Review

Not peer-reviewed version

Trustworthy AI in Digital Health: A Comprehensive Review of Robustness and Explainability

Abdullah Mamun [*] , Shovito Barua Soumma , Hassan Ghasemzadeh

Posted Date: 29 July 2025

doi: 10.20944/preprints202507.2387.v1

Keywords: machine learning; trustworthy AI; digital health; explainable AI; robustness; review article

Review

Trustworthy AI in Digital Health: A Comprehensive Review of Robustness and Explainability

Abdullah Mamun [1,*], Shovito Barua Soumma [2] and Hassan Ghasemzadeh [2]

[1] School of Computing and Augmented Intelligence, Arizona State University
[2] College of Health Solutions, Arizona State University
[*] Correspondence: a.mamun@asu.edu

Abstract

Ensuring trust in AI systems is essential for the safe and ethical integration of machine learning systems into high-stakes domains such as digital health. Key dimensions, including robustness, explainability, fairness, accountability, and privacy, need to be addressed throughout the AI lifecycle, from problem formulation and data collection to model deployment and human interaction. While various contributions address different aspects of trustworthy AI, a focused synthesis on robustness and explainability, especially tailored to the healthcare context, remains limited. This review addresses that need by organizing recent advancements into an accessible framework, highlighting both technical and practical considerations. We present a structured overview of methods, challenges, and solutions, aiming to support researchers and practitioners in developing reliable and explainable AI solutions for digital health. This review article is organized into three main parts. First, we introduce the pillars of trustworthy AI and discuss the technical and ethical challenges, particularly in the context of digital health. Second, we explore application-specific trust considerations across domains such as intensive care, neonatal health, and metabolic health, highlighting how robustness and explainability support trust. Lastly, we present recent advancements in techniques aimed at improving robustness under data scarcity and distributional shifts, as well as explainable AI methods ranging from feature attribution to gradient-based interpretations and counterfactual explanations. This paper is further enriched with detailed discussions of the contributions toward robustness and explainability in digital health, the development of trustworthy AI systems in the era of LLMs, and various evaluation metrics for measuring trust and related parameters such as validity, fidelity, and diversity.

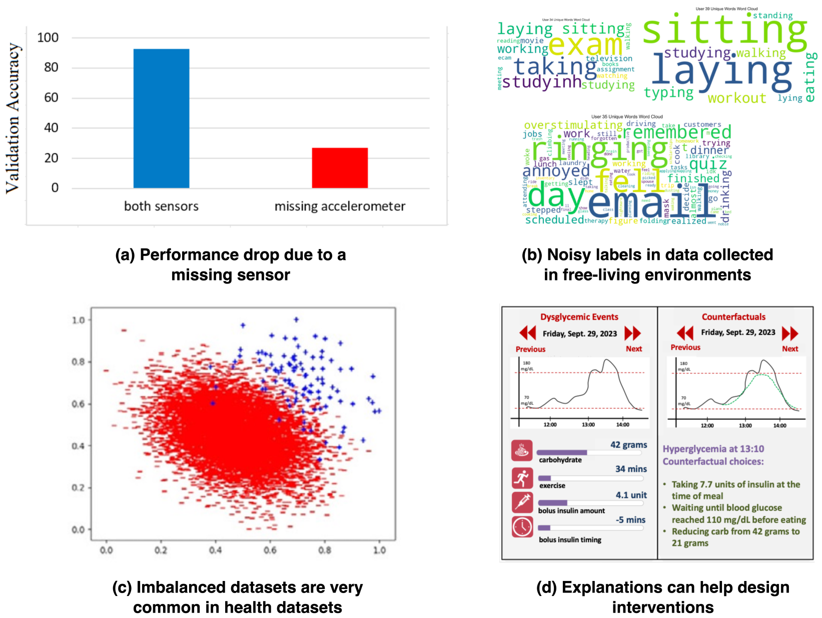
Keywords: machine learning; trustworthy AI; digital health; explainable AI; robustness; review article

1. Introduction

Trustworthy AI focuses on ensuring that AI systems are reliable, transparent, and accountable. By emphasizing fairness, safety, and explainability, it fosters user confidence and promotes ethical decision-making. This approach is essential for ensuring that AI-driven solutions are both effective and equitable, ultimately contributing to better outcomes in a variety of applications across industries.

The National Institute of Standards and Technology (NIST) defines trustworthy AI through a set of core characteristics that include validity and reliability, safety, security and resiliency, accountability, transparency, explainability, privacy, and fairness [4]. These components serve as foundational pillars for developing standards, guidelines, and practices to ensure AI systems are aligned with ethical and societal values. The High-Level Expert Group on AI presented the Ethics Guidelines for Trustworthy Artificial Intelligence [2] in April 2019. The guidelines define trustworthy AI as systems that are lawful, ethical, and robust, both technically and socially. They outline seven key requirements for trustworthy AI: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity and fairness, societal and environmental well-being, and accountability. These principles emphasize empowering human decision-making, ensuring security and reliability, respecting privacy, promoting transparency, avoiding bias, fostering inclusivity and sustainability, and implementing mechanisms for responsibility and redress.

In this article, we discuss in detail two of these components: robustness and explainability. Robustness and explainability are two important characteristics of a machine learning system. Consider a multisensor system where a human activity recognition system is trained with both accelerometer and gyroscope data. During inference, if one of the two sensors (e.g., the accelerometer) is unavailable, due to hardware limitations, being disabled to conserve power, or connectivity issues, the prediction accuracy can degrade drastically (Figure 1a).



(a) Performance drop due to a missing sensor

(b) Noisy labels in data collected in free-living environments

(c) Imbalanced datasets are very common in health datasets

(d) Explanations can help design interventions

Figure 1. Examples implicating the importance of robustness and explainability in AI systems.

A challenge often faced with datasets collected in free-living environments is the presence of noisy labels. In one of our user studies at Arizona State University, we found that different participants expressed activities in different expressions (Figure 1b). E.g., some people used the word 'typing' and some other people may write 'studying' for the same activity (e.g., working on a writing assignment). Furthermore, there can be typing mistakes while recording behavioral labels, e.g., 'studyinh' when intending to write 'studying'. A robust machine learning system would have to overcome these challenges.

Imbalanced datasets are quite common in certain health datasets, for example, diagnosing a rather uncommon disease. When training with a highly imbalanced dataset (Figure 1c), a machine learning model may get biased toward the majority class. To overcome the issue of imbalanced classes and small datasets, data balancing methods and data generation tools can improve robustness.

Finally, in intelligent digital health systems, it is more important than usual to provide reasons and explanations on predictions made by a machine learning system, so that interventions can be safely applied. For example, in Figure 1d, we present a machine learning system that can detect abnormal blood glucose events such as hyperglycemia and hypoglycemia. However, detection alone may not be sufficiently insightful to prevent abnormal events because actionable feedback can provide with model-based reasoning or insightful feedback about how to prevent an undesired outcome. An explanation with a counterfactual explanation or alternative scenarios with proposed changes to improve a health outcome can overcome that limitation. Hence, it is important for an intelligent digital health system to be able to provide meaningful explanations.

Significant progress has been made in applying machine learning and artificial intelligence to healthcare challenges, addressing issues such as sensor failure, treatment adherence, neonatal risk prediction, and physical activity forecasting. These efforts strive to bridge the gap between clinical requirements and computational methods by integrating advanced deep learning techniques with domain-specific insights to improve the reliability, personalization, and explainability of health monitoring systems.

One study [89] addresses the issue of missing data in sensor-based systems caused by sensor malfunctions or power limitations. By proposing an algorithm to reconstruct missing input data, the research demonstrated significant improvements in activity classification accuracy on benchmark datasets like MotionSense [81,82], MobiAct [128], and MHEALTH [19]. This contribution underscores the importance of building robust systems capable of maintaining reliable performance in real-world scenarios.

In the field of neonatal health, related work [85,86] explores the use of deep learning combined with counterfactual explanations to predict adverse labor outcomes. This approach addresses challenges such as class imbalance and explainability, aiming to reduce risks during childbirth and improve neonatal health outcomes. Similarly, research on treatment adherence prediction demonstrates the effectiveness of personalized models in forecasting adherence patterns, providing a framework for designing timely interventions to support patient care.

Efforts in multimodal data integration have also gained attention, particularly in the context of adaptive lifestyle interventions. By combining data from wearable devices, app engagement, and dietary habits, these studies [87,88] enable the prediction of physical activity levels and postprandial glucose trends. The integration of explainable AI techniques and counterfactual reasoning in these systems supports informed decision-making for both users and healthcare providers.

While several prior review articles have surveyed different aspects of trustworthy AI in isolation, there remains a gap in systematically analyzing both robustness and explainability dimensions together in the context of digital health. This review uniquely bridges that gap by providing a comprehensive discussion of robustness and explainability, grounded in the practical demands of healthcare systems. It also critically examines existing evaluation metrics and proposes a taxonomy to guide future research at the intersection of reliable and explainable AI in digital health applications. This paper aims to balance breadth and depth by offering a concise overview of trustworthy AI, with a detailed focus on two key aspects: robustness and explainability.

2. Background

2.1. Prior Reviews on Trustworthy AI

The landscape of trustworthy AI research has been significantly enriched by numerous review efforts. Kaur et al. [65] provide a comprehensive analysis of the diverse requirements for trustworthy AI, including fairness, explainability, accountability, and reliability, offering insights into approaches to mitigate risks and improve societal acceptance. Kumar et al. [71] emphasize the ethical foundations of trustworthy AI, detailing how ethics can be embedded in system design and development, with a focus on practical applications in smart cities. Similarly, Kaur et al. [64] consolidate approaches for trustworthy AI based on the European Union's principles, presenting a structured overview for achieving reliable systems. Liu et al. [79] delve into the alignment of large language models (LLMs), identifying key dimensions of trustworthiness such as safety, fairness, and adherence to social norms, while also conducting empirical evaluations to highlight alignment challenges. Lastly, Fehr et al. [49] assess the transparency of CE-certified [1] medical AI products in Europe, revealing significant documentation gaps and calling for stricter legal requirements to ensure safety and ethical compliance in medical AI. These works collectively advance the understanding and implementation of trustworthy AI across various domains.

Continuing the exploration of trustworthy AI, Li et al. [74] propose a unified framework that integrates fragmented approaches to AI trustworthiness across the system lifecycle, addressing challenges such as robustness, fairness, and privacy preservation. Their comprehensive guide identifies actionable strategies for practitioners and policymakers to enhance the trustworthiness of AI systems. In the healthcare sector, Albahri et al. [6] present a systematic review of the trustworthiness and explainability of AI applications, highlighting the significant transparency and bias risks present in current methodologies. They propose guidelines and a detailed taxonomy for integrating explainable AI (XAI) methods into healthcare systems. Verma et al. [129] focus specifically on counterfactual explanations, offering a rubric to evaluate counterfactual algorithms and identifying promising research directions for this critical aspect of model explainability. Saraswat et al. [112] expand on the role of XAI in Healthcare 5.0, providing a solution taxonomy and case studies that demonstrate the potential of explainable AI techniques in improving operational efficiency and patient trust through privacy-preserving federated learning frameworks. Lastly, Band et al. [18] critically review XAI methodologies such as SHAP [80], LIME [110], and Grad-CAM [115], emphasizing their applicability to healthcare challenges like tumor segmentation and disease diagnosis. They offer a synthesis of usability and reliability studies, underscoring the transformative potential of XAI in medical decision-making. The review paper by Ojha et al. (2025) [102] explores the critical role of uncertainty in AI for healthcare, emphasizing the need for improved uncertainty quantification methods and addressing the gaps in understanding user perceptions of uncertainty and trustworthiness. Together, these works deepen our understanding of the pathways to achieving trustworthy and explainable AI systems, particularly in high-stakes domains like healthcare. A brief summary of the above papers is presented in Table 1.

