhty.v1.13

Dai Y et al (2023) Improving adversarial robustness of medical imaging systems via adding global attention noise. Comput Biol Med 164:107251. https://doi.org/10.1016/j.compbiomed.2023.107251

De Caigny A, Coussement K, De Bock KW, Lessmann S (2020) Incorporating textual information in customer churn prediction models based on a convolutional neural network. Int J Forecast 36(4):1563–1578. https://doi.org/10.1016/j.ijforecast.2019.03.029

deeparmor.com. Deep Armor's Gauntlet powerful security monitoring platform. https://www.deeparmor.com/

Deffayet R, Renders J-M, de Rijke M (2022) Evaluating the robustness of click models to policy distributional shift. ACM Trans Inf Syst 41(4):1–28. https://doi.org/10.1145/3569086

DeVore S, Champion RW (2011) Driving population health through accountable care organizations. Health Aff 30(1):41–50. https://doi.org/10.1377/hlthaff.2010.0935

Dgani Y, Greenspan H, Goldberger J (2018) Training a neural network based on unreliable human annotation of medical images. In: Proceedings—international symposium on biomedical imaging, pp 39–42. IEEE. https://doi.org/10.1109/ISBI.2018.8363518

DI Y, Yang R, Huang M (2021) Fault diagnosis of rotating machinery based on domain adversarial training of neural networks. In: IEEE international symposium on industrial electronics, pp 1–6. IEEE. https://doi.org/10.1109/ISIE45552.2021.9576238

Ding Y, Jia M, Cao Y, Ding P, Zhao X, Lee CG (2023) Domain generalization via adversarial out-domain augmentation for remaining useful life prediction of bearings under unseen conditions. Knowl Based Syst 261:110199. https://doi.org/10.1016/j.knosys.2022.110199

Dong Y, Deng Z, Pang T, Zhu J, Su H (2020a) Adversarial distributional training for robust deep learning. Adv Neural Inf Process Syst 33:8270–8283

Dong Y et al (2020b) Benchmarking adversarial robustness on image classification. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 318–328. https://doi.org/10.1109/CVPR42600.2020.00040
Drenkow N, Sani N, Shpitser I, Unberath M (2021) A systematic review of robustness in deep learning for computer vision: mind the gap? 1–23
Duamwan LM, Bird JJ (2023) Explainable AI for medical image processing: a study on MRI in Alzheimer's disease. In: ACM international conference proceeding series, pp 480–484. https://doi.org/10.1145/3594806.3596521
Egli H, Totschnig L, Samartzis N, Kalaitzopoulos DR (2023) Biker's nodule in women: a case report and review of the literature. Case Rep Womens Health 39:e00539
El-Ghany SA, Azad M, Elmogy M (2023) Robustness fine-tuning deep learning model for cancers diagnosis based on histopathology image analysis. Diagnostics 13(4):699. https://doi.org/10.3390/diagnostics13040699
El Jellouli W et al (2023) The implications of AI in optimizing operating theatre efficiency. Asian J Res Surg 6(2)
El-Sappagh S, Alonso-Moral JM, Abuhmed T, Ali F, Bugarín-Diz A (2023) Trustworthy artificial intelligence in Alzheimer's disease: state of the art, opportunities, and challenges. Artif Intell Rev 56(10):11149–11296. https://doi.org/10.1007/s10462-023-10415-5
Elseddik M et al (2023) Predicting CTS diagnosis and prognosis based on machine learning techniques. Diagnostics 13(3):492. https://doi.org/10.3390/diagnostics13030492
Eren Y, Küçükdemiral İ (2024) A comprehensive review on deep learning approaches for short-term load forecasting. Renew Sustain Energy Rev 189:114031
Essemlali A, St-Onge E, Descoteaux M, Jodoin P-M (2020) Understanding Alzheimer disease's structural connectivity through explainable AI. In: Medical imaging with deep learning, PMLR, pp 217–229
Fang YP, Zio E (2019) An adaptive robust framework for the optimization of the resilience of interdependent infrastructures under natural hazards. Eur J Oper Res 276(3):1119–1136. https://doi.org/10.1016/j.ejor.2019.01.052
Fawaz SM, Belal N, ElRefaey A, Fakhr MW (2021) A comparative study of homomorphic encryption schemes using microsoft SEAL. Journal of Physics: Conference Series, IOP Publishing, p 12021
Fawzi A, Moosavi-Dezfooli S-M, Frossard P (2016) Robustness of classifiers: from adversarial to random noise. Adv Neural Inf Process Syst 29
Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S (2015) Certifying and removing disparate impact. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining, pp 259–268. https://doi.org/10.1145/2783258.2783311
Fihn S, Saria S, Matheny M, Shah N, Liu H, Auerbach A (2019) Deploying Ai in Clinical Settings. Artif Intell Health Care 145:145
Finlayson SG, Chung HW, Kohane IS, Beam AL (2018) Adversarial attacks against medical deep learning systems. arXiv preprint arXiv:1804.05296
Freitas S, Yang D, Kumar S, Tong H, Chau DH (2023) Graph vulnerability and robustness: a survey. IEEE Trans Knowl Data Eng 35(6):5915–5934. https://doi.org/10.1109/TKDE.2022.3163672
Gaboardi M, Hay M, Vadhan S (2020) A Programming Framework for OpenDP. In: Moratuwa Engineering Research Conference (MERCon), pp 578–583
Gadepally R, Gomella A, Gingold E, Lakhani P (2022) Generalization of artificial intelligence models in medical imaging: a case-based review. arXiv preprint arXiv:2211.13230
Ganapavarapu G et al (2023) AI Explainability 360 toolkit for time-series and industrial use cases. In: Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining, pp 5777–5778
Ganin Y, Lempitsky V (2015) Unsupervised domain adaptation by backpropagation. In: 32nd international conference on machine learning, ICML 2015, PMLR, pp 1180–1189
Ganin Y, Larochelle H, Marchand M (2016) 域适应9 (对抗训练2, 和与训练6太像了, 作者都一样, 应该基本上就是一个东西) Domain-Adversarial Training of Neural Networks. J Mach Learn Res 17(1):1–35
Garcia Valencia OA et al (2023) Ethical implications of chatbot utilization in nephrology. J Pers Med 13(9):1363. https://doi.org/10.3390/jpm13091363
Garg N, Schiebinger L, Jurafsky D, Zou J (2018) Word embeddings quantify 100 years of gender and ethnic stereotypes. Proc Natl Acad Sci USA 115(16):E3635–E3644. https://doi.org/10.1073/pnas.1720347115
Gaur L, Bhandari M, Razdan T (2022) Development of image translating model to counter adversarial attacks. DeepFakes. https://doi.org/10.1201/9781003231493-5
Ge M, Syed NF, Fu X, Baig Z, Robles-Kelly A (2021) Towards a deep learning-driven intrusion detection approach for Internet of Things. Comput Netw 186:107784
Gehr T, Mirman M, Drachsler-Cohen D, Tsankov P, Chaudhuri S, Vechev M (2018) AI2: safety and robustness certification of neural networks with abstract interpretation. In: Proceedings—IEEE symposium on security and privacy, pp 3–18. IEEE. https://doi.org/10.1109/SP.2018.00058

