www.ijarp.com ISSN 2456-9992 Page: 01-12

International Journal Advanced Research Publications

Review Article

Volume: 02
Issue: 08

EXPLAINABLE ARTIFICIAL INTELLIGENCE IN HEALTHCARE: A SURVEY OF METHODS, APPLICATIONS, CHALLENGES, AND FUTURE DIRECTIONS

Pranav Kumar Bhadane*[1], Dr. Abhinav Sudhir Thorat[2]

[1]PhD Research Scholar, Department of Computer Science and Engineering, Srinivas University Institute of Engineering and Technology, Mangalore, Karnataka, India.

[2]Assistant Professor, Department of Computer Engineering, Amrutvahini College of Engineering, Maharashtra, India.

Article Received: 20 July 2026, Article Revised: 10 August 2026, Published on: 30 August 2026

***Corresponding Author: Pranav Kumar Bhadane**

PhD Research Scholar, Department of Computer Science and Engineering, Srinivas University Institute of Engineering and Technology, Mangalore, Karnataka, India.

Doi: https://doi-doi.org/101555/ijarp.2807

ABSTRACT

Artificial Intelligence (AI) has significantly transformed healthcare by enabling accurate disease diagnosis, medical image analysis, clinical decision support, personalized medicine, and drug discovery. Despite these advancements, the widespread adoption of AI in clinical practice is limited by the lack of transparency and interpretability of complex machine learning and deep learning models. Explainable Artificial Intelligence (XAI) addresses this challenge by providing human-understandable explanations that improve the transparency, trustworthiness, and reliability of AI-assisted decision-making. This survey presents a concise review of XAI methods and their role in developing trustworthy healthcare systems. It discusses the fundamental concepts of AI, Machine Learning (ML), Deep Learning (DL), Explainable AI, Interpretable AI, Trustworthy AI, and Responsible AI, followed by a taxonomy of XAI techniques based on model type, explanation scope, timing, and explanation type. The survey compares widely used explainability methods, including LIME, SHAP, Grad-CAM, Integrated Gradients, and counterfactual explanations, highlighting their advantages, limitations, and healthcare applications. It further reviews the use of XAI in medical imaging, Electronic Health Records (EHRs), clinical decision support, disease prediction, and personalized medicine, along with benchmark public datasets and evaluation metrics. Finally, the paper discusses current challenges, identifies key research gaps, and outlines future research directions, including self-explainable AI, multimodal XAI, Explainable Large Language Models (LLMs), Agentic AI, and federated explainable learning.

This survey provides researchers and healthcare professionals with a comprehensive overview of the current state of XAI and highlights promising directions for developing transparent, trustworthy, and clinically deployable AI systems.

KEYWORDS: Explainable Artificial Intelligence (XAI), Healthcare, Machine Learning, Deep Learning, Medical Imaging, Clinical Decision Support, Trustworthy AI, Large Language Models.

1. INTRODUCTION

Artificial Intelligence (AI) has emerged as one of the most influential technologies in modern healthcare, transforming disease diagnosis, treatment planning, patient monitoring, and hospital management. The rapid growth of Electronic Health Records (EHRs), medical imaging, wearable devices, genomic sequencing, and Internet of Medical Things (IoMT) technologies has generated massive amounts of healthcare data. Traditional statistical methods often struggle to analyse these heterogeneous and high-dimensional datasets efficiently. Machine Learning (ML) and Deep Learning (DL) algorithms have demonstrated remarkable capabilities in extracting meaningful patterns from medical data, enabling accurate disease prediction and supporting evidence-based clinical decision-making. As a result, AI has become an essential component of healthcare applications such as medical image analysis, clinical decision support systems, personalized medicine, drug discovery, and remote patient monitoring.[11],[12]