Table 1. Prior review works in the area of Trustworthy AI and the contribution of this paper.

Publication	Application	Brief Description
Kumar et al. (2020) [71]	Smart Cities	Highlights the embedding of ethical foundations in AI system design and development, focusing on practical applications in smart cities.
Verma et al. (2020) [129]	Counterfactual Explanations	Offers a rubric to evaluate counterfactual algorithms, identifying research directions for this critical aspect of model explainability.
Kaur et al. (2021) [64]	General AI systems	Consolidates approaches for trustworthy AI based on the European Union's principles, presenting a structured overview for achieving reliable systems.
Kaur et al. (2022) [65]	General AI systems	Provides a comprehensive analysis of the requirements for trustworthy AI, including fairness, explainability, accountability, and reliability, with approaches to mitigate risks and improve societal acceptance.
Saraswat et al. (2022) [112]	Healthcare 5.0 [90]	Provides a taxonomy and case studies demonstrating the potential of explainable AI in improving operational efficiency and patient trust using privacy-preserving federated learning frameworks.
Liu et al. (2023) [79]	Large Language Models (LLMs)	Explores the alignment of LLMs with safety, fairness, and social norms, providing empirical evaluations to highlight alignment challenges.
Band et al. (2023) [18]	Medical Decision-Making	Critically reviews XAI methodologies like SHAP [80], LIME [110], and Grad-CAM [115], emphasizing their usability and reliability for applications like tumor segmentation and disease diagnosis.
Albahri et al. (2023) [6]	Healthcare AI	Systematically reviews trustworthiness and explainability in healthcare AI, highlighting transparency and bias risks and proposing a taxonomy for integrating XAI methods into healthcare systems.
Li et al. (2023) [74]	General AI systems	Proposes a unified framework integrating fragmented approaches to AI trustworthiness, addressing challenges such as robustness, fairness, and privacy preservation.
Fehr et al. (2024) [49]	Medical AI in Europe	Assesses the transparency of CE-certified medical AI products, revealing significant documentation gaps and calling for stricter legal requirements to ensure safety and ethical compliance.
Ojha et al. (2024) [102]	Healthcare AI	Focuses on one specific area of trustworthiness, that is, uncertainty.
This paper	Digital Health	Provides a comprehensive review and discussion of the aspects of trustworthy AI in digital health systems, with a focus on robustness and explainability.

2.2. Branches of Trustworthy AI

Trustworthy AI can be categorized into five interrelated branches, each addressing distinct aspects of ethical and reliable AI systems. Robustness ensures that systems perform reliably even under varied or adverse conditions, such as noisy inputs or missing data. Explainability makes AI models and their decisions understandable, allowing users to comprehend and trust the decisions made by the models. Fairness focuses on preventing biases and ensuring equitable outcomes for all users. Privacy safeguards sensitive user data throughout the AI lifecycle, maintaining confidentiality and security. Finally, accountability establishes mechanisms for responsibility and oversight, ensuring that AI systems operate within ethical and legal boundaries. Together, these branches provide a comprehensive framework for building AI systems that inspire trust and confidence among users.

2.3. Challenges in Different Phases of a Machine Learning System

Challenges in Trustworthy AI arise at various stages of the machine learning pipeline. During the problem definition phase, it is essential to clearly outline the scope and objectives while accounting for ethical and societal implications. Data collection often involves issues such as noisy data, noisy labels, and missing data, which can compromise the quality of training datasets. In feature selection, managing multimodality becomes crucial to effectively integrate heterogeneous data sources. The model training phase encounters difficulties like class imbalance, small datasets, and the risk of overfitting, while model evaluation must address underfitting, overfitting, and performance gaps due to distribution shifts. Finally, the inference stage brings challenges such as missing channels or modalities and ensuring explainability during deployment. Addressing these challenges holistically is key to creating robust and trustworthy AI systems. We present an overview of these challenges in Figure 2.

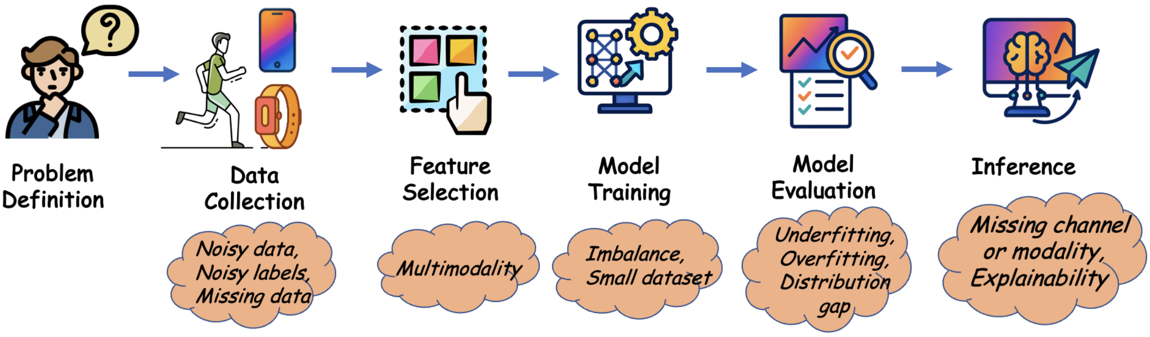


Figure 2. Challenges in machine learning systems in different phases of the development lifecycle.

2.4. Designing Robust Machine Learning Models

Robustness is a fundamental aspect of Trustworthy AI, ensuring that models maintain reliable performance across a wide range of conditions [2,74]. For noisy data, noisy labels, and missing data, techniques such as denoising autoencoders [130], convolutional networks [89], and fuzzy c-means clustering [14] have proven effective in reconstructing or imputing missing information. Masked autoencoders [57], for instance, use random masking and self-supervised learning to predict and reconstruct missing patches in input data, enhancing model reliability. Robustness also extends to multimodal data integration and addressing class imbalance. Multimodal deep learning methods combine data from diverse sources, improving system performance and resilience. Data balancing techniques, such as re-weighting or re-sampling, help tackle class imbalance issues, especially in small datasets [85,116,142].

Several contributions highlight advancements in robustness. For example, systems designed to mitigate sensor failure or system designed to be able to operate with low energy [61,62,89]. Robustness for noisy labels has improved activity recognition in free-living environments [7,92,123], while robustness for multimodal data has applications in activity forecasting [87]. Additionally, addressing class imbalance has led to innovations in neonatal health applications, reducing overfitting and improving

doi:10.20944/preprints202507.2387.v1

outcomes in small datasets. A summary of different contributions in the field developing robust machine learning solutions is provided in Table 2 and Table 3.

Table 2. Summary of robustness methods for missing data, noisy data, and related challenges and their applications.

Name	Authors	Year	Method	Original Application
Denoising Autoencoder [130]	Vincent et al.	2008	Uses noisy data as input and the corresponding clean data as output to train	Denoising images
Masked Autoencoder [57]	He et al.	2021	The positions of the missing patches are used to improve reconstruction.	Recreates missing patches in images
Generalized Butterworth Filter [114]	Selesnick and Burrus	1998	A Butterworth filter is a signal filter with a maximally flat passband response, minimizing ripples and ensuring smooth attenuation to the stopband.	Reduces noise in time-series data
Missing data imputation with Fuzzy c-means clustering [14]	Aydilek and Arslan	2013	A hybrid approach combines fuzzy c-means clustering, support vector regression, and a genetic algorithm to estimate missing values.	Improves imputation performance, outperforming zero imputation and other traditional methods.
ActiLabel [7]	Alinia et al.	2020	Dependency graphs to capture structural similarities and map activity labels between domains.	Improve activity recognition's usability and performance
Missing Sensor Data Reconstruction Algorithm [89]	Mamun et al.	2022	Proposes an algorithm to reconstruct missing input data in sensor-based health monitoring systems, improving prediction accuracy on multiple activity classification benchmarks.	Reconstructing missing sensor data
CIM: Clustering-based Energy-Efficient Data Imputation Method [62]	Hussein and Bhat	2023	CIM detects missing sensors, predicts their clusters for imputation using a mapping table, and determines activities through imputation-aware classification or a reliable activity classifier.	Detecting and imputing missing sensor data
Cross-Domain Conditional Diffusion Models for Time Series Imputation [143]	Zhang et al.	2025	Introduces a diffusion-based method for cross-domain time series imputation that handles domain shifts and missing data via spectral interpolation and consistency alignment.	Time series data

Table 3. Summary of robustness methods for multimodality, class imbalance, overfitting, and applications.

Name	Authors	Year	Method	Data Type
Multimodal deep learning [100]	Ngiam et al.	2011	Bimodal deep autoencoder and Restricted Boltzmann Machine to outperform unimodal classifiers	Video, Audio
SMOTE [31]	Chawla et al.	2002	K-Nearest-neighbor based synthetic data generation	Tabular, Transformed features
AdaSYN [56]	He et al.	2008	Synthetic data generator with higher priority near the decision boundary	Tabular, Transformed features
CTGAN [138]	Xu et al.	2019	A modified GAN for tabular data with different processing for categorical and numerical features	Tabular
Binary Imbalanced Data Classification [142]	Zhai et al.	2021	GAN and discarding of batch data based on silhouette score	Tabular
AIMEN and R-AIMEN [85]	Mamun et al.	2024	CTGAN based data balancing with or without restrictions on similarity and type of the generated data	Tabular
MetaBoost [116]	Shah et al.	2025	Hybrid data balancing method that creates batches of synthetic data from weighted combinations of multiple balancing methods	Tabular
Dropout [124]	Srivastava et al.	2014	Randomly disables a number of neurons of the previous layer during training time. In the inference time, the weights are adjusted to maintain consistency.	Any data type
Knowledge-guided transformer for forecasting [107]	Qi et al.	2021	Uses future knowledge (future promotions) to improve performance of forecasting	Time-series

3. Application-Specific Trust Concerns

While general principles of trustworthy AI, such as fairness, explainability, and robustness, remain applicable across domains, their manifestation in healthcare is often domain-specific. This section explores how trust concerns arise and are addressed within distinct areas of digital health. By contextualizing trust within each domain, we highlight the nuances and practical challenges of building AI systems that clinicians, patients, and regulators can rely on.

3.1. Radiology and Medical Imaging

Medical imaging has been a fertile ground for deep learning-based diagnostics [22,55,144]. However, black-box models pose concerns regarding explainability and safety. Trustworthiness in this domain has been approached through saliency maps, Grad-CAM [115], counterfactual visualization [127], and attention-based explainability [37]. Techniques like domain adaptation and test-time augmentation are used to enhance robustness and consistency [52,131].

