

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

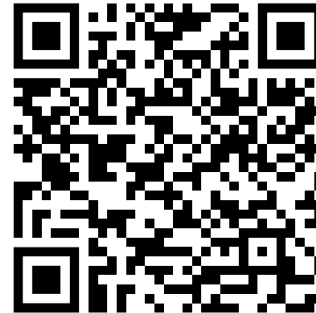
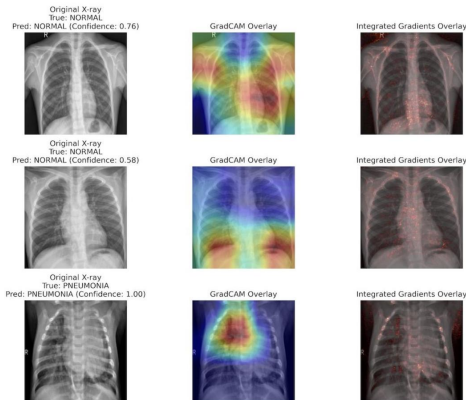
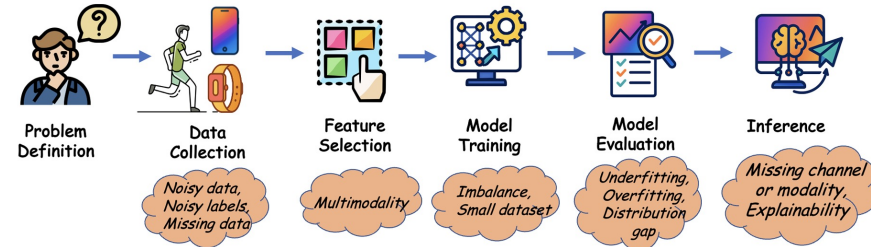


**Abdullah Mamun\***, Shovito Barua Soumma, Hassan Ghasemzadeh

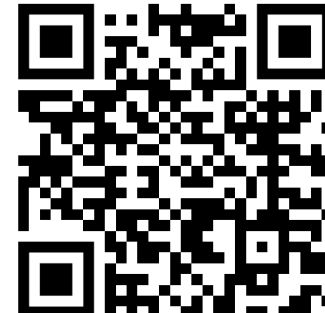
Journal: Progress in Biomedical Engineering (Impact Factor: 7.7)

Publication date: March 6, 2026

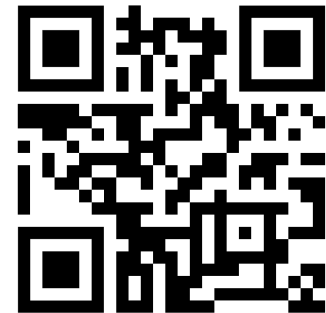
Email: [a.mamun@asu.edu](mailto:a.mamun@asu.edu)



Paper



Preprint



X: @AB9Mamun



# Trustworthy AI in Digital Health

## AI Trust in Digital Health

Ensuring trust in AI systems is essential for safe and ethical integration into high-stakes domains like digital health, addressing robustness, explainability, fairness, accountability, and privacy across the AI lifecycle.

# Introduction to Trustworthy AI

## 1 Core Principles

Trustworthy AI focuses on reliability, transparency, and accountability, emphasizing fairness, safety, and explainability to foster user confidence and ethical decision-making.

## 4 Robustness

Robustness ensures AI systems maintain reliable performance despite noisy inputs, sensor failures, or imbalanced datasets, as seen in multisensor activity recognition and disease diagnosis.

## 2 NIST Definition

NIST defines trustworthy AI through core characteristics including validity, reliability, safety, security, accountability, transparency, explainability, privacy, and fairness.

## 5 Explainability

Explainability is crucial for intelligent digital health systems to provide actionable feedback and model-based reasoning, enabling safer interventions and improved health outcomes through insights like counterfactual explanations.

## 3 EU Guidelines

The High-Level Expert Group on AI outlined 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.

# Prior Reviews on Trustworthy AI



## **Comprehensive Analysis**

Reviews cover diverse requirements like fairness, explainability, accountability, and reliability, offering insights into risk mitigation and societal acceptance.

## **Ethical Foundations**

Ethics are embedded in system design and development, focusing on practical applications like smart cities.