Among these applications, medical imaging has experienced significant advancements through AI-based models. Deep learning architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Vision Transformers (ViTs), and transformer-based models have achieved high accuracy in detecting diseases including pneumonia, breast cancer, brain tumours, diabetic retinopathy, tuberculosis, and cardiac abnormalities using X-ray, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), retinal images, and Electrocardiogram (ECG) signals. AI has also enhanced Natural Language Processing (NLP) applications for analysing clinical notes, discharge summaries, and electronic health records, thereby improving healthcare delivery and patient management.[14],[15]

Despite these achievements, the adoption of AI in clinical practice remains limited by the lack of transparency and interpretability of complex deep learning models. Many modern AI systems operate as black boxes, providing accurate predictions without clearly explaining

their reasoning. In healthcare, where decisions directly affect patient safety and treatment outcomes, clinicians need both reliable predictions and understandable explanations. Explainable Artificial Intelligence (XAI) addresses this challenge by improving the transparency and trustworthiness of AI systems through feature importance, image-region highlighting, decision rules, and counterfactual explanations. Techniques such as LIME, SHAP, Grad-CAM, and Integrated Gradients help clinicians understand AI predictions, validate model outputs, identify potential biases, and support informed clinical decision-making.[6],[9]

Trustworthiness is essential for AI-enabled healthcare because clinicians remain responsible for patient care, making AI a decision-support tool rather than a replacement for human expertise. XAI supports transparency, accountability, fairness, and regulatory compliance by helping clinicians verify AI recommendations and identify potential biases or uncertainties. However, challenges such as lack of standardized evaluation metrics, computational complexity, privacy concerns, dataset bias, and limited clinical integration remain. Emerging technologies including multimodal AI, Large Language Models (LLMs), knowledge graphs, federated learning, and Agentic AI create new opportunities while introducing additional explainability challenges. This survey reviews key XAI concepts, techniques, healthcare applications, datasets, evaluation metrics, challenges, and research gaps, with particular focus on self-explainable models, multimodal XAI, explainable LLMs, federated XAI, and uncertainty-aware clinical decision support. [6],[9],[18]

2. Background

Artificial Intelligence (AI) is a transformative technology in healthcare, supporting disease diagnosis, clinical decision-making, personalized medicine, patient monitoring, and hospital management. The increasing availability of medical data from Electronic Health Records (EHRs), medical imaging, wearable devices, genomic sequencing, and physiological sensors has created a growing need for intelligent systems that can analyse complex information and provide meaningful insights. However, the increasing complexity of AI models has raised concerns about transparency, fairness, accountability, and trust, leading to the development of Machine Learning (ML), Deep Learning (DL), Explainable AI (XAI), Interpretable AI, Trustworthy AI, and Responsible AI. In healthcare, AI analyses large volumes of medical data to support accurate diagnosis, evidence-based treatment, clinical decision support, drug discovery, and efficient healthcare delivery. [11],[12],[15]

Machine Learning (ML) is a subset of AI that enables systems to learn patterns from historical data and make predictions without explicit programming. ML algorithms are extensively used for disease prediction, patient risk assessment, cancer detection, diabetes prediction, and medical image classification. Based on the learning process, ML is categorized into supervised learning, unsupervised learning, and reinforcement learning. Supervised learning utilizes labelled datasets for classification and prediction tasks, whereas unsupervised learning discovers hidden patterns in unlabelled data. Reinforcement learning allows intelligent agents to learn optimal actions through interactions with their environment and is increasingly applied in personalized treatment planning and robotic surgery. While many traditional ML models, such as decision trees and logistic regression, are relatively interpretable, complex ensemble methods often sacrifice interpretability for improved predictive performance.[5],[6]

Deep Learning (DL), a specialized branch of machine learning, employs multi-layer artificial neural networks to automatically learn hierarchical features from large datasets. Unlike traditional ML techniques, deep learning eliminates the need for manual feature engineering by learning directly from raw data. Deep learning models compute outputs by transforming input features through weighted connections and activation functions. The basic mathematical representation of a neural network neuron is given by

$$y = f(Wx + b)$$

Where x is the input feature vector, W represents the weight matrix, b denotes the bias term, and f( ) is the activation function such as ReLU or Sigmoid. This formulation enables deep neural networks to learn complex nonlinear relationships from healthcare data. Deep learning architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, Transformers, and Vision Transformers (ViTs) have demonstrated remarkable success in healthcare applications, including X-ray, CT, and MRI image analysis, ECG interpretation, disease classification, and clinical text analysis. Despite their exceptional predictive performance, these models often function as black-box systems, making it difficult to understand how predictions are generated and limiting their acceptance in clinical practice. [12],[15]