3.2. Cardiovascular Health

AI in cardiology is used in arrhythmia detection and atrial fibrillation prediction from ECG signals [9,51]. Trust hinges on robust signal modeling, explainability, and alignment with clinical workflows. Wearable ECG applications demand calibrated outputs and minimal false alarms to avoid unnecessary interventions.

Wearables generate continuous data streams for health tracking, yet present challenges in sensor fidelity, environmental noise, and user variability [123]. Trustworthy modeling in this domain emphasizes online learning, anomaly detection, as well as designing a system so that it can handle unexpected scenarios such as sensor failure or unavailability during inference [89].

3.3. Metabolic Health

Applications in metabolic health include blood glucose forecasting [42,48,83], lifestyle personalization [11,12], and insulin resistance monitoring. These models typically integrate data from CGMs, diet logs, and wearables [83]. Key challenges include user variability, missing data, class imbalance, and context sensitivity. Explainability techniques, e.g., counterfactual explanations and privacy-preserving learning frameworks (e.g., federated learning) are often employed to enhance trust [10,11,32,83,116].

3.4. Neonatal Health and Pediatrics

AI in pediatric and neonatal care faces the dual challenge of data scarcity and high stakes [85, 86,106,118]. Explainability is crucial, as caregivers and providers demand transparency for decisions affecting vulnerable populations. Methods include what-if analysis, causal feature attribution, and generalizability testing across institutions [85].

3.5. Mental Health and Addiction Recovery

AI systems are increasingly used to monitor mental health and support addiction recovery in free-living environments. To address the challenge of limited labels in substance use detection, recent frameworks leverage self-supervised learning on wearable sensor data, enabling accurate detection of cannabis use with minimal supervision [15]. Stress detection in individuals recovering from Alcohol Use Disorder has similarly benefited from optimized sensor selection and context modeling, revealing skin conductance as a key signal [111]. A broader review of ML in addiction research highlights the use of brain imaging, behavioral phenotyping, and ensemble models like random forests for early screening and monitoring [35]. These approaches emphasize trust through robust learning under label scarcity, personalization, and clinically relevant explanations.

3.6. Brain Health

AI methods for brain health, especially neurodegenerative diseases like Alzheimer's and Parkinson's face distinct trust challenges due to the sensitive, subjective, and progressive nature of these conditions [76,97,136]. In brain health, models using mobile sensing and behavioral data raise privacy concerns and benefit from causal modeling and clinician-in-the-loop systems. To ensure safety and transparency, approaches such as federated learning, uncertainty estimation, and saliency-based explanations are being adopted [122].

3.7. Intensive Care and Monitoring

AI is increasingly used in ICUs for early prediction of sepsis, patient deterioration, and resource management. A deep learning model trained on over 130,000 ICU admissions across multiple countries demonstrated strong generalization and detected sepsis 3.7 hours before onset [95]. External validation and low false alarm rates contributed to clinical trust. Reviews emphasize the importance of explainability, ethical safeguards, and robust modeling under data variability [125].

3.8. Public Health and Epidemiology

In public health, AI supports disease surveillance, epidemic forecasting, and community-level interventions. Studies highlight the promise of citizen science, predictive analytics, and participatory tools to improve health equity [67,103]. Responsible AI frameworks are essential to address fairness, transparency, and ethical risks in real-time surveillance systems [25]. Trust relies on inclusive design, societal accountability, and transparent public engagement.

4. Advances in Trustworthy AI for Digital Health

4.1. Label Scarcity and Data-Efficient Learning

Trustworthy AI in digital health emphasizes the development of robust and explainable systems to address unique challenges in healthcare. For example, label scarcity in free-living environments complicates supervised learning for detecting health-related behaviors. The CUDLE framework addresses this by leveraging self-supervised learning to identify cannabis consumption moments using wearable sensors. By training on augmented data for robust feature extraction and fine-tuning with minimal labeled data, CUDLE achieves higher accuracy compared to traditional supervised methods, even with 75% fewer labels [15]. Similarly, clinical speech AI must account for limited data availability and overfitting risks. Integrating insights from speech science, explainable models, and robust validation frameworks can mitigate these issues and accelerate the translation of speech biomarkers for clinical use [21]. These approaches demonstrate the importance of data-efficient and explainable methodologies in healthcare AI.

4.2. Forecasting and Personalized Interventions

Probabilistic forecasting in public health further underscores the need for trustworthy AI. During the COVID-19 pandemic, ensemble models synthesizing predictions from multiple groups consistently outperformed individual models in forecasting mortality rates. This collaborative effort highlighted the importance of active coordination between public health agencies, academia, and industry for developing reliable modeling capabilities under real-world constraints [39]. In digital health interventions, personalized approaches such as smartphone and text-message-based systems for managing type 2 diabetes showed significant improvements in glycemic control compared to website-based interventions [132]. The higher reach and uptake of these modalities suggest their practicality for real-world deployment, though optimizing delivery mechanisms remains crucial [96].

4.3. Self-Supervised Learning and Cross-Domain Generalization

Self-supervised learning also plays a transformative role in medical AI [15], enabling models to learn from large-scale unannotated data across diverse modalities such as medical images and bioelectrical signals [60,104]. These methods address challenges like limited annotated datasets and biased data collection, facilitating the development of scalable and generalizable AI systems [70]. Additionally, inherent sensor redundancies have been exploited to enhance anomaly detection in sensor-based systems, demonstrating the potential of cross-domain techniques for improving trustworthiness and robustness in digital health applications [58]. Together, these advancements highlight a growing emphasis on creating AI systems that are not only accurate but also ethical, transparent, and adaptable to diverse healthcare contexts.

4.4. Robustness and Clinical Utility

Several additional works have addressed challenges in trustworthy AI for digital health, such as sensor failure and imbalanced datasets in healthcare applications. For example, efforts in AIMEN [85] and medication adherence forecasting [84] have focused on building robust, explainable systems that aim to improve reliability and clinical utility, particularly in the contexts of neonatal health and treatment adherence. These works contribute to enhancing the trustworthiness and effectiveness of AI applications in healthcare.

4.5. Scientific Discovery and Human Oversight

A recent example of progress toward trustworthy AI is Google's Co-Scientist system [50], which introduces a multi-agent framework for automated scientific discovery while maintaining human oversight. Designed to generate, critique, and refine hypotheses grounded in literature, the system exemplifies transparency and accountability through its debate-style evaluation and citation-based reasoning. Importantly, the inclusion of scientists in the loop ensures explainability and alignment with domain knowledge, supporting responsible deployment in high-stakes fields such as biomedical research. While broader ethical considerations remain underexplored, the Co-Scientist represents a promising step toward collaborative, robust, and explainable AI in scientific and healthcare contexts.

5. Explainability in Machine Learning

The growing importance of explainability in machine learning (ML) and artificial intelligence (AI) has led to the development of several methods to make complex models more transparent and their predictions understandable to users. These techniques, ranging from feature importance scores to counterfactual explanations, help bridge the gap between high-performing but opaque models and their practical, responsible deployment in real-world applications, especially in critical fields like healthcare. A summary of different explainable AI methods is provided in Table 4. Below, we review various explainable AI approaches, with a particular emphasis on counterfactual explanations and their applications.

Table 4. An overview of a few different explainable AI methods.

Name	Authors	Year	Method	Data Type
LIME [110]	Ribeiro et al.	2016	Surrogate interpretable model for estimating effect.	Tabular, Text, Image
Shapley [117]	L. Shapley	1953	Measure effect of a feature on different coalitions of all other features.	Tabular, Time-series, Image
SHAP [80]	Lundberg and Lee	2017	Efficient approximation based on Shapley (practical when the number of features is large).	Tabular, Time-series, Image
GradCAM [115]	Selvaraju et al.	2017	Generates class-specific heatmaps by calculating the gradients of the target class w.r.t. feature maps.	Image
Layer-wise Relevance Propagation [24]	Binder et al.	2016	Assigns relevance scores to input features by propagating the output decision backwards.	Image, Tabular, Text
Integrated Gradients [126]	Sundararajan et al.	2017	A path-based attribution method that assigns feature importance by accumulating gradients from a baseline to the input.	Image
NICE [26]	Brughmans et al.	2023	Counterfactual explanation based on a modified nearest unlike neighbor.	Tabular
DiCE [98]	Mothilal et al.	2020	Counterfactual explanations that optimize on validity, proximity, diversity, and sparsity.	Tabular
Semi-factual explanation [13]	Aryal and Keane	2024	Finds alternate feature values on the same class. Can be counterfactual (CF)-free or CF-guided.	Tabular
Multi-Objective Counterfactuals [41]	Dandl et al.	2020	Counterfactual method satisfying 4 objective functions: validity, distance, sparsity, distribution sanity.	Tabular

5.1. Explainable AI Methods in ML Models

Explainable AI focuses on making machine learning models transparent and ensuring that users can understand the reasoning behind predictions. Methods such as Shapley values [117] provide feature importance scores, which help users discern the factors influencing model predictions. Another prominent technique is counterfactual explanations, which explore how altering input features would lead to different outcomes, offering actionable insights for decision-making. Tools like DiCE [98] generate diverse counterfactuals, enhancing both explainability and usability in practical settings. Further advancements include NICE (Nearest Instance Counterfactual Explanations) [26], which produces intuitive and transparent explanations by finding the nearest unlike neighbor and then modifying it. These innovations have been applied in critical domains, such as neonatal health, where NICE and SHAP were used to explain health predictions [85], offering clinicians a better understanding of risk factors and outcomes. Similarly, DiCE and other methods for counterfactual explanations have been applied to multimodal hyperglycemia prediction, improving interpretability and reducing the risks associated with glucose management [11,83].

5.2. Concept and Visual Explanations

In the realm of explainable AI (XAI), recent work has focused on enhancing the depth and quality of solutions through both local and global explanations. For instance, Achtibat et al. [5] propose Concept Relevance Propagation (CRP), a novel method that bridges the gap between local and global XAI approaches. By addressing both the "where" and "what" aspects of model predictions, CRP provides human-understandable explanations via techniques like concept atlases and subspace analysis. Similarly, Dreyer et al. [45] extend CRP to segmentation and object detection tasks with L-CRP, helping locate and visualize relevant learned concepts and uncovering biases in models like DeepLabV3+ [33] and YOLOv6 [75]. Saliency maps [119] and deconvnets [141] were popular technique for visualizing convolutional networks. Amorim et al. [8] developed a framework that assesses the faithfulness of saliency maps, focusing on tumor localization in medical images, while Chattopadhay et al. [29] introduced Grad-CAM++, an improved saliency map method for better localization of multiple object occurrences in images. These approaches underscore the importance of improving visual explanations to provide more transparent and explainable models.

5.3. Regularization and Novel Frameworks for Model Robustness

Regularization techniques are pivotal in enhancing model robustness and explainability. Choi et al. [36] explored the impact of orthogonality constraints on feature representations in deep learning models. Their work introduced an Orthogonal Sphere regularizer, which improves feature diversity, enhances semantic localization, and reduces calibration errors across various datasets. Concurrently, Du et al. [46] tackled the issue of hallucinations in large language models (LLMs) with their HaloScope framework. This method uses unlabeled LLM generations to train a classifier to detect hallucinations, achieving superior performance without the need for additional data collection. Both contributions highlight the critical role of structural constraints and novel frameworks in enhancing the explainablity and reliability of AI systems.