## **EU Principles and Trustworthy AI**

Approaches are consolidated for trustworthy AI based on EU principles, providing a structured overview for reliable systems.

## **LLM Alignment and Trust**

Reviews identify key dimensions of trustworthiness for LLMs, such as safety, fairness, and adherence to social norms, highlighting alignment challenges.

## **Transparency Gaps in Medical AI**

Reviews reveal significant documentation gaps in medical AI products and call for stricter legal requirements to ensure safety and ethical compliance.

# Branches of Trustworthy AI

## **Robustness**

Robustness ensures reliable system performance under varied or adverse conditions, such as noisy inputs or missing data.

## **Explainability**

Explainability makes AI models and their decisions understandable, allowing users to comprehend and trust model decisions.

## **Fairness**

Fairness prevents biases and ensures equitable outcomes for all users.

## **Privacy**

Privacy safeguards sensitive user data throughout the AI lifecycle, maintaining confidentiality and security.

## **Accountability**

Accountability establishes mechanisms for responsibility and oversight, ensuring AI systems operate within ethical and legal boundaries.

# Challenges in Machine Learning System Phases

## 1 Problem Definition

Problem definition requires clear scope and ethical considerations.

## 2 Data Challenges

Data collection faces challenges from noisy data, noisy labels, and missing data.

## 3 Feature Selection

Feature selection necessitates managing multimodality to integrate heterogeneous data sources effectively.

## 4 Model Training

Model training confronts class imbalance, small datasets, and overfitting.

## 5 Model Evaluation

Model evaluation must address underfitting, overfitting, and performance gaps due to distribution shifts.

## 6 Inference Challenges

Inference faces challenges like missing channels/modalities and ensuring explainability.

# Designing Robust ML Models

## Robustness in AI Systems

Robustness is fundamental for AI, ensuring reliable performance across diverse conditions.

## Handling Data Imperfections

Techniques like denoising autoencoders, convolutional networks, and fuzzy c-means clustering effectively handle noisy data, labels, and missing information.

## Masked Autoencoders

Masked autoencoders use self-supervised learning to reconstruct missing data, enhancing model reliability.

## Multimodal Integration & Imbalance

Robustness extends to multimodal data integration and class imbalance, addressed by multimodal deep learning and data balancing techniques.



# Trust Concerns Across Health Domains

## Medical Imaging: Trust & Explainability

Deep learning models in medical imaging raise explainability and safety concerns due to their black-box nature, addressed by techniques like saliency maps and Grad-CAM.

## Cardiovascular Health: Signal Robustness

AI in cardiology, used for arrhythmia detection, relies on robust signal modeling and explainability for clinical workflow alignment.

## Wearables: Data Challenges & Trust

Wearable data in health tracking presents challenges in sensor fidelity and noise, emphasizing online learning and anomaly detection for trustworthiness.

## Metabolic Health: Variability & Privacy

Metabolic health applications integrating CGMs and wearables face challenges like user variability and missing data, mitigated by explainability and privacy-preserving learning.

## Neonatal Health: High Stakes & Transparency

AI in pediatric and neonatal care requires crucial explainability due to data scarcity and high stakes, using what-if analysis and causal feature attribution.

## Mental Health: Scarcity & Personalization

AI for mental health and addiction recovery emphasizes robust learning under label scarcity, personalization, and clinically relevant explanations using self-supervised learning.

# AI in Critical Care & Public Health

## 1 Intensive Care: Prediction & Trust

AI in ICUs supports early prediction of sepsis and patient deterioration, requiring strong generalization and low false alarm rates for clinical trust.

## 2 Public Health: Surveillance & Ethics

Public health AI aids disease surveillance and epidemic forecasting, emphasizing responsible AI frameworks for fairness, transparency, and ethical risk management.

# Label Scarcity and Data-Efficient Learning



## **Prioritizing Robustness & Explainability**

Trustworthy AI in digital health prioritizes robust and explainable systems, crucial for addressing unique healthcare challenges like label scarcity.

## **CUDLE Framework for Label Scarcity**

The CUDLE framework leverages self-supervised learning for accurate health behavior detection with minimal labels, achieving higher accuracy than traditional methods.