For multi-class disease classification, deep learning models commonly use the Softmax function to convert output scores into class probabilities.

$$P(y = i) = \frac{e^{zi}}{\sum_{j=1}^{K} e^{zi}}$$

Where zi represents the output score for class and is the total number of disease classes.

Fig. 1 XAI **Workflow** illustrates the overall work flow of Explainable Artificial Intelligence in healthcare. Patient data are processed by AI or deep learning models to generate predictions, which are subsequently interpreted using XAI techniques. The resulting explanations assist clinicians in making transparent, reliable, and trustworthy medical decisions.

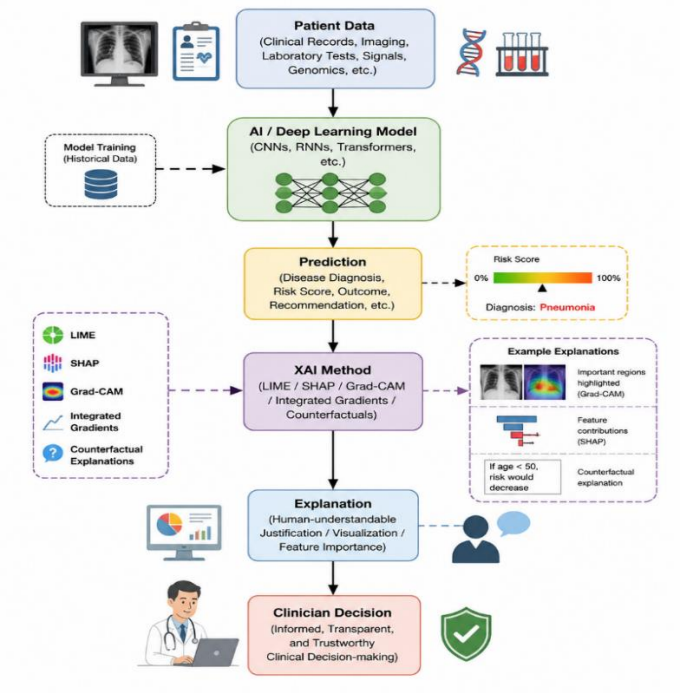

Fig. 1 XAI Workflow.

Explainable Artificial Intelligence (XAI) addresses the limitations of complex AI models by providing human-understandable explanations for their predictions. It identifies important features, highlights relevant regions in medical images, and explains factors influencing model decisions. Common techniques include SHAP, LIME, Grad-CAM, Integrated Gradients, attention visualization, and counterfactual explanations. These methods can help clinicians validate predictions, identify potential biases, improve transparency, and increase confidence in AI-assisted decisions. XAI can be classified as intrinsic or post-hoc, local or global, and model-specific or model-agnostic. Interpretable AI uses inherently transparent models such as decision trees and linear regression, while Trustworthy AI emphasizes

reliability, fairness, robustness, privacy, security, and human oversight. Responsible AI further incorporates ethical governance, regulatory compliance, inclusiveness, and patient safety, supporting the development of transparent and clinically reliable healthcare systems. [6],[9],[16],[17]

3. Explainability Taxonomy and Comparison of Major XAI Techniques

Explainable Artificial Intelligence (XAI) comprises a diverse set of techniques designed to improve the transparency and interpretability of Artificial Intelligence models. In healthcare, selecting an appropriate explainability method depends on the underlying AI model, the level of explanation required, and the target clinical application. XAI methods are commonly classified based on four perspectives: model type, explanation scope, timing of explanation, and explanation type. Based on model type, AI models are categorized as white-box or black-box. White-box models, such as Decision Trees, Linear Regression, Logistic Regression, and Rule-Based Systems, are inherently interpretable and allow users to understand the reasoning behind predictions directly. These models are easy to validate and suitable for clinical decision-making but may have lower predictive performance for complex healthcare data. In contrast, black-box models, including Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Vision Transformers (ViTs), provide high predictive accuracy but lack transparency, making their decisions difficult to interpret.