5.4. Counterfactual Explanations in XAI

Counterfactual explanations have become a focal point of research in XAI, as they help clarify what changes to the input data could lead to different model predictions. Aryal [13] introduces the concept of semi-factual explanations, which highlight input changes that do not alter the model's output. These explanations offer insights into the stability of model predictions, thus enhancing trust and interpretability. In parallel, Bajaj et al. [16] have worked on generating robust counterfactual explanations for Graph Neural Networks (GNNs), ensuring that the explanations remain stable even in noisy environments. Their method identifies subgraphs that strongly influence predictions while maintaining human interpretability. Brughmans et al. [26] have developed NICE, a tool designed to produce counterfactuals for tabular data, optimizing for sparsity, proximity, and plausibility, thus ensuring that the explanations are both efficient and applicable to a wide range of models.

5.4.1. Benchmarking and Frameworks for Counterfactuals

Several studies have also focused on benchmarking and unifying counterfactual explanation methods. Guidotti [53] provides a comprehensive review and benchmarking of counterfactual explanation techniques, categorizing them based on properties such as stability, diversity, and actionability. This analysis highlights trade-offs in counterfactual generation and calls for methods that can balance multiple desirable properties. These surveys and frameworks are essential for refining counterfactual techniques and ensuring they are actionable, efficient, and interpretable for both technical and non-technical users.

5.4.2. Applications of Counterfactual Explanations in Healthcare

Counterfactual explanations play a vital role in healthcare applications, particularly in systems like AIMEN [85] and GlucoLens [83]. These systems use counterfactual reasoning to provide actionable insights for clinicians and patients, such as demonstrating the effects of alternative interventions or behavioral changes. By showing how small changes in input features could lead to different outcomes, counterfactual explanations not only improve the interpretability of AI-driven decisions but also enhance user trust in these systems. The use of counterfactuals in healthcare systems aligns with the ongoing need to make AI both more transparent and user-friendly, particularly in high-stakes domains such as healthcare and clinical decision-making.

6. Explainable AI Techniques

6.1. LIME (Local Interpretable Model-Agnostic Explanations)

LIME [110] provides explanations for predictions by fitting interpretable models (e.g., linear regression) locally around the data point of interest. It perturbs the input data to create a neighborhood and observes the behavior of the black-box model. The resulting simplified model highlights the contribution of individual features, offering an easy-to-understand explanation of the prediction. LIME's flexibility allows it to work with any machine learning model, making it widely applicable. An example of a LIME plot is presented in Figure 3.

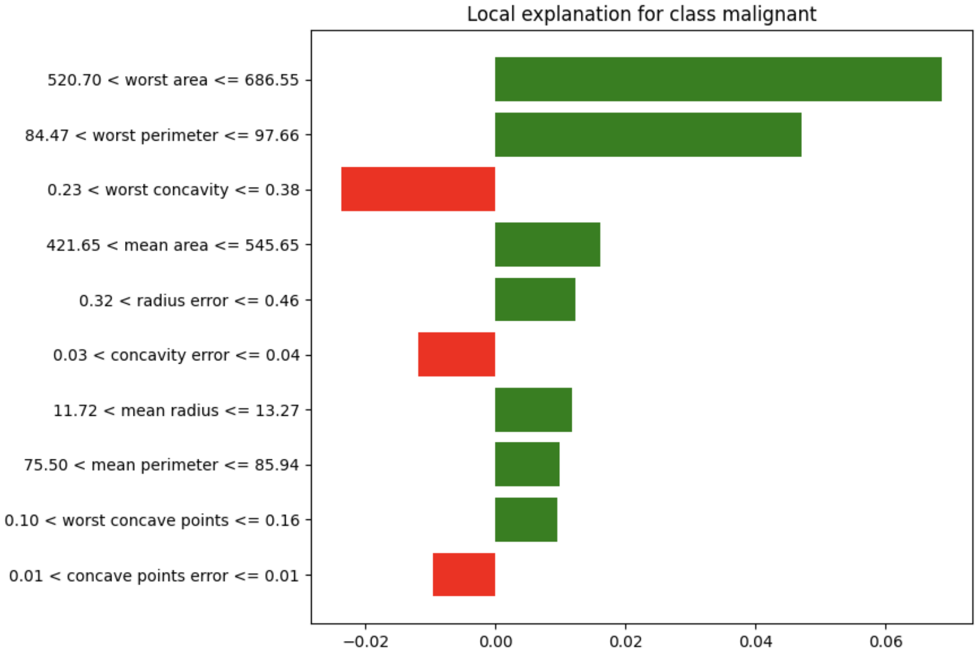


Figure 3. Example of a LIME plot for breast cancer detection. In this example, we applied LIME on the publicly available breast cancer dataset [3,134].

6.2. Shapley Values

Shapley values [117] are derived from cooperative game theory and represent a method to fairly distribute a total value among contributors (features in machine learning). They calculate the marginal contribution of a feature by averaging over all permutations of feature inclusion. While computationally intensive, Shapley values form the theoretical basis for tools like SHAP, ensuring fairness and consistency in feature attribution.

6.3. SHAP (SHapley Additive exPlanations)

SHAP [80] employs concepts from cooperative game theory, specifically Shapley values, to quantify the contribution of each feature to a model's prediction. By averaging over all possible feature subsets, SHAP guarantees a fair distribution of the prediction value among the features. Its additive nature makes it compatible with both linear and complex models, providing consistent and theoretically sound explanations. An example of a SHAP plot is presented in Figure 4.

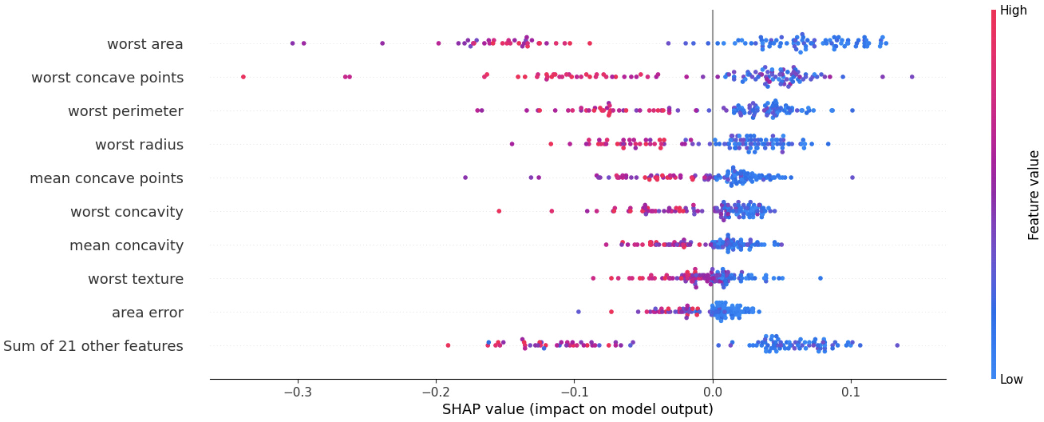


Figure 4. Example of a SHAP BeeSwarm plot for breast cancer detection. We applied SHAP on the breast cancer dataset [3,134].

6.4. LRP (Layer-Wise Relevance Propagation)

LRP [24] is a technique for explaining neural networks by redistributing the model's output backward through the layers to the input. Using relevance propagation rules, it assigns relevance scores to each neuron and feature, indicating their contribution to the prediction. LRP is particularly useful for understanding the decision-making process in deep networks like CNNs and RNNs, especially in image and text tasks.

6.5. GradCAM (Gradient-Weighted Class Activation Mapping)

GradCAM [115] generates visual explanations for CNN-based models by highlighting important regions in input images. It computes gradients of the model's output with respect to the feature maps of the last convolutional layer, using these gradients to create heatmaps. The resulting heatmaps overlay the input image to show areas that strongly influence the model's decision, making it highly intuitive for interpreting visual data. An example of the GRADCAM plot is presented in Figure 5.

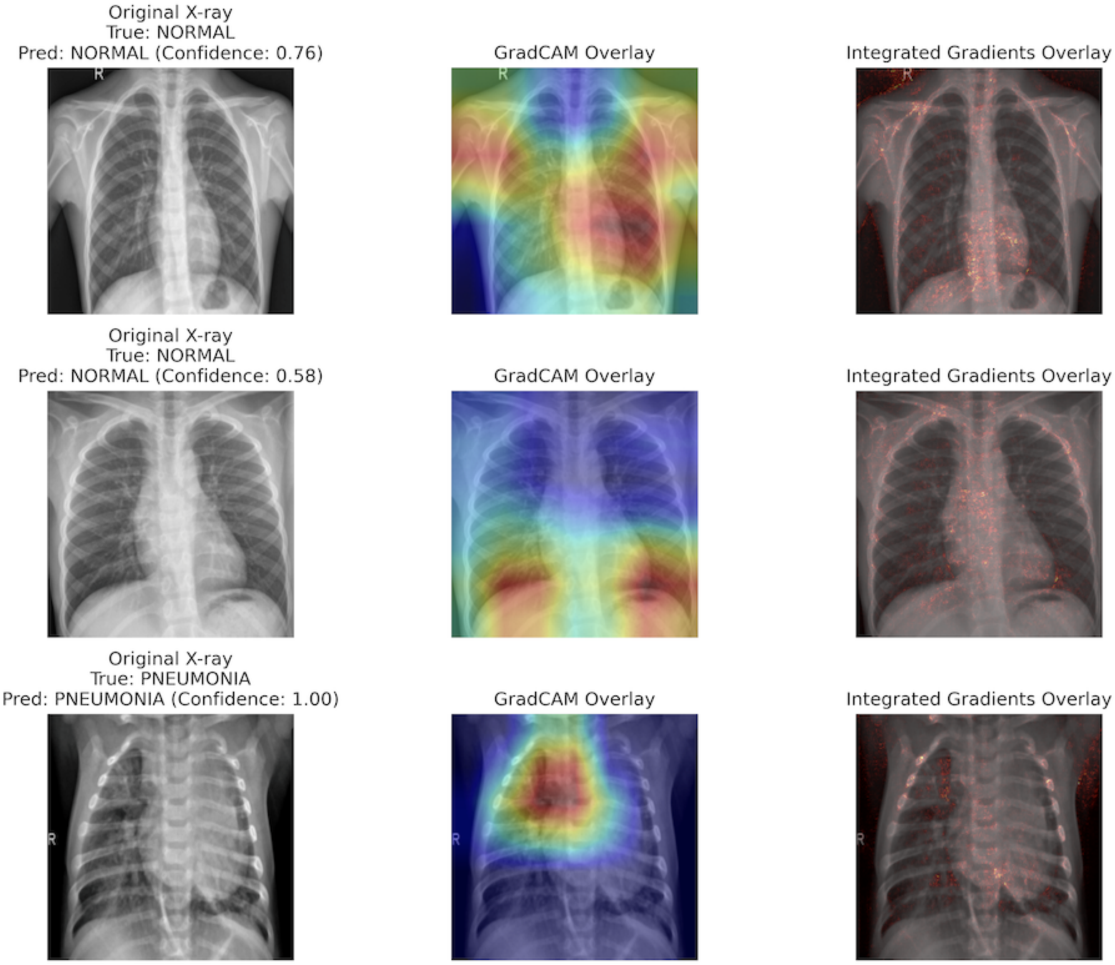

Original X-ray
True: NORMAL
Pred: NORMAL (Confidence: 0.76)
GradCAM Overlay
Integrated Gradients Overlay
Original X-ray
True: NORMAL
Pred: NORMAL (Confidence: 0.58)
GradCAM Overlay
Integrated Gradients Overlay
Original X-ray
True: PNEUMONIA
Pred: PNEUMONIA (Confidence: 1.00)
GradCAM Overlay
Integrated Gradients Overlay
R

Figure 5. Example of GradCAM and Integrated Gradients plots for detecting pneumonia from chest X-rays from a publicly available dataset [66]. In this example, we trained a ResNet18 model that achieves a test accuracy of 82% and a macro average test F1 score of 0.77. Then, GradCAM and Integrated Gradients were applied for the explanation of the model's predictions. In the first normal case, the explanation methods highlight the overall X-ray as most of the X-ray suggests a normal finding. In the second example, the model predicts a normal finding correctly, but with lower confidence, and both explanation methods highlight the areas that led the model to predict normal. In the third example, the model correctly predicts pneumonia with high confidence, and GradCAM and Integrated Gradients refer to the areas responsible for this diagnosis, because of the possibility of reticular opacities, a common indicator of pneumonia.