Ghaffari Laleh N et al (2022) Adversarial attacks and adversarial robustness in computational pathology. Nat Commun 13(1):5711. https://doi.org/10.1038/s41467-022-33266-0
Ghamizi S, Cordy M, Papadakis M, Le Traon Y (2023) On evaluating adversarial robustness of chest X-ray classification: pitfalls and best practices. In: CEUR workshop proc, vol 3381
Ghosh S, Shah D, More N, Choppadandi M, Ranglani D, Kapusetti G (2021) Clinical validation of the medical devices: a general prospective. In: BioSensing, theranostics, and medical devices: from laboratory to point-of-care testing, pp 265–297. https://doi.org/10.1007/978-981-16-2782-8_11
Ghosh D, Chowdhury K, Muhuri S (2023) Finding correlation between diabetic retinopathy and diabetes during pregnancy based on computer-aided diagnosis: a review. Multimed Tools Appl. https://doi.org/10.1007/s11042-023-16449-9
Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L (2019) Explaining explanations: an overview of interpretability of machine learning. In: Proceedings—2018 IEEE 5th international conference on data science and advanced analytics, DSAA 2018, pp 80–89. IEEE. https://doi.org/10.1109/DSAA.2018.00018.
Giuffrè M, Shung DL (2023) Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. NPJ Digit Med 6(1):186. https://doi.org/10.1038/s41746-023-00927-3
Goel K, Rajani N, Vig J, Taschdjian Z, Bansal M, Ré C (2021) Robustness gym: unifying the NLP evaluation landscape. In: NAACL-HLT 2021—2021 conference of the North American chapter of the association for computational linguistics: human language technologies, demonstrations, pp 42–55. https://doi.org/10.18653/v1/2021.naacl-demos.6
Gojić G, Vincan V, Kundačina O, Mišković D, Dragan D (2023) Non-adversarial robustness of deep learning methods for computer vision. In: Proceedings—10th international conference on electrical, electronic and computing engineering, IcETRAN 2023. https://doi.org/10.1109/IcETRAN59631.2023.10192125
Goodfellow IJ, Shlens J, Szegedy C (2015) Explaining and harnessing adversarial examples. In: 3rd international conference on learning representations, ICLR 2015—conference track proceedings
Goodfellow I, Papernot N, McDaniel P (2016) Cleverhans V0.1: an adversarial machine learning library. arXiv preprint arXiv:1610.00768, vol 1, no i, pp 1–18
Greco A, Strisciuglio N, Vento M, Vigilante V (2023) Benchmarking deep networks for facial emotion recognition in the wild. Multimed Tools Appl 82(8):11189–11220. https://doi.org/10.1007/s11042-022-12790-7
Gretton A et al (2012) Optimal kernel choice for large-scale two-sample tests. Adv Neural Inf Process Syst 2:1205–1213
Gretton A, Smola A, Huang J, Schmittfull M, Borgwardt K, Schölkopf B (2013) Covariate shift by Kernel mean matching. Dataset Shift Mach Learn 3(4):131–160. https://doi.org/10.7551/mitpress/9780262170055.003.0008
Hamon R, Junklewitz H, Sanchez I (2020) Robustness and explainability of artificial intelligence. Joint Res Centre 207:40
Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. Adv Neural Inf Process Syst 29:3323–3331
Harrison CJ, Sidey-Gibbons CJ (2021) Machine learning in medicine: a practical introduction to natural language processing. BMC Med Res Methodol 21(1):1–18. https://doi.org/10.1186/s12874-021-01347-1
Hendrycks D, Gimpel K (2017) A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: 5th international conference on learning representations
Higgins DC, Johner C (2023) Validation of artificial intelligence containing products across the regulated healthcare industries. Ther Innov Regul Sci 57(4):797–809. https://doi.org/10.1007/s43441-023-00530-4
Himeur Y et al (2023) Video surveillance using deep transfer learning and deep domain adaptation: towards better generalization. Eng Appl Artif Intell 119:105698. https://doi.org/10.1016/j.engappai.2022.105698
Holtz B, Nelson V, Poropatich RK (2023) Artificial intelligence in health: enhancing a return to patient-centered communication. Telemed e-Health 29(6):795–797. https://doi.org/10.1089/tmj.2022.0413
Holzinger A et al (2022) Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. Inf Fusion 79:263–278. https://doi.org/10.1016/j.inffus.2021.10.007
Hong Y, Zeng ML (2023) International classification of diseases (ICD). Knowl Organ 49(7):496–528
Hu Q (2021) A survey of adversarial example toolboxes. In: Proceedings—2021 2nd international conference on computing and data science, CDS 2021, pp 603–608. IEEE. https://doi.org/10.1109/CDS52072.2021.00109
Hu X, Chu L, Pei J, Liu W, Bian J (2021) Model complexity of deep learning: a survey. Knowl Inf Syst 63:2585–2619
Hu Q, Zhang G, Qin Z, Cai Y, Yu G, Li GY (2022) Robust semantic communications against semantic noise. In: IEEE vehicular technology conference, pp 1–6. IEEE. https://doi.org/10.1109/VTC2022-Fall57202.2022.10012843

Huang R, Li Y (2023) Adversarial attack mitigation strategy for machine learning-based network attack detection model in power system. IEEE Trans Smart Grid 14(3):2367–2376. https://doi.org/10.1109/TSG.2022.3217060

Huang G, Li Y, Pleiss G, Liu Z, Hopcroft JE, Weinberger KQ (2017) Snapshot ensembles: Train 1, get M for free. In: 5th international conference on learning representations, ICLR 2017—conference track proceedings

Huang X et al (2020) A survey of safety and trustworthiness of deep neural networks: verification, testing, adversarial attack and defence, and interpretability. Comput Sci Rev 37:100270. https://doi.org/10.1016/j.cosrev.2020.100270

IBM (2022) IBM Federated Learning—IBM Documentation. Accessed 21 Nov 2022. https://www.ibm.com/docs/en/cloud-paks/cp-data/4.5.x?topic=models-federated-learning

Ige T, Marfo W, Tonkinson J, Adewale S, Matti BH (2023) Adversarial sampling for fairness testing in Deep Neural Network. Int J Adv Comput Sci Appl 14(2):7–13. https://doi.org/10.14569/IJACSA.2023.0140202

Jahan S et al (2023a) Explainable AI-based Alzheimer's prediction and management using multimodal data. PLoS ONE 18(11):e0294253. https://doi.org/10.1371/journal.pone.0294253

Jahan S, Saif Adib MR, Mahmud M, Kaiser MS (2023b) Comparison between explainable AI algorithms for Alzheimer's disease prediction using EfficientNet models. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer, New York, pp 357–368. https://doi.org/10.1007/978-3-031-43075-6_31

Javaid M, Haleem A, Pratap Singh R, Suman R, Rab S (2022) Significance of machine learning in healthcare: Features, pillars and applications. Int J Intell Netw 3:58–73. https://doi.org/10.1016/j.ijin.2022.05.002

Javaid M, Haleem A, Singh RP, Suman R (2023) Towards insighting cybersecurity for healthcare domains: a comprehensive review of recent practices and trends. Cyber Secur Appl 1:100016. https://doi.org/10.1016/j.csa.2023.100016

Javed H, Muqeet HA, Shehzad M, Jamil M, Khan AA, Guerrero JM (2021) Optimal energy management of a campus microgrid considering financial and economic analysis with demand response strategies. Energies 14(24):8501. https://doi.org/10.3390/en14248501