## **Clinical Speech AI Development**

Clinical speech AI must integrate insights from speech science, explainable models, and robust validation frameworks to mitigate limited data and overfitting.

# Forecasting and Personalized Interventions

## Ensemble Models in Public Health

Probabilistic forecasting in public health, especially during the COVID-19 pandemic, demonstrated that ensemble models consistently outperformed individual models in mortality rate forecasting.

## Collaborative Modeling Imperative

This highlights the importance of active coordination between public health agencies, academia, and industry for reliable modeling under real-world constraints.

## Personalized Interventions Efficacy

Personalized digital health interventions, such as smartphone and text-message-based systems for managing type 2 diabetes, significantly improved glycemic control compared to website-based interventions.

# Self-Supervised Learning and Cross-Domain Generalization

## Transformative Role of Self-Supervised Learning

Self-supervised learning transforms medical AI by enabling models to learn from large-scale unannotated data across diverse modalities like medical images and bioelectrical signals.

## Addressing Data Challenges for Scalability

These methods address challenges like limited annotated datasets and biased data collection, facilitating the development of scalable and generalizable AI systems.

## Enhancing Trustworthiness via Sensor Redundancy

Inherent sensor redundancies have been exploited to enhance anomaly detection in sensor-based systems, improving trustworthiness and robustness in digital health applications.

# Robustness, Utility, and Oversight

1

## Robust and Explainable Systems

Work in AIMEN and medication adherence forecasting focuses on building robust, explainable systems to improve reliability and clinical utility in neonatal health and treatment adherence.

2

## AI Co-Scientist System

Google's Co-Scientist system introduces a multi-agent framework for automated scientific discovery with human oversight.

3

## Transparency and Accountability

This system exemplifies transparency and accountability through debate-style evaluation and citation-based reasoning.

4

## Human Oversight and Domain Knowledge

The inclusion of scientists ensures explainability and alignment with domain knowledge, supporting responsible deployment in high-stakes fields.



# Explainability in Machine Learning

## Goal of Explainable AI

Explainable AI (XAI) methods aim to make complex models transparent, bridging the gap between high-performing opaque models and responsible real-world deployment.

## Range of Techniques

Techniques range from feature importance scores to counterfactual explanations, crucial in critical fields like healthcare.

# Explainable AI Methods

## Feature Importance: Shapley & SHAP

Shapley values and SHAP quantify feature contribution to predictions, ensuring fairness and consistency.

## Counterfactual Explanations: Actionable Insights

Counterfactual explanations show how altering input features changes outcomes, providing actionable insights for decision-making.

## Practical Tools: DiCE & NICE

Tools like DiCE and NICE generate diverse, intuitive, and transparent counterfactuals, enhancing explainability and usability.

## Applications in Healthcare

These methods have been applied in critical domains like neonatal health and multimodal hyperglycemia prediction, improving interpretability and safety.

# Counterfactual Explanations in XAI

1

## **Semi-Factual Explanations**

Semi-factual explanations highlight input changes that do not alter a model's output, offering insights into prediction stability and enhancing trust.

2

## **Robust Counterfactuals for GNNs**

Robust counterfactual explanations for Graph Neural Networks (GNNs) ensure stability in noisy environments by identifying influential subgraphs.

3

## **NICE for Tabular Data**

NICE generates counterfactuals for tabular data, optimizing for sparsity, proximity, and plausibility to ensure efficient and applicable explanations.

# Benchmarking & Frameworks for Counterfactuals

1

## Benchmarking Counterfactuals

Benchmarking efforts categorize counterfactual explanation techniques based on stability, diversity, and actionability, highlighting trade-offs.

2

## Balancing Properties for Actionability

Analysis calls for methods that balance multiple desirable properties for actionable and efficient explanations.



# Applications of Counterfactual Explanations in Healthcare

## Role in Healthcare

Counterfactual explanations are vital in healthcare systems like AIMEN and GlucoLens, providing actionable insights for clinicians and patients.

## Enhancing Interpretability and Trust

They demonstrate effects of interventions or behavioral changes, improving interpretability and enhancing user trust.