6.6. Integrated Gradients

Integrated Gradients [126] is a widely used attribution technique designed to explain predictions made by deep neural networks in a theoretically grounded manner. It satisfies key axioms such as sensitivity and implementation invariance, which ensure that the attributions are consistent and faithful to the model's behavior. The method estimates feature importance by tracing a straight-line path from a baseline input (often a zero or black image) to the actual input, and accumulating gradients along this path. This process results in attribution scores for each feature, capturing how changes in the input influence the model's output. The resulting attribution map highlights the most influential input features, offering a more stable and interpretable explanation compared to basic gradient-based methods. This makes Integrated Gradients especially valuable in sensitive domains like medical imaging and other applications requiring transparent decision-making. An example visualization of this method is shown in Figure 5.

6.7. NICE (Nearest Instance Counterfactual Explanations)

NICE [26] focuses on generating counterfactual explanations by identifying the nearest instance in the feature space that results in a different prediction. It minimizes the distance between the original and counterfactual inputs while ensuring that the counterfactual belongs to a different class. This approach ensures that the generated explanations are not only interpretable but also closely aligned with the data distribution, making them realistic and actionable.

6.8. DiCE (Diverse Counterfactual Explanations)

DiCE [98] generates counterfactual explanations to highlight how minimal changes to input features could alter a model's prediction. It focuses on creating diverse explanations, providing multiple plausible alternatives for user exploration. By balancing proximity, diversity, and feasibility, DiCE is well-suited for use in sensitive domains like finance and healthcare, where interpretability is crucial.

6.9. CFNOW

CFNOW [43] adopts a two-step optimization strategy for generating counterfactual explanations. The first step, counterfactual (CF) search, identifies an initial solution that alters the original classification, focusing on finding a feasible counterfactual without ensuring it is optimal. The second step, CF improvement, refines this initial solution using an objective function, such as minimizing the distance between the original and counterfactual instances, to enhance its quality. CFNOW supports four approaches: greedy simple (CGS), greedy optimized (CGO), random simple (CRS), and random optimized (CRO). While simple methods prioritize speed and computational efficiency, optimized methods refine the counterfactual for improved plausibility and adherence to specific objectives. CFNOW processes diverse data types, including tabular, image, and textual inputs, by converting them into tabular-like features. It supports binary and multiclass classification tasks with options to compare factual and counterfactual classes. This flexibility makes CFNOW a powerful tool for generating realistic and adaptable counterfactual explanations across a variety of domains.

7. Trustworthy AI in the Era of LLMs

Emerging technologies leveraging advancements in large language models (LLMs) are rapidly reshaping numerous domains. The advent of systems like ChatGPT has showcased the potential of LLMs to enhance research, education, and medical practices by democratizing access to information and improving decision-making capabilities [38]. However, these advancements come with challenges, such as the risks of misinformation and ethical concerns related to accountability. Beyond healthcare, the development of Llama 3 models demonstrates strides in multilingual support, coding, and multimodal functionalities, enabling applications in diverse fields from content generation to knowledge synthesis [47]. By integrating image, video, and speech processing, these systems push the boundaries of AI capabilities, further amplified by novel approaches like retrieval-augmented generation (RAG), which combines parametric and non-parametric memories to enhance specificity and factual accuracy in language generation [73].

The focus on reasoning capabilities has also gained momentum with models like DeepSeek-R1, which employs reinforcement learning to refine logical inference and problem-solving skills [54]. This approach complements the zero-shot reasoning capabilities explored in foundational works on chain-of-thought prompting, which reveal the untapped cognitive potential of LLMs [68]. Furthermore, specialized methods like RNAS-CL [99] improve the robustness of neural architectures through cross-layer knowledge distillation, addressing challenges of adversarial vulnerability in AI systems [99]. Together, these innovations underscore a transformative era where AI systems evolve to be not only versatile but also reliable and ethical, fostering advancements across healthcare, education, mental health, and beyond.

In our opinion, the explanations for any prediction or output by language models can be divided into three categories, i) the explanations are generated by the same language model [34], ii) the explanations are generated by a non-language model, iii) the explanations are generated by another language model [120]. The previous publications usually divide the first category again into two subcategories [34]: a) chain of thought (CoT) [68,101,133], or b) post-hoc [28,105] .

As these technologies become integral to critical decision-making processes, ensuring their trustworthiness is paramount [17,79]. In the context of LLMs, transparency refers to the ability to understand how these models arrive at their conclusions, a challenge given their complex, opaque nature. The recent reasoning models are being created in a way so that a detailed explanation of the steps are provided with the LLM response [20,54,63,108]. It was found that providing reasonings behind the decision often improves the accuracy of the prediction of large language models [68]. However, the more important aspect of the reasoning is building trust of humans by providing additional information (the reasoning itself) that can be cross-checked and verified [77,78]. The development of interpretable models and methods for explaining their predictions is vital to building user trust and ensuring that these systems are used responsibly. Alongside explainability, fairness is a crucial component of trustworthy AI, requiring that these systems be free from biases that could lead to discriminatory outcomes, particularly in sensitive domains like healthcare [59]. Researchers are focusing on methods for detecting and mitigating biases, ensuring that LLMs treat all users equitably [30,40,140].

Moreover, accountability plays a central role in the trustworthiness of LLMs, particularly in high-stakes healthcare contexts where decisions can directly impact patient outcomes. LLMs are now being used more and more for different purposes, including therapy [23,109], diagnosis [72,139], or assisting doctors [27,135]. This requires clear guidelines for the responsible use of AI in medicine, including strong human oversight, clinical validation, and safeguards against misinformation or unsafe outputs. As LLMs become more integrated into diagnostic support, patient communication, and documentation, embedding ethical frameworks into their development and deployment will be essential. Such frameworks must align with core healthcare principles, such as beneficence, non-maleficence, autonomy, and justice, while addressing pressing concerns around data privacy, security, bias, and explainability.

Multimodal foundation models (MMFMs) are vital in areas like autonomous driving and healthcare but exhibit vulnerabilities such as biases and unsafe content [137]. The MMDT (Multimodal DecodingTrust) platform addresses these issues by comprehensively evaluating MMFMs' safety, fairness, privacy, and robustness, identifying areas for improvement [137].

8. Evaluation Methods and Metrics of Trustworthy AI

8.1. Evaluation Methods

Evaluating explainable AI methods is challenging due to the absence of ground truth in real-world data, unlike supervised learning tasks. Doshi-Velez and Kim (2017) [44] propose three levels of evaluation: (1) Application level, where explanations are tested by end-users (e.g., radiologists using fracture detection software), (2) Human level, using simplified tasks with laypersons, and (3) Function level, relying on proxies such as model characteristics (e.g., tree depth). Evaluation further involves examining properties of explanation methods (e.g., expressiveness, translucency, portability, complexity) and individual explanations (e.g., accuracy, fidelity, stability, comprehensibility [94]. These properties provide a framework for assessing how well explanations align with model predictions, user understanding, and task-specific requirements, emphasizing the importance of comprehensibility, certainty, and novelty in fostering trustworthy AI.

Z-Inspection is a method introduced by Zicari et al. (2021) [145] that integrates holistic and analytic approaches to evaluate the ethical impact of AI technologies on users, society, and the environment. Its process involves three phases: Set Up, which establishes the assessment framework; Assess, which identifies and maps ethical, technical, and legal issues; and Resolve, which addresses these issues and provides recommendations for ethical AI maintenance. Kowald et al. (2024) [69] provides a comprehensive perspective on trustworthy AI by focusing on evaluation aspects across the entire AI lifecycle. It offers a unified approach that addresses technical, human-centric, and legal requirements, including transparency, fairness, accountability, and human agency, while outlining open challenges and opportunities for further research.

8.2. Metrics of Evaluation

Evaluating trustworthy AI methods is essential to ensure their utility, reliability, and alignment with user needs. Effective metrics assess various aspects of explanations, such as their accuracy, simplicity, and robustness, providing a comprehensive understanding of the strengths and limitations of the methods. This subsection outlines key metrics, including validity, fidelity, consistency, sparsity, and robustness, which collectively capture the quality of explanations from different perspectives.

8.2.1. Trust

Schmidt and Biessman [113] introduce a trust coefficient to quantify the extent to which human decisions are influenced by machine learning (ML) model predictions relative to ground truth labels:

$$\bar{T} = \frac{\mathrm{ITR}_{\hat{Y}_{\mathrm{ML}}}}{\mathrm{ITR}_{Y}}$$

Here, $\mathrm{ITR}_{\hat{Y}_{\mathrm{ML}}}$ measures the mutual information between human decisions and model predictions, while ITR_{Y} measures it with respect to the true labels. A coefficient greater than one suggests over-reliance on the model, whereas a value below one indicates skepticism or reduced trust. We refer the readers to the paper by Schmidt and Biessman [113] for more details about the calculations of $\mathrm{ITR}_{\hat{Y}_{\mathrm{ML}}}$ and ITR_{Y}.

Survey-based assessments [121] provide a validated framework for measuring public trust and openness toward AI in healthcare. They can evaluate attitudes across multiple dimensions, including diagnosis, treatment, and decision-making support, while also capturing concerns related to privacy, equity, care quality, and the human element of medicine. Such multi-dimensional measures offer critical insights into the factors that influence patients' acceptance of AI-enabled technologies in learning health systems.

8.2.2. Validity

Validity refers to whether a generated counterfactual truly changes the model's prediction compared to the original input. Formally, a counterfactual example $\mathbf{x}'$ is considered valid if it results in a different predicted class than the original instance $\mathbf{x}$ [53], i.e.,

$$b(\mathbf{x}') \neq b(\mathbf{x}),$$

where $b(\cdot)$ denotes the prediction function.