Javed H, Muqeet HA, Javed T (2024) Ethical frameworks for machine learning in sensitive healthcare applications. IEEE Access 12(2023):16233–16254. https://doi.org/10.1109/ACCESS.2023.3340884

Jayabalan J, Jeyanthi N (2022) Scalable blockchain model using off-chain IPFS storage for healthcare data security and privacy. J Parallel Distrib Comput 164:152–167. https://doi.org/10.1016/j.jpdc.2022.03.009

Ji Y, Bowman B, Howie Huang H (2019) Securing malware cognitive systems against adversarial attacks. In: Proceedings—2019 IEEE international conference on cognitive computing, ICCC 2019—Part of the 2019 IEEE world congress on services, pp 1–9. https://doi.org/10.1109/ICCC.2019.00014

Joel MZ et al (2022) Using adversarial images to assess the robustness of deep learning models trained on diagnostic images in Oncology. JCO Clin Cancer Inform 6(6):e2100170. https://doi.org/10.1200/cci.21.00170

Joel MZ et al (2023) Comparing detection schemes for adversarial images against deep learning models for cancer imaging. Cancers 15(5):1548. https://doi.org/10.3390/cancers15051548

Johann LI et al (2023) A systematic collection of medical image datasets for deep learning. ACM Comput Surv 56(5):1–51. https://doi.org/10.1145/3615862

Juraev F, Abuhamad M, Woo SS, Thiruvathukal GK, Abuhmed T (2024) Impact of architectural modifications on deep learning adversarial robustness. arXiv preprint arXiv:2405.01934

Kaelbling LP, Littman ML, Moore AW (1996) Reinforcement learning: a survey. J Artif Intell Res 4:237–285

Kajić V, Esmaeelpour M, Považay B, Marshall D, Rosin PL, Drexler W (2012) Automated choroidal segmentation of 1060 nm OCT in healthy and pathologic eyes using a statistical model. Biomed Opt Express 3(1):86. https://doi.org/10.1364/boe.3.000086

Kamal MS, Northcote A, Chowdhury L, Dey N, Crespo RG, Herrera-Viedma E (2021) Alzheimer's patient analysis using image and gene expression data and explainable-AI to present associated genes. IEEE Trans Instrum Meas 70:1–7

Kass NE, Faden RR (2018) Ethics and learning health care: the essential roles of engagement, transparency, and accountability. Learn Health Syst 2(4):e10066. https://doi.org/10.1002/lrh2.10066

Kaviani S, Han KJ, Sohn I (2022) Adversarial attacks and defenses on AI in medical imaging informatics: a survey. Expert Syst Appl 198:116815. https://doi.org/10.1016/j.eswa.2022.116815

Kawaguchi K, Bengio Y, Kaelbling L (2022) Generalization in deep learning. Math Aspects Deep Learn 1(8):112–148. https://doi.org/10.1017/9781009025096.003

Kennedy DM, Caselli RJ, Berry LL (2011) A roadmap for improving healthcare service quality. J Healthc Manag 56(6):385–400. https://doi.org/10.1097/00115514-201111000-00007

Khakzar A, Albarqouni S, Navab N (2019) Learning interpretable features via adversarially robust optimization. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer, New York, pp 793–800. https://doi.org/10.1007/978-3-030-32226-7_88

Khalid N, Qayyum A, Bilal M, Al-Fuqaha A, Qadir J (2023a) Privacy-preserving artificial intelligence in healthcare: techniques and applications. Comput Biol Med 158:106848. https://doi.org/10.1016/j.compbiomed.2023.106848

Khalid N, Qayyum A, Bilal M, Al-Fuqaha A, Qadir J (2023b) Privacy-preserving artificial intelligence in healthcare: techniques and applications. Elsevier Ltd., Amsterdam. https://doi.org/10.1016/j.compbiomed.2023.106848

Khodabandehloo E, Riboni D, Alimohammadi A (2021) HealthXAI: collaborative and explainable AI for supporting early diagnosis of cognitive decline. Futur Gener Comput Syst 116:168–189. https://doi.org/10.1016/j.future.2020.10.030

Kireev K, Andriushchenko M, Troncoso C, Flammarion N (2023) Transferable adversarial robustness for categorical data via universal robust embeddings, arXiv preprint arXiv:2306.04064

Klaise J, Van Looveren A, Vacanti G, Coca A (2021) Alibi explain: algorithms for explaining machine learning models. J Mach Learn Res 22(1):8194–8200

Koçak B, Cuocolo R, Dos Santos DP, Stanzione A, Ugga L (2023) Must-have qualities of clinical research on artificial intelligence and machine learning. Balkan Med J 40(1):3–12. https://doi.org/10.4274/balkanmedj.galenos.2022.2022-11-51

Kornblith S, Norouzi M, Lee H, Hinton G (2019) Similarity of neural network representations revisited. In: 36th international conference on machine learning, ICML 2019, PMLR, pp 6156–6175

Krizhevsky A, Sutskever I, Hinton GE (2017) ImageNet classification with deep convolutional neural networks. Commun ACM 60(6):84–90. https://doi.org/10.1145/3065386

Kuadey NAE, Maale GT, Kwantwi T, Sun G, Liu G (2021) DeepSecure: detection of distributed denial of service attacks on 5G network slicing—deep learning approach. IEEE Wirel Commun Lett 11(3):488–492

Kurakin A, Goodfellow IJ, Bengio S (2017) Adversarial machine learning at scale. In: 5th international conference on learning representations, ICLR 2017—conference track proceedings

Laine K, Player R (2013) Simple Encrypted Arithmetic Library—SEAL (v2 . 0). In: Financial Cryptography and data security: FC 2017 international workshops, WAHC, BITCOIN, VOTING, WTSC, and TA, Sliema, Malta, April 7, 2017, Revised Selected Papers 21. Springer, New York, pp 3–18

Laine K, Player R (2016) Simple encrypted arithmetic library-seal (v2. 0). Technical report

Lakshminarayana S, Karachiwala JS, Teng TZ, Tan R, Yau DKY (2019) Performance and resilience of cyber-physical control systems with reactive attack mitigation. IEEE Trans Smart Grid 10(6):6640–6654. https://doi.org/10.1109/TSG.2019.2909357

Lane ND, Georgiev P, Qendro L (2015) DeepEar: robust smartphone audio sensing in unconstrained acoustic environments using deep learning. In: UbiComp 2015—proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing, pp 283–294. https://doi.org/10.1145/2750858.2804262.