Based on the explanation scope, XAI methods provide either global or local explanations. Global explanations describe the overall behaviour of an AI model, such as identifying the most influential clinical features across an entire dataset, whereas local explanations explain individual predictions, such as highlighting the lung region responsible for diagnosing pneumonia from a chest X-ray.[1],[2]

According to the timing of explanation, methods are classified as intrinsic or post-hoc. Intrinsic explainability is achieved through inherently interpretable models, while post-hoc methods generate explanations after prediction using techniques such as SHAP [2], LIME [1], Grad-CAM [3], and Integrated Gradients [4]. XAI methods can also be categorized by explanation type, including visual explanations, feature importance, rule-based, counterfactual, example-based, and concept-based explanations, each offering different levels of interpretability for healthcare applications.[10]

Several XAI techniques have been widely adopted in healthcare. LIME and SHAP are model-agnostic approaches suitable for disease prediction and clinical decision support, while Grad-

CAM and Integrated Gradients are extensively used for interpreting medical imaging models. TCAV provides concept-based explanations, whereas Counterfactual Explanations offer actionable recommendations by showing how slight changes in patient features can alter predictions. Attention mechanisms, Deep LIFT, Anchors, and Occlusion Sensitivity further enhance model interpretability across medical imaging, Electronic Health Records (EHRs), and Natural Language Processing (NLP). Although these techniques improve transparency and clinician trust, they differ in computational complexity, stability, and applicability. Therefore, selecting an appropriate XAI technique requires balancing interpretability, computational efficiency, and clinical usefulness according to the specific healthcare application.[10]

Table 1 Survey of Existing Research on XAI in Healthcare.

Technique / Paper	Year	Model Type	Explanation Type	Key Contribution	Advantages	Limitations	Healthcare Applications
LIME (Why Should I Trust You?)	2016	Model-Agnostic	Local	Explains individual predictions using locally interpretable surrogate models	Simple, model-independent, easy to interpret	Unstable explanations; computational overhead	Disease prediction, clinical decision support, EHR analysis
SHAP (A Unified Approach to Interpreting Model Predictions)	2017	Model-Agnostic	Local & Global	Uses Shapley values to quantify feature importance	Consistent, theoretically sound, accurate explanations	Computationally expensive for large models	Medical diagnosis, risk prediction, personalized medicine
Grad-CAM	2017	CNN-Based	Visual	Generates heatmaps highlighting image regions influencing predictions	Easy visualization, suitable for CNNs	Limited to image-based deep learning models	X-ray, CT, MRI, tumour detection
Integrated Gradients	2017	Deep Learning	Feature Attribution	Computes feature importance using integrated gradients	Faithful attribution, mathematically justified	Sensitive to baseline selection; computationally intensive	Medical imaging, ECG analysis, disease classification
XAI Taxonomy Framework	2020	General AI	Taxonomy	Provides comprehensive classification of XAI	Comprehensive overview of XAI landscape	Does not introduce a new explanation algorithm	General healthcare AI research