## Alignment with AI Transparency

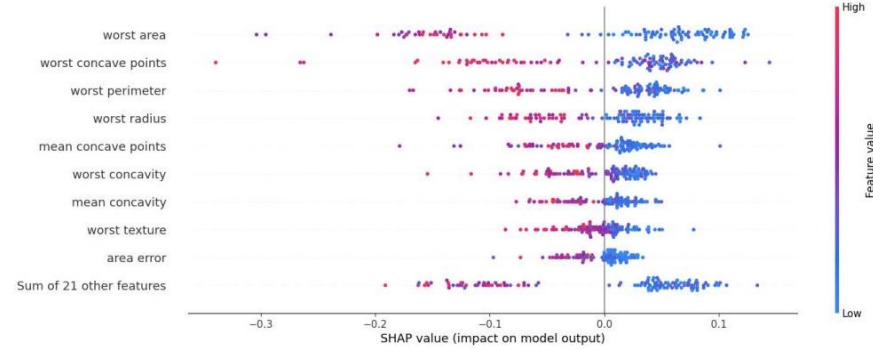
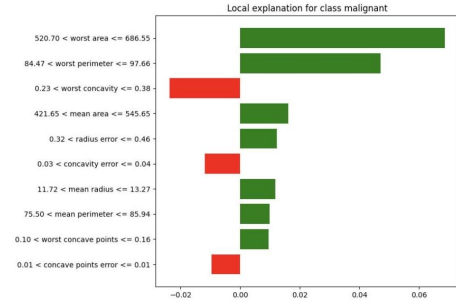
Counterfactuals align with the need for transparent and user-friendly AI in high-stakes domains.

## Local Interpretable Model-agnostic Explanations

LIME provides explanations for predictions by fitting interpretable models locally, perturbing input data to create a neighborhood and observing black-box model behavior.

### Feature Contribution & Flexibility

It highlights individual feature contributions and is flexible across any machine learning model.



# Shapley Values

## 1 Cooperative Game Theory Basis

Shapley values, from cooperative game theory, fairly distribute total value among contributors (features) by calculating marginal contribution averaged over all permutations.

## 2 Theoretical Foundation for SHAP

They form the theoretical basis for tools like SHAP, ensuring fairness and consistency in feature attribution, despite being computationally intensive.

# GradCAM (Gradient-weighted Class Activation Mapping)



1

## Visual Explanations for CNNs

GradCAM generates visual explanations for CNNs by highlighting important regions in input images using gradients of the model's output with respect to feature maps.

2

## Intuitive Visual Interpretation

Heatmaps overlay input images to show areas influencing model decisions, providing intuitive interpretation for visual data.

# Integrated Gradients

## 1 Axiomatic Attribution

Integrated Gradients is an attribution technique that explains deep neural network predictions in a theoretically grounded manner.

## 3 Stable and Interpretable Explanations

The resulting attribution map highlights influential input features, offering stable and interpretable explanations.

## 2 Feature Importance Estimation

It estimates feature importance by accumulating gradients along a straight-line path from a baseline to the actual input.

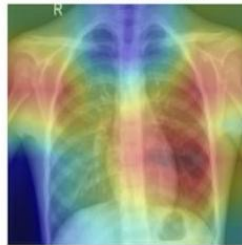
## 4 Value in Sensitive Domains

It is valuable in sensitive domains like medical imaging, requiring transparent decision-making.

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



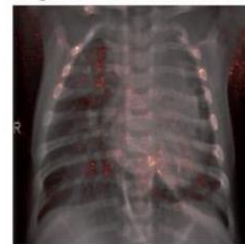
Original X-ray  
True: PNEUMONIA  
Pred: PNEUMONIA (Confidence: 1.00)



GradCAM Overlay



Integrated Gradients Overlay



# NICE (Nearest Instance Counterfactual Explanations)

## **Nearest Instance Counterfactuals**

NICE generates counterfactual explanations by identifying the nearest instance in the feature space that results in a different prediction.

## **Distance Minimization**

It minimizes distance while ensuring the counterfactual belongs to a different class.

## **Actionable and Realistic Explanations**

This approach ensures interpretable, realistic, and actionable explanations aligned with data distribution.