Larson DB, Magnus DC, Lungren MP, Shah NH, Langlotz CP (2020) Ethics of using and sharing clinical imaging data for artificial intelligence: a proposed framework. Radiology 295(3):675–682. https://doi.org/10.1148/radiol.2020192536

Lee K, Lee K, Lee H, Shin J (2018) A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Adv Neural Inf Process Syst 7167–7177

Lestas I, Vinnicombe G (2005) Scalable robustness for consensus protocols with heterogeneous dynamics. IFAC proceedings volumes (IFAC-PapersOnline), vol 16, no 1, pp 185–190. https://doi.org/10.3182/20050703-6-cz-1902.00975

Li B, Tsao Y, Sim KC (2013) An investigation of spectral restoration algorithms for deep neural networks based noise robust speech recognition. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, pp 3002–3006. IEEE. https://doi.org/10.21437/interspeech.2013-278

Li H, Wang YF, Wan R, Wang S, Li TQ, Kot AC (2020) Domain generalization for medical imaging classification with linear-dependency regularization. Adv Neural Inf Process Syst 2020:3118–3129

Li X et al (2022) Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond. Knowl Inf Syst 64(12):3197–3234. https://doi.org/10.1007/s10115-022-01756-8

Lin J, Njilla LL, Xiong K (2022) Secure machine learning against adversarial samples at test time. EURASIP J Inf Secur 2022(1):1

Linardatos P, Papastefanopoulos V, Kotsiantis S (2021) Explainable AI: a review of machine learning interpretability methods. Entropy 23(1):1–45. https://doi.org/10.3390/e23010018

Litjens G et al (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60–88. https://doi.org/10.1016/j.media.2017.07.005

Liu Y, Peng J, James JQ, Wu Y (2019) PPGAN: privacy-preserving generative adversarial network. In: 2019 IEEE 25Th international conference on parallel and distributed systems (ICPADS), pp 985–989. IEEE
Liu Z, Fang L, Jiang D, Qu R (2022) A machine-learning-based fault diagnosis method with adaptive secondary sampling for multiphase drive systems. IEEE Trans Power Electron 37(8):8767–8772. https://doi.org/10.1109/TPEL.2022.3153797
Liu Z, Chen Y, Zhang Y, Ran S, Cheng C, Yang G (2023) Diagnosis of arrhythmias with few abnormal ECG samples using metric-based meta learning. Comput Biol Med 153:106465. https://doi.org/10.1016/j.compbiomed.2022.106465
Liu C et al (2024) A comprehensive study on robustness of image classification models: benchmarking and rethinking. Int J Comput Vis. https://doi.org/10.1007/s11263-024-02196-3
Lo SK, Lu Q, Zhu L, Paik H-Y, Xu X, Wang C (2022) Architectural patterns for the design of federated learning systems. J Syst Softw 191:111357
Lombardi A et al (2022) A robust framework to investigate the reliability and stability of explainable artificial intelligence markers of Mild Cognitive Impairment and Alzheimer's Disease. Brain Inform 9(1):1–17. https://doi.org/10.1186/s40708-022-00165-5
Ludwig H et al (2020) IBM federated learning: an enterprise framework White Paper V0.1. arXiv preprint arXiv:2007.10987
Lundqvist O, Fabricio Oliveira Advisor Fabricio Oliveira S (2023) A robust optimization approach against adversarial attacks on medical images. Thesis AaltoDoc, p 63
Ma L, Liang L (2023) Increasing-margin adversarial (IMA) training to improve adversarial robustness of neural networks. Comput Methods Prog Biomed 240:107687
Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A (2018) Towards deep learning models resistant to adversarial attacks. In: 6th international conference on learning representations, ICLR 2018—conference track proceedings
Magrabi F et al (2019) Artificial intelligence in clinical decision support: challenges for evaluating ai and practical implications. Yearb Med Inform 28(1):128–134. https://doi.org/10.1055/s-0039-1677903
Mahoto NA, Shaikh A, Sulaiman A, Al Reshan MS, Rajab A, Rajab K (2023) A machine learning based data modeling for medical diagnosis. Biomed Signal Process Control 81:104481. https://doi.org/10.1016/j.bspc.2022.104481
Manogaran G, Vijayakumar V, Varatharajan R, Malarvizhi Kumar P, Sundarasekar R, Hsu CH (2018) Machine learning based big data processing framework for cancer diagnosis using hidden markov model and GM clustering. Wirel Pers Commun 102(3):2099–2116. https://doi.org/10.1007/s11277-017-5044-z
Marinagi C, Reklitis P, Trivellas P, Sakas D (2023) The impact of industry 4.0 technologies on key performance indicators for a resilient supply chain 4.0. Sustainability 15(6):5185. https://doi.org/10.3390/su15065185
Maron RC et al (2021) A benchmark for neural network robustness in skin cancer classification. Eur J Cancer 155:191–199. https://doi.org/10.1016/j.ejca.2021.06.047
Marulli F, Marrone S, Verde L (2022) Sensitivity of machine learning approaches to fake and untrusted data in healthcare domain. J Sens Actuator Netw 11(2):21. https://doi.org/10.3390/jsan11020021
Masud M et al (2021) A lightweight and robust secure key establishment protocol for internet of medical things in COVID-19 patients care. IEEE Internet Things J 8(21):15694–15703. https://doi.org/10.1109/JIOT.2020.3047662
Md Nor N, Che Hassan CR, Hussain MA (2020) A review of data-driven fault detection and diagnosis methods: applications in chemical process systems. Rev Chem Eng 36(4):513–553. https://doi.org/10.1515/revce-2017-0069
Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2021) A survey on bias and fairness in machine learning. ACM Comput Surv 54(6):1–35
Meier BM, Rice H, Bandara S (2021) Monitoring attacks on health care as a basis to facilitate accountability for human rights violations. Health Hum Rights 23(1):55–70
Mewa T (2020) Fairness through awareness' by Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, & Rich Zemel *Cis.Pubpub.Org*
Miller DD (2019) The medical AI insurgency: what physicians must know about data to practice with intelligent machines. NPJ Digit Med 2(1):62
Misra S, Huddy J, Hanna G, Oliver N (2017) Validation and regulation of point of care devices for medical applications. In: Medical biosensors for Point of Care (POC) applications. Elsevier, Amsterdam, pp 27–44. https://doi.org/10.1016/B978-0-08-100072-4.00002-2
Miyato T, Dai AM, Goodfellow I (2017) Adversarial training methods for semi-supervised text classification. In: 5th international conference on learning representations, ICLR 2017—Conference Track Proceedings

Mok TCW, Chung ACS (2019) Learning data augmentation for brain tumor segmentation with coarse-to-fine generative adversarial networks. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer, New York, pp 70–80. https://doi.org/10.1007/978-3-030-11723-8_7

Molnar C, Casalicchio G, Bischl B (2020) Interpretable machine learning—a brief history, state-of-the-art and challenges. In: Communications in computer and information science. Springer, New York, pp 417–431. https://doi.org/10.1007/978-3-030-65965-3_28

Morley J et al (2021) The ethics of AI in health care: a mapping review. Philos Stud Ser 144:313–346. https://doi.org/10.1007/978-3-030-81907-1_18

Moskalenko V, Moskalenko A (2022) Neural network based image classifier resilient to destructive perturbation influences—architecture and training method. Radioelectron Comput Syst 2022(3):95–109. https://doi.org/10.32620/reks.2022.3.07

Moussa GS, Owais M, Dabbour E (2022) Variance-based global sensitivity analysis for rear-end crash investigation using deep learning. Accid Anal Prev 165:106514. https://doi.org/10.1016/j.aap.2021.106514

Muhammad A, Bae SH (2022) A survey on efficient methods for adversarial robustness. IEEE Access 10:118815–118830. https://doi.org/10.1109/ACCESS.2022.3216291

Muhammad I, Yan Z (2015) Supervised machine learning approaches: a survey. ICTACT J Soft Comput 5(3)

Mumby PJ, Chollett I, Bozec YM, Wolff NH (2014) Ecological resilience, robustness and vulnerability: how do these concepts benefit ecosystem management? Curr Opin Environ Sustain 7:22–27. https://doi.org/10.1016/j.cosust.2013.11.021