According to Verma et al. [129], a valid counterfactual must be classified in the desired target class. To evaluate this quantitatively, Mothilal et al. [98] define validity as the fraction of unique counterfactuals that produce a different outcome than the original input:

$$\text{Validity} = \frac{|\{\mathbf{c} \in C \mid f(\mathbf{c}) \neq f(\mathbf{x})\}|}{k},$$

where C is the set of generated counterfactuals, k is the total number generated, and $f(\cdot)$ is the model's prediction function. Only unique counterfactuals need to be considered to avoid overcounting identical examples returned by the method.

8.2.3. Fidelity

Fidelity measures how well an explanation approximates the black-box model's predictions, and is crucial for trust, as low-fidelity explanations fail to reflect the model's true behavior and thus lack utility [94]. There are multiple computational measures for fidelity, such as simulated experiments, sanity checks, and comparative evaluation [93]. Miró-Nicolau et al. [91] discussed four metrics of fidelity in their recent review paper: region perturbation, faithfulness correlation, faithfulness estimation, and infidelity. The LIME paper by Riberio et al. [110] provides a loss function for unfaithfulness of an explanation, which is the opposite of fidelity. An explanation is a simplified model $g \in \mathcal{G}$ (e.g., a linear model or decision tree) that approximates the predictions of a complex model f in the vicinity of an instance x, using interpretable components. To balance local fidelity and human interpretability, LIME minimizes the unfaithfulness of g to f, while keeping the complexity $\Omega(g)$ low:

$$\xi(x) = \arg\min_{g \in \mathcal{G}} L(f, g, \pi_x) + \Omega(g)$$

where $L(f, g, \pi_x)$ is a loss function that measures the lack of fidelity between f and g in the locality defined by π_x, and $\Omega(g)$ is a measure of the complexity of the explanation model [110].

8.2.4. Proximity

Mothilal et al. [98] define proximity between a counterfactual example c_i and the original input x as the average feature-wise distance, computed separately for continuous and categorical features. For evaluation, they use the original feature scale (rather than the normalized scale used during CF generation) to enhance interpretability.

$$\text{Proximity}_{\text{cont}} = -\frac{1}{k} \sum_{i=1}^{k} \text{dist}_{\text{cont}}(c_i, x)$$

$$\text{Proximity}_{\text{cat}} = 1 - \frac{1}{k} \sum_{i=1}^{k} \text{dist}_{\text{cat}}(c_i, x)$$

In general, a higher proximity or a lower distance between the original example and its corresponding counterfactual example is preferred, so that the intervention to avoid an unwanted outcome (e.g., hyperglycemia) does not require drastic changes to the input features.

8.2.5. Sparsity

Sparsity measures how many features are changed in the counterfactual (CF) examples compared to the original input, providing a complementary perspective to proximity. Specifically, it quantifies the average fraction of unchanged features across all CF examples [98].

$$\text{Sparsity} = 1 - \frac{1}{k \cdot d} \sum_{i=1}^{k} \sum_{l=1}^{d} 1_{[c_i^l \neq x^l]}$$

In this equation, if the value of sparsity is high, the counterfactual will be more sparse, or fewer features will need to be changed. Because of that, the average number of mismatched features is subtracted from 1. The higher value of the sparsity metric according to this equation is better, because that would mean fewer feature changes and less burden on the user to adhere to the intervention.

While calculating the metric according to this equation help optimize counterfactuals, a more human-friendly metric often reported is how many features, on average, are changed [85]. It can be calculated by the following equation:

$$\text{Features changed} = \frac{1}{k} \sum_{i=1}^{k} \sum_{l=1}^{d} 1_{[c_i^l \neq x^l]}$$

Note that the division by the number of features is absent here, so that regardless of how many input features were used, the user can understand how many parameters would need to be changed.

8.2.6. Diversity

Different users of a system can have different preferences and abilities to apply interventions to overcome a health challenge. While counterfactuals need to be in close proximity to the original example, it is important for the counterfactuals of the same example to be diverse enough, when possible, so that the user can have different options.

Diversity of counterfactual (CF) examples is computed by measuring the average pairwise distance between CF examples, using either continuous or categorical feature-wise distance. A count-based diversity metric is also used that captures the average fraction of differing features between each pair of CFs [98].

$$\text{Diversity} = \Delta = \frac{1}{\binom{k}{2}} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \text{dist}(c_i, c_j)$$

$$\text{Count-Diversity} = \frac{1}{\binom{k}{2} \cdot d} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \sum_{l=1}^{d} 1_{[c_i^l \neq c_j^l]}$$

These equations take pairwise distances among the counterfactual examples and take the average by dividing the summation of the distances by the number of pairs, which is the number of choices for pairs from k counterfactuals: calculated by $\binom{k}{2}$. For categorical values, the distance is calculated by adding the number of mismatched feature values and dividing by the number of features.

9. Conclusion

Building trustworthy AI systems for digital health requires addressing both robustness and explainability as foundational properties. This review synthesized recent developments across algorithmic innovations, evaluation frameworks, and practical deployment challenges to highlight the current landscape and emerging opportunities. Robustness in clinical settings must account for data distribution shifts, sensor noise, and adversarial scenarios, while explainability must go beyond model transparency to support clinician understanding, justification, and decision-making.

We emphasize that trustworthiness is not a monolithic concept, but a composite of verifiable properties that must be operationalized and measured systematically. As AI becomes increasingly integrated into healthcare workflows, aligning technical goals with human-centered requirements will be essential. Ongoing research in causal reasoning, uncertainty quantification, and interactive explanation offers promising paths forward. Ultimately, advancing trustworthy AI in digital health is not only a technical challenge but a societal imperative.

References

1. CE Marking trade.gov. https://www.trade.gov/ce-marking. [Accessed July 2025].
2. Ethics guidelines for trustworthy AI — digital-strategy.ec.europa.eu. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. [Accessed July 2025].
3. load_breast_cancer — scikit-learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html. [Accessed July 2025].
4. Trustworthy and Responsible AI — nist.gov. https://www.nist.gov/trustworthy-and-responsible-ai. [Accessed July 2025].
5. Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. From attribution maps to human-understandable explanations through concept relevance propagation. *Nature Machine Intelligence*, 5(9):1006–1019, 2023.
6. Ahmed Shihab Albahri, Ali M Duhaim, Mohammed A Fadhel, Alhamzah Alnoor, Noor S Baqer, Laith Alzubaidi, Osamah Shihab Albahri, Abdullah Hussein Alamoodi, Jinshuai Bai, Asma Salhi, et al. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion*, 96:156–191, 2023.

7. Parastoo Alinia, Iman Mirzadeh, and Hassan Ghasemzadeh. Actilabel: A combinatorial transfer learning framework for activity recognition. *arXiv preprint arXiv:2003.07415*, 2020.
8. José P Amorim, Pedro H Abreu, João Santos, Marc Cortes, and Victor Vila. Evaluating the faithfulness of saliency maps in explaining deep learning models using realistic perturbations. *Information Processing & Management*, 60(2):103225, 2023.
9. Yaqoob Ansari, Omar Mourad, Khalid Qaraqe, and Erchin Serpedin. Deep learning for ecg arrhythmia detection and classification: an overview of progress for period 2017–2023. *Frontiers in Physiology*, 14:1246746, 2023.
10. Asiful Arefeen and Hassan Ghasemzadeh. Designing user-centric behavioral interventions to prevent dysglycemia with novel counterfactual explanations. *arXiv preprint arXiv:2310.01684*, 2023.
11. Asiful Arefeen, Saman Khamesian, Maria Adela Grando, Bithika Thompson, and Hassan Ghasemzadeh. Glyman: Glycemic management using patient-centric counterfactuals. In *2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–5. IEEE, 2024.
12. Asiful Arefeen, Saman Khamesian, Maria Adela Grando, Bithika Thompson, and Hassan Ghasemzadeh. Glytwin: Digital twin for glucose control in type 1 diabetes through optimal behavioral modifications using patient-centric counterfactuals. *arXiv preprint arXiv:2504.09846*, 2025.
13. Saugat Aryal. Semi-factual explanations in ai. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23379–23380, 2024.
14. Ibrahim Berkan Aydilek and Ahmet Arslan. A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. *Information Sciences*, 233:25–35, 2013.
15. Reza Rahimi Azghan, Nicholas C Glodosky, Ramesh Kumar Sah, Carrie Cuttler, Ryan McLaughlin, Michael J Cleveland, and Hassan Ghasemzadeh. Cudle: Learning under label scarcity to detect cannabis use in uncontrolled environments using wearables. *IEEE Sensors Journal*, 2025.
16. Mohit Bajaj, Lingyang Chu, Zi Yu Xue, Jian Pei, Lanjun Wang, Peter Cho-Ho Lam, and Yong Zhang. Robust counterfactual explanations on graph neural networks. *Advances in Neural Information Processing Systems*, 34:5644–5655, 2021.
17. Stephanie Baker and Wei Xiang. Explainable ai is responsible ai: How explainability creates trustworthy and socially responsible artificial intelligence. *arXiv preprint arXiv:2312.01555*, 2023.
18. Shahab S Band, Atefeh Yarahmadi, Chung-Chian Hsu, Meghdad Biyari, Mehdi Sookhak, Rasoul Ameri, Iman Dehzangi, Anthony Theodore Chronopoulos, and Huey-Wen Liang. Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. *Informatics in Medicine Unlocked*, 40:101286, 2023.
19. Oresti Banos, Rafael Garcia, and Alejandro Saez. Mhealth dataset. *UCI machine learning repository*, 2014.
20. Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, et al. Llama-nemotron: Efficient reasoning models. *arXiv preprint arXiv:2505.00949*, 2025.
21. Visar Berisha and Julie M Liss. Responsible development of clinical speech ai: Bridging the gap between clinical research and technology. *NPJ Digital Medicine*, 7(1):208, 2024.
22. Jorge Bernal, Nima Tajkbaksh, Francisco Javier Sanchez, Bogdan J Matuszewski, Hao Chen, Lequan Yu, Quentin Angermann, Olivier Romain, Bjørn Rustad, Ilangko Balasingham, et al. Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge. *IEEE transactions on medical imaging*, 36(6):1231–1249, 2017.
23. Desirée Bill and Theodor Eriksson. Fine-tuning a llm using reinforcement learning from human feedback for a therapy chatbot application, 2023.
24. Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. In *Artificial Neural Networks and Machine Learning–ICANN 2016: 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6-9, 2016, Proceedings, Part II 25*, pages 63–71. Springer, 2016.
25. Ann Borda, Andreea Molnar, Cristina Neesham, and Patty Kostkova. Ethical issues in ai-enabled disease surveillance: perspectives from global health. *Applied Sciences*, 12(8):3890, 2022.
26. Dieter Brughmans, Pieter Leyman, and David Martens. Nice: an algorithm for nearest instance counterfactual explanations. *Data mining and knowledge discovery*, 38(5):2665–2703, 2024.