# DiCE (Diverse Counterfactual Explanations)



## **Minimal Feature Changes**

DiCE generates counterfactual explanations to show how minimal changes to input features can alter a model's prediction.

## **Diverse Alternatives for Users**

It creates diverse explanations, offering multiple plausible alternatives for user exploration.

## **Suitability for Sensitive Domains**

DiCE balances proximity, diversity, and feasibility, making it suitable for sensitive domains like finance and healthcare where interpretability is crucial.

# CFNOW

## Two-Step Optimization

CFNOW uses a two-step optimization: CF search finds an initial solution for classification change, then CF improvement refines it by minimizing distance.

## Flexible Approach

It supports greedy and random approaches, prioritizing speed or refinement for plausibility.

## Versatile Data Handling

CFNOW processes diverse data types (tabular, image, text) for binary and multiclass tasks, offering flexible and adaptable counterfactual explanations.



# Trustworthy AI in the Era of LLMs

## 1 LLMs Reshaping Domains

LLMs like ChatGPT are reshaping research, education, and medicine by enhancing information access and decision-making.

## 2 Multimodal and Multilingual LLMs

Advancements like Llama 3 show strides in multilingual support, coding, and multimodal functionalities, enabling diverse applications.

## 3 RAG for Enhanced Accuracy

Retrieval-augmented generation (RAG) combines parametric and non-parametric memories to enhance specificity and factual accuracy in language generation.

## 4 Advanced Reasoning Models

Reasoning models like DeepSeek-R1 refine logical inference, while chain-of-thought prompting reveals untapped cognitive potential.

## 5 Transparency in LLMs

Transparency in LLMs involves understanding how models arrive at conclusions, a challenge given their complex nature.

## 6 Fairness and Bias Mitigation

Fairness in LLMs requires detecting and mitigating biases to ensure equitable outcomes, especially in sensitive domains like healthcare.

# Metrics of Evaluation

## 1 Importance of Evaluation

Evaluating trustworthy AI methods ensures utility, reliability, and alignment with user needs.

## 2 Assessing Explanation Quality

Effective metrics assess accuracy, simplicity, and robustness of explanations, providing a comprehensive understanding of strengths and limitations.

# Trust Metrics

1

## Quantifying Trust Influence

The trust coefficient quantifies how human decisions are influenced by ML model predictions relative to ground truth labels.

2

## Interpreting Trust Coefficient

A coefficient greater than one suggests over-reliance, while a value below one indicates skepticism.

3

## Survey-Based Trust Assessment

Survey-based assessments provide a validated framework for measuring public trust and openness toward AI in healthcare.



# Conclusion

## Foundational Properties

Trustworthy AI in digital health requires addressing robustness and explainability as foundational properties.

## Synthesized Developments

The review synthesizes recent developments in algorithmic innovations, evaluation frameworks, and practical deployment challenges.

## Robustness in Clinical Settings

Robustness in clinical settings must account for data distribution shifts, sensor noise, and adversarial scenarios.

## Explainability for Clinicians

Explainability must go beyond model transparency to support clinician understanding, justification, and decision-making.

## Verifiable Trustworthiness

Trustworthiness is a composite of verifiable properties that must be operationalized and measured systematically.

## Societal Imperative

Advancing trustworthy AI in digital health is not only a technical challenge but a societal imperative.

# *Thank You!*



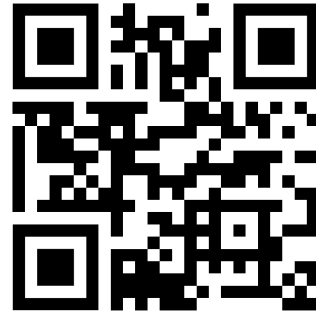
*Email: [a.mamun@asu.edu](mailto:a.mamun@asu.edu)*



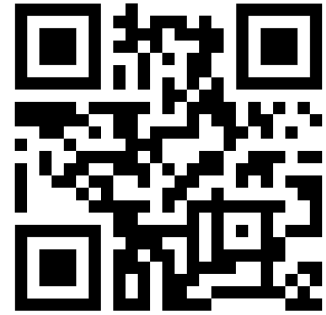
*Paper*



*Preprint*



*abdullah-mamun.com*



*X: @AB9Mamun*