Muoka GW et al (2023) A comprehensive review and analysis of deep learning-based medical image adversarial attack and defense. Mathematics 11(20):4272. https://doi.org/10.3390/math11204272

Na HJ, Park JS (2021) Accented speech recognition based on end-to-end domain adversarial training of neural networks. Appl Sci 11(18):8412. https://doi.org/10.3390/app11188412

Naik N et al (2022) Legal and ethical consideration in artificial intelligence in healthcare: who takes responsibility? Front Surg 9:266. https://doi.org/10.3389/fsurg.2022.862322

Najafi A, Maeda SI, Koyama M, Miyato T (2019) Robustness to adversarial perturbations in learning from incomplete data. Adv Neural Inf Process Syst 32

Nan C, Sansavini G (2017) A quantitative method for assessing resilience of interdependent infrastructures. Reliab Eng Syst Saf 157:35–53. https://doi.org/10.1016/j.ress.2016.08.013

Natarajan D, Dai W (2021) Seal-embedded: a homomorphic encryption library for the internet of things. IACR Trans. Cryptogr Hardw Embed Syst 756–779

Natsiavas P, Malousi A, Bousquet C, Jaulent MC, Koutkias V (2019) Computational advances in drug safety: systematic and mapping review of knowledge engineering based approaches. Front Pharmacol 10:415. https://doi.org/10.3389/fphar.2019.00415

Navarro et al (2021) Evaluating the robustness of self-supervised learning in medical imaging. arXiv preprint arXiv:2105.06986

Naveed A (2023) Transforming clinical trials with informatics and AI/ML: a data-driven approach. Int J Comput Sci Technol 7(1):485–503

Ngamkhanong C et al (2022) Data-driven prediction of stability of rock tunnel heading: an application of machine learning models. Infrastructures 7(11):148. https://doi.org/10.3390/infrastructures7110148

Ngiam KY, Khor IW (2019) Big data and machine learning algorithms for health-care delivery. Lancet Oncol 20(5):e262–e273. https://doi.org/10.1016/S1470-2045(19)30149-4

Nguyen CT et al (2022) Transfer learning for wireless networks: a comprehensive survey. Proc IEEE 110(8):1073–1115. https://doi.org/10.1109/JPROC.2022.3175942

Nicholson PW (2017) Artificial intelligence in health care: applications and legal issues. SciTech Lawyer 14(1):10–13

Nicolae M-I et al (2018) Adversarial robustness toolbox v1.0.0, arXiv preprint arXiv:1807.01069

Ning J, Li Y, Guo Z (2023) Evaluating similitude and robustness of deep image denoising models via adversarial attack. arXiv preprint arXiv:2306.16050

Niyirora R, Ji W, Masengesho E, Munyaneza J, Niyonyungu F, Nyirandayisabye R (2022) Intelligent damage diagnosis in bridges using vibration-based monitoring approaches and machine learning: a systematic review. Results Eng 16:100761. https://doi.org/10.1016/j.rineng.2022.100761

Nowrozy R, Ahmed K, Wang H, Mcintosh T (2023) Towards a universal privacy model for electronic health record systems: an ontology and machine learning approach. In: Informatics, MDPI, p 60. https://doi.org/10.3390/informatics10030060

Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science 366(6464):447–453. https://doi.org/10.1126/science.aax2342

Oktian YE, Lee SG, Lee HJ, Lam JH (2017) Distributed SDN controller system: a survey on design choice. Comput Netw 121:100–111. https://doi.org/10.1016/j.comnet.2017.04.038

Otoum S (2019) Machine learning-driven intrusion detection techniques in critical infrastructures monitored by sensor networks. Université d'Ottawa/University of Ottawa, p 144
Ovaisi Z, Heinecke S, Li J, Zhang Y, Zheleva E, Xiong C (2022) Rgrecsys: a toolkit for robustness evaluation of recommender systems. In: Proceedings of the fifteenth ACM international conference on web search and data mining, pp 1597–1600
Oymak S (2019) Stochastic gradient descent learns state equations with nonlinear activations. In: Proceedings of machine learning research, PMLR, pp 2551–2579
Pandey A, Jain K (2022) A robust deep attention dense convolutional neural network for plant leaf disease identification and classification from smart phone captured real world images. Ecol Inform 70:101725. https://doi.org/10.1016/j.ecoinf.2022.101725
Pandey R, Zhou Y, Govindaraju V (2015) Deep secure encoding: an application to face recognition. arXiv preprint arXiv:1506.04340
Pandey RK, Zhou Y, Kota BU, Govindaraju V (2016) Deep secure encoding for face template protection. In: IEEE computer society conference on computer vision and pattern recognition workshops, pp 77–83. https://doi.org/10.1109/CVPRW.2016.17
Pansota MS, Khan HA, Rehman A (2021) A comparative analysis of artificial intelligence and machine learning approach to estimate currents in electrical power transmission lines. Univ Wah J Sci Technol 5:72–80
Papernot N et al (2016) Technical Report on the CleverHans v2.1.0 Adversarial Examples Library, arXiv preprint arXiv:1610.00768
Partalas I, Tsoumakas G, Vlahavas I (2008) Focused ensemble selection: a diversity-based method for greedy ensemble selection. Front Artif Intell Appl. https://doi.org/10.3233/978-1-58603-891-5-117
Patrini G, Rozza A, Menon AK, Nock R, Qu L (2017) Making deep neural networks robust to label noise: a loss correction approach. In: Proceedings—30th IEEE conference on computer vision and pattern recognition, CVPR 2017, pp 2233–2241. https://doi.org/10.1109/CVPR.2017.240
Pavlitska S, Lambing N, Zöllner JM (2023) Adversarial attacks on traffic sign recognition: a survey, arXiv preprint arXiv:2307.08278. https://doi.org/10.1109/ICECCME57830.2023.10252727
Pintor M, Demetrio L, Sotgiu A, Melis M, Demontis A, Biggio B (2022) secml: secure and explainable machine learning in Python. SoftwareX, vol 18, https://doi.org/10.1016/j.softx.2022.101095
Pitas I (2021) Privacy protection, ethics, robustness and regulatory issues in autonomous systems. In: 2021 10th Mediterranean conference on embedded computing (MECO), pp 1–1. IEEE. https://doi.org/10.1109/meco52532.2021.9460216.
Price W, Nicholson II (2019) Medical AI and contextual bias. Harv JL Tech 33:65
Priya KV, Dinesh PJ (2023) A detailed study on adversarial attacks and defense mechanisms on various deep learning models. In: Proceedings of the ACCTHPA 2023—conference on advanced computing and communication technologies for high performance applications, pp 1–6. IEEE. https://doi.org/10.1109/ACCTHPA57160.2023.10083378
Pronovost PJ, Armstrong CM, Demski R, Peterson RR, Rothman PB (2018) Next level of board accountability in health care quality. J Health Organ Manag 32(1):2–8. https://doi.org/10.1108/jhom-09-2017-0238
Qayyum A, Qadir J, Bilal M, Al-Fuqaha A (2021) Secure and robust machine learning for healthcare: a survey. IEEE Rev Biomed Eng 14:156–180. https://doi.org/10.1109/RBME.2020.3013489
Qiu J, Oppelt MP, Nissen M, Anneken L, Breininger K, Eskofier B (2022) Improving deep learning-based cardiac abnormality detection in 12-lead ECG with data augmentation. In: Proceedings of the annual international conference of the IEEE engineering in medicine and biology society, EMBS, pp 945–949. IEEE. https://doi.org/10.1109/EMBC48229.2022.9871969
Qiu Y et al (2023) Two-stage distributionally robust optimization-based coordinated scheduling of integrated energy system with electricity-hydrogen hybrid energy storage. Prot Control Mod Power Syst 8(2):1–14
Raghunathan A, Steinhardt J, Liang P (2018) Semidefinite relaxations for certifying robustness to adversarial examples. Adv Neural Inf Process Syst 31:10877–10887
Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH (2018) Ensuring fairness in machine learning to advance health equity. Ann Intern Med 169(12):866–872. https://doi.org/10.7326/M18-1990
Rajpurkar P et al (2017) CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225
Ramage D, McMahan B (2017) Federated learning: collaborative machine learning without centralized training data. https://blog.research.google/2017/04/federated-learning-collaborative.html
Rasheed K, Qayyum A, Ghaly M, Al-Fuqaha A, Razi A, Qadir J (2022) Explainable, trustworthy, and ethical machine learning for healthcare: a survey. Comput Biol Med 149:106043. https://doi.org/10.1016/j.compbiomed.2022.106043
Rashid KMJ (2023) Optimize the Taguchi method, the signal-to-noise ratio, and the sensitivity. Int J Stat Appl Math 8(6):64–70. https://doi.org/10.22271/maths.2023.v8.i6a.1406