27. Muhammed Kayra Bulut and Banu Diri. Artificial intelligence revolution in turkish health consultancy: Development of llm-based virtual doctor assistants. In *2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP)*, pages 1–6. IEEE, 2024.
28. Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31, 2018.
29. Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.
30. Shreyas Chaudhari, Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, Ameet Deshpande, and Bruno Castro da Silva. Rlhf deciphered: A critical analysis of reinforcement learning from human feedback for llms. *ACM Computing Surveys*, 2024.
31. Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.
32. Narmatha Chellamani, Saleh Ali Albelwi, Manimurugan Shanmuganathan, Palanisamy Amirthalingam, and Anand Paul. Diabetes: Non-invasive blood glucose monitoring using federated learning with biosensor signals. *Biosensors*, 15(4):255, 2025.
33. Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.
34. Yanda Chen, Ruiqi Zhong, Narutatsu Ri, Chen Zhao, He He, Jacob Steinhardt, Zhou Yu, and Kathleen McKeown. Do models explain themselves? counterfactual simulatability of natural language explanations. In *Proceedings of the 41st International Conference on Machine Learning*, pages 7880–7904, 2024.
35. Bijoy Chhetri, Lalit Mohan Goyal, and Mamta Mittal. How machine learning is used to study addiction in digital healthcare: A systematic review. *International Journal of Information Management Data Insights*, 3(2):100175, 2023.
36. Hongjun Choi, Anirudh Som, and Pavan Turaga. Role of orthogonality constraints in improving properties of deep networks for image classification. *arXiv preprint arXiv:2009.10762*, 2020.
37. Minjae Chung, Jong Bum Won, Ganghyun Kim, Yujin Kim, and Utku Ozbulak. Evaluating visual explanations of attention maps for transformer-based medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 110–120. Springer, 2024.
38. Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, et al. The future landscape of large language models in medicine. *Communications medicine*, 3(1):141, 2023.
39. Estee Y Cramer, Evan L Ray, Velma K Lopez, Johannes Bracher, Andrea Brennen, Alvaro J Castro Rivadeneira, Aaron Gerding, Tilmann Gneiting, Katie H House, Yuxin Huang, et al. Evaluation of individual and ensemble probabilistic forecasts of covid-19 mortality in the united states. *Proceedings of the National Academy of Sciences*, 119(15):e2113561119, 2022.
40. James L Cross, Michael A Choma, and John A Onofrey. Bias in medical ai: Implications for clinical decision-making. *PLOS Digital Health*, 3(11):e0000651, 2024.
41. Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. Multi-objective counterfactual explanations. In *International conference on parallel problem solving from nature*, pages 448–469. Springer, 2020.
42. John Daniels, Pau Herrero, and Pantelis Georgiou. A multitask learning approach to personalized blood glucose prediction. *IEEE Journal of Biomedical and Health Informatics*, 26(1):436–445, 2021.
43. Raphael Mazzine Barbosa de Oliveira, Kenneth Sörensen, and David Martens. A model-agnostic and data-independent tabu search algorithm to generate counterfactuals for tabular, image, and text data. *European Journal of Operational Research*, 317(2):286–302, 2024.
44. Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.
45. Maximilian Dreyer, Reduan Achtibat, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. Revealing hidden context bias in segmentation and object detection through concept-specific explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3828–3838, 2023.
46. Xuefeng Du, Chaowei Xiao, and Sharon Li. Haloscope: Harnessing unlabeled llm generations for hallucination detection. *Advances in Neural Information Processing Systems*, 37:102948–102972, 2025.

47. Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
48. Ebrahim Farahmand, Reza Rahimi Azghan, Nooshin Taheri Chatrudi, Eric Kim, Gautham Krishna Gudur, Edison Thomaz, Giulia Pedrielli, Pavan Turaga, and Hassan Ghasemzadeh. Attengluco: Multimodal transformer-based blood glucose forecasting on ai-readi dataset. *arXiv preprint arXiv:2502.09919*, 2025.
49. Jana Fehr, Brian Citro, Rohit Malpani, Christoph Lippert, and Vince I Madai. A trustworthy ai reality-check: the lack of transparency of artificial intelligence products in healthcare. *Frontiers in Digital Health*, 6:1267290, 2024.
50. Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.
51. Chengjian Guan, Angwei Gong, Yan Zhao, Chen Yin, Lu Geng, Linli Liu, Xiuchun Yang, Jingchao Lu, and Bing Xiao. Interpretable machine learning model for new-onset atrial fibrillation prediction in critically ill patients: a multi-center study. *Critical Care*, 28(1):349, 2024.
52. Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021.
53. Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 38(5):2770–2824, 2024.
54. Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
55. Navid Hasani, Michael A Morris, Arman Rhamim, Ronald M Summers, Elizabeth Jones, Eliot Siegel, and Babak Saboury. Trustworthy artificial intelligence in medical imaging. *PET clinics*, 17(1):1, 2022.
56. Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. Ieee, 2008.
57. Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.
58. Tianjia He, Lin Zhang, Fanxin Kong, and Asif Salekin. Exploring inherent sensor redundancy for automotive anomaly detection. In *2020 57th ACM/IEEE Design Automation Conference (DAC)*, pages 1–6. IEEE, 2020.
59. Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.
60. Shih-Cheng Huang, Anuj Pareek, Malte Jensen, Matthew P Lungren, Serena Yeung, and Akshay S Chaudhari. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *NPJ Digital Medicine*, 6(1):74, 2023.
61. Dina Hussein, Taha Belkhouja, Ganapati Bhat, and Janardhan Rao Doppa. Energy-efficient missing data recovery in wearable devices: A novel search-based approach. In *2023 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, pages 1–6. IEEE, 2023.
62. Dina Hussein and Ganapati Bhat. Cim: A novel clustering-based energy-efficient data imputation method for human activity recognition. *ACM Transactions on Embedded Computing Systems*, 22(5s):1–26, 2023.
63. Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.
64. Davinder Kaur, Suleyman Uslu, and Arjan Durresi. Requirements for trustworthy artificial intelligence–a review. In *Advances in Networked-Based Information Systems: The 23rd International Conference on Network-Based Information Systems (NBiS-2020) 23*, pages 105–115. Springer, 2021.
65. Davinder Kaur, Suleyman Uslu, Kaley J Rittichier, and Arjan Durresi. Trustworthy artificial intelligence: a review. *ACM computing surveys (CSUR)*, 55(2):1–38, 2022.
66. Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131, 2018.

67. Abby C King, Zakaria N Doueiri, Ankita Kaulberg, and Lisa Goldman Rosas. The promise and perils of artificial intelligence in advancing participatory science and health equity in public health. *JMIR Public Health and Surveillance*, 11(1):e65699, 2025.
68. Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.
69. Dominik Kowald, Sebastian Scher, Viktoria Pammer-Schindler, Peter Müllner, Kerstin Waxnegger, Lea Demelius, Angela Fessl, Maximilian Toller, Inti Gabriel Mendoza Estrada, Ilija Šimić, et al. Establishing and evaluating trustworthy ai: overview and research challenges. *Frontiers in Big Data*, 7:1467222, 2024.
70. Rayan Krishnan, Pranav Rajpurkar, and Eric J Topol. Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering*, 6(12):1346–1352, 2022.
71. Abhishek Kumar, Tristan Braud, Sasu Tarkoma, and Pan Hui. Trustworthy ai in the age of pervasive computing and big data. In *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 1–6. IEEE, 2020.
72. Wojciech Kusa, Edoardo Mosca, and Aldo Lipani. “dr llm, what do i have?”: The impact of user beliefs and prompt formulation on health diagnoses. In *Proceedings of the Third Workshop on NLP for Medical Conversations*, pages 13–19, 2023.
73. Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.
74. Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. Trustworthy ai: From principles to practices. *ACM Computing Surveys*, 55(9):1–46, 2023.
75. Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, et al. Yolov6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*, 2022.
76. Qian Li, Xi Yang, Jie Xu, Yi Guo, Xing He, Hui Hu, Tianchen Lyu, David Marra, Amber Miller, Glenn Smith, et al. Early prediction of alzheimer’s disease and related dementias using real-world electronic health records. *Alzheimer's & Dementia*, 19(8):3506–3518, 2023.
77. Zhiming Li, Yushi Cao, Xiufeng Xu, Junzhe Jiang, Xu Liu, Yon Shin Teo, Shang-Wei Lin, and Yang Liu. Llms for relational reasoning: How far are we? In *Proceedings of the 1st International Workshop on Large Language Models for Code*, pages 119–126, 2024.
78. Zhenwen Liang, Ye Liu, Tong Niu, Xiangliang Zhang, Yingbo Zhou, and Semih Yavuz. Improving llm reasoning through scaling inference computation with collaborative verification. *arXiv preprint arXiv:2410.05318*, 2024.
79. Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy llms: a survey and guideline for evaluating large language models’ alignment. *arXiv preprint arXiv:2308.05374*, 2023.
80. Scott Lundberg and Su-in Lee. Shap: A unified approach to interpreting model predictions. *Advances in neural information processing systems*, pages 1–10, 2017.
81. Mohammad Malekzadeh, Richard G. Clegg, Andrea Cavallaro, and Hamed Haddadi. Protecting sensory data against sensitive inferences. In *Proceedings of the 1st Workshop on Privacy by Design in Distributed Systems*, W-P2DS’18, pages 2:1–2:6, New York, NY, USA, 2018. ACM.
82. Mohammad Malekzadeh, Richard G. Clegg, Andrea Cavallaro, and Hamed Haddadi. Mobile sensor data anonymization. In *Proceedings of the International Conference on Internet of Things Design and Implementation*, IoTDI ’19, pages 49–58, New York, NY, USA, 2019. ACM.
83. Abdullah Mamun, Asiful Arefeen, Susan B. Racette, Dorothy D. Sears, Corrie M. Whisner, Matthew P. Buman, and Hassan Ghasemzadeh. Llm-powered prediction of hyperglycemia and discovery of behavioral treatment pathways from wearables and diet, 2025.
84. Abdullah Mamun, Diane J Cook, and Hassan Ghasemzadeh. Aimi: Leveraging future knowledge and personalization in sparse event forecasting for treatment adherence. *arXiv preprint arXiv:2503.16091*, 2025.
85. Abdullah Mamun, Lawrence D Devoe, Mark I Evans, David W Britt, Judith Klein-Seetharaman, and Hassan Ghasemzadeh. Use of what-if scenarios to help explain artificial intelligence models for neonatal health. *arXiv preprint arXiv:2410.09635*, 2024.
86. Abdullah Mamun, Chia-Cheng Kuo, David W Britt, Lawrence D Devoe, Mark I Evans, Hassan Ghasemzadeh, and Judith Klein-Seetharaman. Neonatal risk modeling and prediction. In *2023 IEEE 19th International Conference on Body Sensor Networks (BSN)*, pages 1–4. IEEE, 2023.