				methods, challenges, and opportunities			
Black-Box Explanation Survey	2019	General AI	Survey	Reviews model-specific and model-agnostic explanation methods	Comprehensive comparison of existing techniques	No experimental evaluation	AI transparency and healthcare applications
Medical XAI Survey	2021	Healthcare AI	Survey	Reviews XAI techniques specifically for healthcare	Highlights clinical applications and challenges	Limited discussion on LLMs and foundation models	Medical diagnosis, radiology, pathology
Explainable Deep Learning	2019	Deep Learning	Visual & Feature-Based	Summarizes visualization and interpretation methods for deep neural networks	Covers multiple XAI techniques for DL	General overview rather than application-specific	Medical imaging and deep learning systems
Medical Large Language Models Survey	2024	Medical LLMs	Trustworthy AI & XAI	Reviews explainability, safety, and trustworthiness of Medical LLMs	Covers latest foundation models in healthcare	Explainability evaluation remains qualitative	Clinical assistants, medical NLP, patient care
XAI Meets LLMs	2024	GPT-4, LLaMA, Claude	SHAP, LIME, Attention, Causal Analysis	Bridges traditional XAI with Large Language Models	Introduces explainability for foundation models	Lack of standardized evaluation metrics	Medical AI, Clinical LLMs, Decision Support

4. Healthcare Applications, Public Datasets, and Evaluation Metrics

Explainable Artificial Intelligence (XAI) has become an essential component of modern healthcare by improving the transparency, reliability, and trustworthiness of AI-assisted clinical decisions. XAI is extensively applied in medical imaging, where techniques such as Grad-CAM, SHAP, and Integrated Gradients highlight clinically significant regions in X-ray, MRI, CT, and ultrasound images, supporting the diagnosis of diseases such as pneumonia, brain tumours, stroke, and lung cancer. Beyond imaging, XAI enhances ECG analysis, Electronic Health Records (EHRs), Clinical Decision Support Systems (CDSS), cancer detection, diabetes prediction, neurological disease diagnosis, drug discovery, ICU patient monitoring, and personalized medicine by providing interpretable explanations that help

clinicians validate AI predictions and make informed treatment decisions. The performance of AI models is commonly evaluated using Accuracy and F1-score.

Accuracy

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

where TP, TN, FP and FN denote true positives, true negatives, false positives, and false negatives, respectively.

F1-score

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The F1-score provides a balanced evaluation by combining precision and recall, making it particularly useful for imbalanced healthcare datasets.

The development and evaluation of XAI models rely on publicly available healthcare datasets such as MIT-BIH, ChestX-ray14, CheXpert, MIMIC-IV, BraTS, HAM10000, and EyePACS, which provide diverse data for medical imaging, physiological signal analysis, and electronic health records. These benchmark datasets facilitate the development of robust and reproducible AI models. Model evaluation involves both predictive performance and explanation quality. Traditional AI metrics, including Accuracy, Precision, Recall, F1-score, and AUC, assess predictive capability, whereas explainability metrics such as Fidelity, Faithfulness, Stability, Robustness, Consistency, Human Interpretability, and Clinical Utility evaluate the reliability and usefulness of generated explanations. Together, these metrics ensure that AI systems are not only accurate but also transparent, trustworthy, and clinically applicable for real-world healthcare deployment.

5. CHALLENGES

Despite significant progress, several challenges hinder the widespread adoption of Explainable Artificial Intelligence (XAI) in healthcare. Complex black-box models often lack transparency, reducing clinician trust in AI-assisted decisions. The absence of standardized evaluation frameworks makes it difficult to assess the quality and reliability of explanations. Additional concerns include dataset bias, fairness across diverse patient populations, privacy and security of sensitive medical data, and compliance with healthcare regulations. Furthermore, many explainability techniques are computationally expensive and difficult to

integrate into existing hospital workflows. Addressing these challenges is essential for developing transparent, trustworthy, efficient, and clinically deployable AI systems.[13],[17]

6. Future Research Directions and Research Gap Analysis

Although Explainable Artificial Intelligence (XAI) has improved transparency in healthcare AI, several challenges remain. Future research should focus on self-explainable AI, multimodal XAI, Explainable Large Language Models (LLMs), Agentic AI, and knowledge graph-enhanced XAI to generate clinically meaningful explanations from medical images, EHRs, laboratory, and genomic data. Other promising areas include counterfactual explanations, patient-friendly XAI, uncertainty-aware XAI, federated learning, and real-time XAI for ICUs. Standardized evaluation frameworks are also needed to assess explanation quality and clinical utility.