Rauber J, Brendel W, Bethge M (2017) Foolbox: a python toolbox to benchmark the robustness of machine learning models. arXiv preprint arXiv:1707.04131

Rauber J, Zimmermann R, Bethge M, Brendel W (2020) Foolbox Native: fast adversarial attacks to benchmark the robustness of machine learning models in PyTorch, TensorFlow, and JAX. J Open Source Softw 5(53):2607. https://doi.org/10.21105/joss.02607

Reddy Y, Viswanath P, Reddy BE (2018) Semi-supervised learning: a brief review. Int J Eng Technol 7(1.8):81

Ren J et al (2019) Likelihood ratios for out-of-distribution detection. Adv Neural Inf Process Syst 32:14707–14718

Rodriguez D, Nayak T, Chen Y, Krishnan R, Huang Y (2022) On the role of deep learning model complexity in adversarial robustness for medical images. BMC Med Inform Decis Mak 22(Suppl 2):160

Roland T et al (2022) Domain shifts in machine learning based covid-19 diagnosis from blood tests. J Med Syst 46(5):23. https://doi.org/10.1007/s10916-022-01807-1

Rosa L, Silva F, Analide C (2022) Explainable artificial intelligence on smart human mobility: a comparative study approach. In: International symposium on distributed computing and artificial intelligence. Springer, New York, pp 91–101

Rouhani BD, Riazi MS, Koushanfar F (2018) Deepsecure: scalable provably-secure deep learning. In: Proceedings of the 55th annual design automation conference, pp 1–6

Roy S, Meena T, Lim SJ (2022) Demystifying supervised learning in healthcare 4.0: a new reality of transforming diagnostic medicine. Diagnostics 12(10):2549. https://doi.org/10.3390/diagnostics12102549

Roy S, Mehera R, Pal RK, Bandyopadhyay SK (2023a) Hyperparameter optimization for deep neural network models: a comprehensive study on methods and techniques. Innov Syst Softw Eng. https://doi.org/10.1007/s11334-023-00540-3

Roy A, Horstmann J, Ntoutsi E (2023b) Multi-dimensional discrimination in law and machine learning—a comparative overview. In: ACM international conference proceeding series, pp 89–100. https://doi.org/10.1145/3593013.3593979

Rudin C, Chen C, Chen Z, Huang H, Semenova L, Zhong C (2022) Interpretable machine learning: fundamental principles and 10 grand challenges. Stat Surv 16:1–85. https://doi.org/10.1214/21-SS133

Rueckert D, Schnabel JA (2020) Model-based and data-driven strategies in medical image computing. Proc IEEE 108(1):110–124. https://doi.org/10.1109/JPROC.2019.2943836

Ruiz N et al (2022) Simulated adversarial testing of face recognition models. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 4135–4145. https://doi.org/10.1109/CVPR52688.2022.00411

Sabour S, Frosst N, Hinton GE (2017) Dynamic routing between capsules. Adv Neural Inf Process Syst 3857–3867

Sakai T, Shimizu N (2019) Covariate shift adaptation on learning from positive and unlabeled data. In: 33rd AAAI conference on artificial intelligence, AAAI 2019, 31st innovative applications of artificial intelligence conference, IAAI 2019 and the 9th AAAI symposium on educational advances in artificial intelligence, EAAI 2019, pp 4838–4845. https://doi.org/10.1609/aaai.v33i01.33014838

Sarfraz A, Pansota MS, Fahal NAM, Sarfaraz A, Javed H (2021) Analytical solution of stochastic real-time power dispatch with large scale wind farms. Pak J Eng Technol 4(3):18–26. https://doi.org/10.51846/vol4iss3pp18-26

Sattigeri P, Hoffman SC, Chenthamarakshan V, Varshney KR (2019) Fairness GAN: generating datasets with fairness properties using a generative adversarial network. IBM J Res Dev 63(4–5):1–3. https://doi.org/10.1147/JRD.2019.2945519

Shaikh F et al (2021a) Current landscape of imaging and the potential role for artificial intelligence in the management of COVID-19. Curr Probl Diagn Radiol 50(3):430–435. https://doi.org/10.1067/j.cpradiol.2020.06.009

Shaikh F et al (2021b) Artificial intelligence-based clinical decision support systems using advanced medical imaging and radiomics. Curr Probl Diagn Radiol 50(2):262–267. https://doi.org/10.1067/j.cpradiol.2020.05.006

Sheehan B et al (2013) Informing the design of clinical decision support services for evaluation of children with minor blunt head trauma in the emergency department: a sociotechnical analysis. J Biomed Inform 46(5):905–913. https://doi.org/10.1016/j.jbi.2013.07.005

Shen J, Li W, Deng S, Zhang T (2021) Supervised and unsupervised learning of directed percolation. Phys Rev E 103(5):52140. https://doi.org/10.1103/PhysRevE.103.052140

Shi X et al (2022) Robust convolutional neural networks against adversarial attacks on medical images. Pattern Recognit 132:108923

Shi C, Veitch V, Blei DM (2021) Invariant representation learning for treatment effect estimation. In: Proceedings of machine learning research, PMLR, pp 1546–1555

Shi B, Hsu W-N, Mohamed A (2022) Robust self-supervised audio-visual speech recognition. arXiv preprint arXiv:2201.01763

Shibly KH, Hossain MD, Inoue H, Taenaka Y, Kadobayashi Y (2023) Towards autonomous driving model resistant to adversarial attack. Appl Artif Intell 37(1):2193461. https://doi.org/10.1080/08839514.2023.2193461

Shim M, Hwang HJ, Lee SH (2023) Toward practical machine-learning-based diagnosis for drug-naïve women with major depressive disorder using EEG channel reduction approach. J Affect Disord 338:199–206. https://doi.org/10.1016/j.jad.2023.06.007

Shimodaira H (2000) Improving predictive inference under covariate shift by weighting the log-likelihood function. J Stat Plan Inference 90(2):227–244. https://doi.org/10.1016/s0378-3758(00)00115-4