87. Abdullah Mamun, Krista S Leonard, Matthew P Buman, and Hassan Ghasemzadeh. Multimodal time-series activity forecasting for adaptive lifestyle intervention design. In *2022 IEEE-EMBS International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pages 1–4. IEEE, 2022.
88. Abdullah Mamun, Krista S Leonard, Megan E Petrov, Matthew P Buman, and Hassan Ghasemzadeh. Multimodal physical activity forecasting in free-living clinical settings: Hunting opportunities for just-in-time interventions. *arXiv preprint arXiv:2410.09643*, 2024.
89. Abdullah Mamun, Seyed Iman Mirzadeh, and Hassan Ghasemzadeh. Designing deep neural networks robust to sensor failure in mobile health environments. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2442–2446. IEEE, 2022.
90. Elliot Mbunge, Benhildah Muchemwa, Sipho'esihle Jiyane, and John Batani. Sensors and healthcare 5.0: transformative shift in virtual care through emerging digital health technologies. *global health journal*, 5(4):169–177, 2021.
91. Miquel Miró-Nicolau, Antoni Jaume-i Capó, and Gabriel Moyà-Alcover. A comprehensive study on fidelity metrics for xai. *Information Processing & Management*, 62(1):103900, 2025.
92. Seyed Iman Mirzadeh, Jessica Ardo, Ramin Fallahzadeh, Bryan Minor, Lorraine Evangelista, Diane Cook, and Hassan Ghasemzadeh. Labelmerger: Learning activities in uncontrolled environments. In *2019 First International Conference on Transdisciplinary AI (TransAI)*, pages 64–67. IEEE, 2019.
93. Sina Mohseni, Niloofar Zarei, and Eric D Ragan. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(3-4):1–45, 2021.
94. Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.
95. Michael Moor, Nicolas Bennett, Drago Plečko, Max Horn, Bastian Rieck, Nicolai Meinshausen, Peter Bühlmann, and Karsten Borgwardt. Predicting sepsis using deep learning across international sites: a retrospective development and validation study. *EClinicalMedicine*, 62, 2023.
96. George Moschonis, George Siopis, Jenny Jung, Evette Eweka, Ruben Willems, Dominika Kwasnicka, Bernard Yeboah-Asiamah Asare, Vimarsha Kodithuwakku, Nick Verhaeghe, Rajesh Vedanthan, et al. Effectiveness, reach, uptake, and feasibility of digital health interventions for adults with type 2 diabetes: a systematic review and meta-analysis of randomised controlled trials. *The Lancet Digital Health*, 5(3):e125–e143, 2023.
97. Sayyed Mostafa Mostafavi, Shovito Barua Soumma, Daniel Peterson, Shyamal H Mehta, and Hassan Ghasemzadeh. Detection and severity assessment of parkinson's disease through analyzing wearable sensor data using gramian angular fields and deep convolutional neural networks. *Sensors*, 25(11):3421, 2025.
98. Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 607–617, 2020.
99. Utkarsh Nath, Yancheng Wang, Pavan Turaga, and Yingzhen Yang. Rnas-cl: Robust neural architecture search by cross-layer knowledge distillation. *International Journal of Computer Vision*, 132(12):5698–5717, 2024.
100. Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, Andrew Y Ng, et al. Multimodal deep learning. In *ICML*, volume 11, pages 689–696, 2011.
101. Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. Show your work: Scratchpads for intermediate computation with language models. 2021.
102. Jaya Ojha, Oriana Presacan, Pedro G. Lind, Eric Monteiro, and Anis Yazidi. Navigating uncertainty: A user-perspective survey of trustworthiness of ai in healthcare. *ACM Transactions on Computing for Healthcare*, 6(3):1–32, 2025.
103. David B Olawade, Ojima J Wada, Aanuoluwapo Clement David-Olawade, Edward Kunonga, Olawale Abaire, and Jonathan Ling. Using artificial intelligence to improve public health: a narrative review. *Frontiers in Public Health*, 11:1196397, 2023.
104. Cheng Ouyang, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu, and Daniel Rueckert. Self-supervised learning for few-shot medical image segmentation. *IEEE Transactions on Medical Imaging*, 41(7):1837–1848, 2022.
105. Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8779–8788, 2018.

106. Ninlapa Pruksanusak, Natthicha Chainarong, Siriwan Boripan, and Alan Geater. Comparison of the predictive ability for perinatal acidemia in neonates between the nichd 3-tier fhr system combined with clinical risk factors and the fetal reserve index. *Plos one*, 17(10):e0276451, 2022.
107. Xinyuan Qi, Kai Hou, Tong Liu, Zhongzhong Yu, Sihao Hu, and Wenwu Ou. From known to unknown: Knowledge-guided transformer for time-series sales forecasting in alibaba. *arXiv preprint arXiv:2109.08381*, 2021.
108. Abhinav Rastogi, Albert Q Jiang, Andy Lo, Gabrielle Berrada, Guillaume Lample, Jason Rute, Joep Barmentlo, Karmesh Yadav, Kartik Khandelwal, Khyathi Raghavi Chandu, et al. Magistral. *arXiv preprint arXiv:2506.10910*, 2025.
109. Xiaoyu Ren, Yuanchen Bai, Huiyu Duan, Lei Fan, Erkang Fei, Geer Wu, Pradeep Ray, Menghan Hu, Chenyuan Yan, and Guangtao Zhai. Chatasd: Llm-based ai therapist for asd. In *International Forum on Digital TV and Wireless Multimedia Communications*, pages 312–324. Springer, 2023.
110. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.
111. Ramesh Kumar Sah, Michael J Cleveland, and Hassan Ghasemzadeh. Stress monitoring in free-living environments. *IEEE Journal of Biomedical and Health Informatics*, 2023.
112. Deepti Saraswat, Pronaya Bhattacharya, Ashwin Verma, Vivek Kumar Prasad, Sudeep Tanwar, Gulshan Sharma, Pitshou N Bokoro, and Ravi Sharma. Explainable ai for healthcare 5.0: opportunities and challenges. *IEEE Access*, 10:84486–84517, 2022.
113. Philipp Schmidt and Felix Biessmann. Quantifying interpretability and trust in machine learning systems. 2019.
114. Ivan W Selesnick and C Sidney Burrus. Generalized digital butterworth filter design. *IEEE Transactions on signal processing*, 46(6):1688–1694, 2002.
115. Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.
116. Sanyam Paresh Shah, Abdullah Mamun, Shovito Barua Soumma, and Hassan Ghasemzadeh. Enhancing metabolic syndrome prediction with hybrid data balancing and counterfactuals. *arXiv preprint arXiv:2504.06987*, 2025.
117. Lloyd S Shapley et al. A value for n-person games. 1953.
118. Sherif A Shazly, Bijan J Borah, Che G Ngufor, Vanessa E Torbenson, Regan N Theiler, and Abimbola O Famuyide. Impact of labor characteristics on maternal and neonatal outcomes of labor: a machine-learning model. *Plos one*, 17(8):e0273178, 2022.
119. Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.
120. Chandan Singh, Aliyah R Hsu, Richard Antonello, Shailee Jain, Alexander G Huth, Bin Yu, and Jianfeng Gao. Explaining black box text modules in natural language with language models. *arXiv preprint arXiv:2305.09863*, 2023.
121. Bryan A Sisk, Alison L Antes, Sunny C Lin, Paige Nong, and James M DuBois. Validating a novel measure for assessing patient openness and concerns about using artificial intelligence in healthcare. *Learning Health Systems*, 9(1):e10429, 2025.
122. Shovito Barua Soumma, SM Alam, Rudmila Rahman, Umme Niraj Mahi, Abdullah Mamun, Sayyed Mostafa Mostafavi, and Hassan Ghasemzadeh. Freezing of gait detection using gramian angular fields and federated learning from wearable sensors. *arXiv preprint arXiv:2411.11764*, 2024.
123. Shovito Barua Soumma, Abdullah Mamun, and Hassan Ghasemzadeh. Domain-informed label fusion surpasses llms in free-living activity classification (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2025.
124. Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.
125. Charithea Stylianides, Andria Nicolaou, Waqar Aziz Sulaiman, Christina-Athanasia Alexandropoulou, Ilias Panagiotopoulos, Konstantina Karathanasopoulou, George Dimitrakopoulos, Styliani Kleanthous, Eleni Politi, Dimitris Ntalaperas, et al. Ai advances in icu with an emphasis on sepsis prediction: An overview. *Machine Learning and Knowledge Extraction*, 7(1):6, 2025.
126. Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

127. Jayaraman J Thiagarajan, Kowshik Thopalli, Deepta Rajan, and Pavan Turaga. Training calibration-based counterfactual explainers for deep learning models in medical image analysis. *Scientific reports*, 12(1):597, 2022.
128. George Vavoulas, Charikleia Chatzaki, Thodoris Malliotakis, Matthew Pediaditis, and Manolis Tsiknakis. The mobiact dataset: Recognition of activities of daily living using smartphones. In *International conference on information and communication technologies for ageing well and e-health*, volume 2, pages 143–151. SciTePress, 2016.
129. Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*, 2(1):1, 2020.
130. Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.
131. Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019.
132. Jessica L Watterson, Hector P Rodriguez, Stephen M Shortell, and Adrian Aguilera. Improved diabetes care management through a text-message intervention for low-income patients: mixed-methods pilot study. *JMIR diabetes*, 3(4):e8645, 2018.
133. Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
134. William Wolberg, Olvi Mangasarian, Nick Street, and W. Street. Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository, 1993. DOI: https://doi.org/10.24432/C5DW2B.
135. Chengyan Wu, Zehong Lin, Wenlong Fang, and Yuyan Huang. A medical diagnostic assistant based on llm. In *China Health Information Processing Conference*, pages 135–147. Springer, 2023.
136. Xiaolong Wu, Lin Ma, Penghu Wei, Yongzhi Shan, Piu Chan, Kailiang Wang, and Guoguang Zhao. Wearable sensor devices can automatically identify the on-off status of patients with parkinson's disease through an interpretable machine learning model. *Frontiers in Neurology*, 15:1387477, 2024.
137. Chejian Xu, Jiawei Zhang, Zhaorun Chen, Chulin Xie, Mintong Kang, Yujin Potter, Zhun Wang, Zhuowen Yuan, Alexander Xiong, Zidi Xiong, et al. Mmdt: Decoding the trustworthiness and safety of multimodal foundation models. In *The Thirteenth International Conference on Learning Representations*.
138. Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32, 2019.
139. Bufang Yang, Siyang Jiang, Lilin Xu, Kaiwei Liu, Hai Li, Guoliang Xing, Hongkai Chen, Xiaofan Jiang, and Zhenyu Yan. Drhouse: An llm-empowered diagnostic reasoning system through harnessing outcomes from sensor data and expert knowledge. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(4):1–29, 2024.
140. Yifan Yang, Xiaoyu Liu, Qiao Jin, Furong Huang, and Zhiyong Lu. Unmasking and quantifying racial bias of large language models in medical report generation. *Communications medicine*, 4(1):176, 2024.
141. Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.
142. Junhai Zhai, Jiaxing Qi, and Chu Shen. Binary imbalanced data classification based on diversity oversampling by generative models. *Information Sciences*, 585:313–343, 2022.
143. Kexin Zhang, Baoyu Jing, K Selçuk Candan, Dawei Zhou, Qingsong Wen, Han Liu, and Kaize Ding. Cross-domain conditional diffusion models for time series imputation. *arXiv preprint arXiv:2506.12412*, 2025.
144. Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging*, 39(6):1856–1867, 2019.
145. Roberto V Zicari, John Brodersen, James Brusseau, Boris Düdder, Timo Eichhorn, Todor Ivanov, Georgios Kararigas, Pedro Kringen, Melissa McCullough, Florian Möslein, et al. Z-inspection®: a process to assess trustworthy ai. *IEEE Transactions on Technology and Society*, 2(2):83–97, 2021.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.