The research gap analysis shows that existing studies often prioritize predictive accuracy over explanation quality, clinician usability, and real-world deployment. Multimodal integration, uncertainty estimation, patient-centered explanations, standardized evaluation, and clinical validation remain insufficiently explored. Addressing these gaps can support the development of trustworthy, transparent, and clinically applicable XAI systems for next-generation healthcare. [18],[19]

7. CONCLUSION

Explainable Artificial Intelligence (XAI) has become a critical component for developing transparent, trustworthy, and clinically reliable AI systems in healthcare. Although Artificial Intelligence, Machine Learning, and Deep Learning have significantly improved disease diagnosis, medical imaging, clinical decision support, drug discovery, and personalized medicine, their black-box nature limits clinician trust and widespread adoption. XAI overcomes this limitation by providing interpretable explanations that enable healthcare professionals to understand, validate, and confidently use AI-generated predictions.

This survey reviewed the fundamental concepts of XAI, its taxonomy, major explainability techniques, healthcare applications, benchmark datasets, evaluation metrics, current challenges, and future research directions. The analysis indicates that no single XAI technique is universally applicable, and the choice of method depends on the AI model, data modality, and clinical requirements. Despite notable progress, challenges such as the lack of standardized evaluation frameworks, limited multimodal explainability, insufficient clinical validation, privacy concerns, and workflow integration remain unresolved. Future research

should focus on Explainable Large Language Models (LLMs), Agentic AI, self-explainable models, federated learning, and uncertainty-aware AI to develop ethical, transparent, and clinically deployable healthcare systems that improve patient outcomes and strengthen trust in AI-assisted decision-making.

8. REFERENCES

1. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier," Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), pp. 1135–1144, 2016.
2. S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017.
3. R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," Proc. IEEE Int. Conf. Computer Vision (ICCV), pp. 618–626, 2017.
4. M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," Proc. 34th Int. Conf. Machine Learning (ICML), pp. 3319–3328, 2017.
5. F. Doshi-Velez and B. Kim, "Towards a Rigorous Science of Interpretable Machine Learning," arXiv:1702.08608, 2017.
6. A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," Information Fusion, vol. 58, pp. 82–115, 2020.
7. R. Guidotti et al., "A Survey of Methods for Explaining Black Box Models," ACM Computing Surveys, vol. 51, no. 5, pp. 1–42, 2019.
8. A. Holzinger et al., "What Do We Need to Build Explainable AI Systems for the Medical Domain?" arXiv:1712.09923, 2017.
9. E. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 11, pp. 4793–4813, 2021.
10. W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K. R. Müller, Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer, 2019.
11. E. J. Topol, "High-Performance Medicine: The Convergence of Human and Artificial Intelligence," Nature Medicine, vol. 25, pp. 44–56, 2019.
12. A. Esteva et al., "A Guide to Deep Learning in Healthcare," Nature Medicine, vol. 25, pp. 24–29, 2019.

13. C. J. Kelly et al., "Key Challenges for Delivering Clinical Impact with Artificial Intelligence," BMC Medicine, vol. 17, no. 195, 2019.
14. P. Rajpurkar et al., "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," arXiv:1711.05225, 2017.
15. R. Miotto et al., "Deep Learning for Healthcare: Review, Opportunities and Challenges," Briefings in Bioinformatics, vol. 19, no. 6, pp. 1236–1246, 2018.
16. European Commission, Ethics Guidelines for Trustworthy AI, Brussels, Belgium, 2019.
17. National Institute of Standards and Technology (NIST), AI Risk Management Framework (AI RMF 1.0), NIST AI 100-1, 2023.
18. H. Li et al., "A Survey on Medical Large Language Models: Technology, Application, Trustworthiness and Future Directions," arXiv:2406.03712, 2024.
19. X. Zhao et al., "XAI Meets LLMs: A Survey of the Relation between Explainable AI and Large Language Models," arXiv:2407.15248, 2024.
20. D. Li et al., "Explainability for Large Language Models: A Survey," ACM Computing Surveys, 2024.