Silva SH, Najafirad P (2020) Opportunities and challenges in deep learning adversarial robustness: a survey. arXiv preprint arXiv:2007.00753

Singh G, Gehr T, Mirman M, Püschel M, Vechev M (2018) Fast and effective robustness certification. Adv Neural Inf Process Syst 10802–10813

Singh G, Gehr T, Püschel M, Vechev M (2019) Boosting robustness certification of neural networks. In: 7th international conference on learning representations, ICLR 2019

Singla S, Nushi B, Shah S, Kamar E, Horvitz E (2021) Understanding failures of deep networks via robust feature extraction. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 12848–12857. https://doi.org/10.1109/CVPR46437.2021.01266

Song H, Kim M, Park D, Shin Y, Lee JG (2022) Learning from noisy labels with deep neural networks: a survey. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2022.3152527

Sugimoto M, Hikichi S, Takada M, Toi M (2023) Machine learning techniques for breast cancer diagnosis and treatment: a narrative review. Ann Breast Surg 7:7–7. https://doi.org/10.21037/abs-21-63

Taguchi G (1995) Quality engineering (Taguchi methods) for the development of electronic circuit technology. IEEE Trans Reliab 44(2):225–229

Tanesini A (2021) The measure and mismeasure of the self. Oxford University Press, Oxford, pp 1–18. https://doi.org/10.1093/oso/9780198858836.003.0001

Tang X, Li Y, Sun Y, Yao H, Mitra P, Wang S (2020) Transferring robustness for graph neural network against poisoning attacks. In: WSDM 2020—Proceedings of the 13th international conference on web search and data mining, pp 600–608. https://doi.org/10.1145/3336191.3371851

Thomas AW, Ré C, Poldrack RA (2022) Interpreting mental state decoding with deep learning models. Trends Cogn Sci 26(11):972–986. https://doi.org/10.1016/j.tics.2022.07.003

Tian G (2023) OpenDP Programming Framework for Renyi Privacy Filters and Odometers. Accessed 7 Apr 2023. https://dash.harvard.edu/handle/1/37371627

Tian G. OpenDP Programming Framework for Renyi Privacy Filters and Odometers

Trewin S (2018) AI fairness for people with disabilities: point of view. arXiv preprint arXiv:1811.10670

Tsai MJ, Lin PY, Lee ME (2023) Adversarial attacks on medical image classification. Cancers 15(17):4228. https://doi.org/10.3390/cancers15174228

Tu S et al (2021) ModPSO-CNN: an evolutionary convolution neural network with application to visual recognition. Soft Comput 25(3):2165–2176. https://doi.org/10.1007/s00500-020-05288-7

Ullah A, Rehman SU, Tu S, Mehmood RM, Fawad, Ehatisham-Ul-haq M (2021) A hybrid deep CNN model for abnormal arrhythmia detection based on cardiac ECG signal. Sensors 21(3):1–13. https://doi.org/10.3390/s21030951

Ur Rehman S, Tu S, Huang Y, Yang Z (2016) Face recognition: A novel un-supervised convolutional neural network method. In: Proceedings of 2016 IEEE international conference of online analysis and computing science, ICOACS 2016, pp 139–144. IEEE. https://doi.org/10.1109/ICOACS.2016.7563066

ur Rehman S, Tu S, ur Rehman O, Huang Y, Magurawalage CMS, Chang CC (2018) Optimization of CNN through novel training strategy for visual classification problems. Entropy 20(4):290. https://doi.org/10.3390/e20040290

ur Rehman S et al (2019) Unsupervised pre-trained filter learning approach for efficient convolution neural network. Neurocomputing 365:171–190. https://doi.org/10.1016/j.neucom.2019.06.084

Urruty N, Tailliez-Lefebvre D, Huyghe C (2016) Stability, robustness, vulnerability and resilience of agricultural systems. a review. Agron Sustain Dev 36(1):1–15. https://doi.org/10.1007/s13593-015-0347-5

Vaishnavi P, Eykholt K, Rahmati A (2022) Transferring adversarial robustness through robust representation matching. In: Proceedings of the 31st USENIX security symposium, security 2022, pp 2083–2098

Van Biesebroeck J (2007) Robustness of productivity estimates. J Ind Econ 55(3):529–569

Verbraeken J, Wolting M, Katzy J, Kloppenburg J, Verbelen T, Rellermeyer JS (2020) A survey on distributed machine learning. ACM Comput Surv 53(2):1–33. https://doi.org/10.1145/3377454

Verma A, Rao K, Eluri V (2020) Regulating AI in public health: systems challenges and perspectives. ORF Occas Pap 261:1–46

Walonoski J, Scanlon R, Dowling C, Hyland M, Ettema R, Posnack S (2018) Validation and testing of fast healthcare interoperability resources standards compliance: data analysis. JMIR Med Inform 6(4):e10870. https://doi.org/10.2196/10870

Wang J (2021) Adversarial Examples in Physical World. In: IJCAI international joint conference on artificial intelligence. Chapman and Hall/CRC, Boca Raton, pp 4925–4926. https://doi.org/10.24963/ijcai.2021/694

Wang Y, Wang Y (2023) Robustness and reliability of machine learning systems: a comprehensive review engineering. Eng Open 1(2):90–95

Wang B et al (2021) Establishment of a knowledge-and-data-driven artificial intelligence system with robustness and interpretability in laboratory medicine. SSRN Electron J 4(5):2100204. https://doi.org/10.2139/ssrn.3928504

Wang X, Wang H, Yang D (2022a) Measure and improve robustness in NLP models: a survey. In: NAACL 2022—2022 conference of the north american chapter of the association for computational linguistics: human language technologies, proceedings of the conference, pp 4569–4586. https://doi.org/10.18653/v1/2022.naacl-main.339

Wang X et al (2022b) SurvMaximin: robust federated approach to transporting survival risk prediction models. J Biomed Inform 134:104176. https://doi.org/10.1016/j.jbi.2022.104176

Wang N, Cheng M, Ning K (2022c) Overcoming regional limitations: transfer learning for cross-regional microbial-based diagnosis of diseases. Gut 72(10):2004–2006. https://doi.org/10.1136/gutjnl-2022-328216

Wang M, Yang N, Gunasinghe DH, Weng N (2023a) On the robustness of ML-based network intrusion detection systems: an adversarial and distribution shift perspective. Computers 12(10):209. https://doi.org/10.3390/computers12100209

Wang D, Xiao H, Wu D (2023b) Application of unsupervised adversarial learning in radiographic testing of aeroengine turbine blades. NDT E Int 134:102766. https://doi.org/10.1016/j.ndteint.2022.102766

Weng WH (2020) Machine learning for clinical predictive analytics. Leveraging Data Science for Global Health, pp 199–217. https://doi.org/10.1007/978-3-030-47994-7_12

Wilson S, Fischer T, Sunderhauf N, Dayoub F (2023) Hyperdimensional feature fusion for out-of-distribution detection. In: Proceedings—2023 IEEE winter conference on applications of computer vision, WACV 2023, pp 2643–2653. https://doi.org/10.1109/WACV56688.2023.00267

Windmann A, Steude H, Niggemann O (2023) Robustness and generalization performance of deep learning models on cyber-physical systems: a comparative study. arXiv preprint arXiv:2306.07737

Woldeyohannes HD (2021) Review on 'Adversarial Robustness Toolbox (ART) v1. 5. x.': ART attacks against supervised learning algorithms case study

Wu Y, Zhang L, Wu X (2019a) Counterfactual fairness: unidentification, bound and algorithm. In: IJCAI international joint conference on artificial intelligence, pp 1438–1444. https://doi.org/10.24963/ijcai.2019/199

Wu L, Hsieh CJ, Li S, Sharpnack J (2019b) Stochastic shared embeddings: data-driven regularization of embedding layers. Adv Neural Inf Process Syst 32

Xie C, Wu Y, Van Der Maaten Y, Yuille AL, He K (2019) Feature denoising for improving adversarial robustness. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 501–509. https://doi.org/10.1109/CVPR.2019.00059

Xie C, Tan M, Gong B, Wang J, Yuille AL, Le QV (2020) Adversarial examples improve image recognition. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 816–825. https://doi.org/10.1109/CVPR42600.2020.00090

Xing F, Silosky M, Ghosh D, Chin BB (2023) Location-aware encoding for lesion detection in ^{68} Ga-DOTATATE positron emission tomography images. IEEE Trans Biomed Eng. https://doi.org/10.1109/TBME.2023.3297249

Xiong L, Liu X, Liu Y, Zhuo F (2022) Modeling and stability issues of voltage-source converter-dominated power systems: a review. CSEE J Power Energy Syst 8(6):1530–1549. https://doi.org/10.17775/CSEEJPES.2020.03590

Xu J, Chen J, You S, Xiao Z, Yang Y, Lu J (2021a) Robustness of deep learning models on graphs: a survey. AI Open 2:69–78. https://doi.org/10.1016/j.aiopen.2021.05.002

Xu M, Zhang T, Li Z, Liu M, Zhang D (2021b) Towards evaluating the robustness of deep diagnostic models by adversarial attack. Med Image Anal 69:101977. https://doi.org/10.1016/j.media.2021.101977

Xu M, Zhang T, Zhang D (2022) Medrdf: a robust and retrain-less diagnostic framework for medical pretrained models against adversarial attack. IEEE Trans Med Imaging 41(8):2130–2143

Xue C, Dou Q, Shi X, Chen H, Heng PA (2019) Robust learning at noisy labeled medical images: APPLIED to skin lesion classification. In: Proceedings—international symposium on biomedical imaging, pp 1280–1283. IEEE. https://doi.org/10.1109/ISBI.2019.8759203

Xue C, Deng Q, Li X, Dou Q, Heng PA (2020) Cascaded robust learning at imperfect labels for chest X-ray segmentation. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer, New York, pp 579–588. https://doi.org/10.1007/978-3-030-59725-2_56

Xue C, Yu L, Chen P, Dou Q, Heng PA (2022) Robust medical image classification from noisy labeled data with global and local representation guided co-training. IEEE Trans Med Imaging 41(6):1371–1382. https://doi.org/10.1109/TMI.2021.3140140

Yadav RK, Singh P, Kashtriya P (2022) Diagnosis of breast cancer using machine learning techniques—a survey. Procedia Comput Sci 218:1434–1443. https://doi.org/10.1016/j.procs.2023.01.122

Yan JN, Gu Z, Lin H, Rzeszotarski JM (2020) Silva: interactively assessing machine learning fairness using causality. In: Proceedings of the 2020 chi conference on human factors in computing systems, pp 1–13

Yang S, Zhou X (2022) PGS-server: accuracy, robustness and transferability of polygenic score methods for biobank scale studies. Brief Bioinform 23(2):bbac039. https://doi.org/10.1093/bib/bbac039

Yao S, Kang Q, Zhou MC, Rawa MJ, Abusorrah A (2023) A survey of transfer learning for machinery diagnostics and prognostics. Artif Intell Rev 56(4):2871–2922. https://doi.org/10.1007/s10462-022-10230-4

Ye Q et al (2022) Robust weakly supervised learning for COVID-19 recognition using multi-center CT images. Appl Soft Comput 116:108291. https://doi.org/10.1016/j.asoc.2021.108291

Yi R, Tang L, Tian Y, Liu J, Wu Z (2023) Identification and classification of pneumonia disease using a deep learning-based intelligent computational framework. Neural Comput Appl 35(20):14473–14486

Yuan Y, Wei J, Huang H, Jiao W, Wang J, Chen H (2023) Review of resampling techniques for the treatment of imbalanced industrial data classification in equipment condition monitoring. Eng Appl Artif Intell 126:106911

Zamir AR et al (2020) Robust learning through cross-task consistency. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 11194–11203. https://doi.org/10.1109/CVPR42600.2020.01121

Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer, New York, pp 818–833. https://doi.org/10.1007/978-3-319-10590-1_53

Zhang X, Su H, Yang L, Zhang S (2015) Fine-grained histopathological image analysis via robust segmentation and large-scale retrieval. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 5361–5368. https://doi.org/10.1109/CVPR.2015.7299174

Zhang C, Bengio S, Hardt M, Recht B, Vinyals O (2021) Understanding deep learning (still) requires rethinking generalization. Commun ACM 64(3):107–115. https://doi.org/10.1145/3446776

Zhang Z, Yang Z, Yau DKY, Tian Y, Ma J (2023a) Data security of machine learning applied in low-carbon smart grid: a formal model for the physics-constrained robustness. Appl Energy 347:121405. https://doi.org/10.1016/j.apenergy.2023.121405

Zhang S et al (2023b) Robust failure diagnosis of microservice system through multimodal data. IEEE Trans Serv Comput. https://doi.org/10.1109/TSC.2023.3290018

Zhao Y, Gao D, Yao Y, Zhang Z, Mao B, Yao X (2023) Robust deep learning models against semantic-preserving adversarial attack. In: Proceedings of the international joint conference on neural networks, vol 2023. https://doi.org/10.1109/IJCNN54540.2023.10191198

Zhong X et al (2019) Deep transfer learning-based prostate cancer classification using 3 Tesla multi-parametric MRI. Abdomin Radiol 44(6):2030–2039. https://doi.org/10.1007/s00261-018-1824-5

Zhou ZH, Wu J, Tang W (2002) Ensembling neural networks: Many could be better than all. Artif Intell 137(1–2):239–263. https://doi.org/10.1016/S0004-3702(02)00190-X

Zhou SK et al (2021) A review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises. Proc IEEE 109(5):820–838. https://doi.org/10.1109/JPROC.2021.3054390

Zhu Q, Başar T (2015) Game-theoretic methods for robustness, security, and resilience of cyberphysical control systems: games-in-games principle for optimal cross-layer resilient control systems. IEEE Control Syst 35(1):46–65. https://doi.org/10.1109/MCS.2014.2364710

Zhu H, Shi J, Wu J (2019) Pick-and-learn: automatic quality evaluation for noisy-labeled image segmentation. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer, New York, pp 576–584. https://doi.org/10.1007/978-3-030-32226-7_64

Zhu Z, Zhang L, Pei K, Chen S (2023) A robust and lightweight voice activity detection algorithm for speech enhancement at low signal-to-noise ratio. Digital Signal Process 141:104151. https://doi.org/10.1016/j.dsp.2023.104151

Żurański AM, Martinez Alvarado JI, Shields BJ, Doyle AG (2021) Predicting reaction yields via supervised learning. Acc Chem Res 54(8):1856–1865. https://doi.org/10.1021/acs.accounts.0c00770