

THE POTENTIAL OF XAI TO REVOLUTIONIZE HEALTHCARE: APPLICATIONS AND CHALLENGES

Akanksha Sarwan¹, Sandeep Sarwan²

¹Lecturer, Department of Electronics and Communication, India.

²Government Polytechnic, Ambikapur, Surguja, Chhattisgarh, India.

DOI: <https://www.doi.org/10.58257/IJPREMS39163>

ABSTRACT

Explainable Artificial Intelligence (XAI) is crucial in healthcare, enhancing transparency and trust in AI-driven medical applications. Traditional AI models often function as "black boxes," offering limited insight into their decision-making processes, which can hinder clinician and patient confidence. XAI addresses this by providing clear, interpretable explanations, thereby improving the reliability of AI-assisted healthcare. XAI employs various techniques to clarify AI models' operations. These include intrinsic methods, which design inherently interpretable models like decision trees, and post-hoc methods that explain outputs of complex models after training. For instance, Local Interpretable Model-agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) analyze input-output relationships to determine each feature's contribution to a prediction, enhancing model transparency. XAI techniques improve interpretability in analyzing medical images, such as X-rays and MRIs. In clinical decision support systems, XAI offers transparent reasoning behind AI-generated diagnoses, aiding clinicians. It aids in identifying potential drug candidates by elucidating how AI models predict molecular interactions and efficacy, thereby accelerating drug development and enhancing understanding of biological mechanisms. XAI has the potential to transform healthcare by making AI systems more transparent and trustworthy. Addressing challenges related to complexity, standardization, and ethics is essential for successful XAI integration into clinical practice, ultimately leading to improved patient outcomes and increased clinician confidence in AI-assisted healthcare solutions.

Keywords: Explainable Artificial Intelligence, Machine Learning, SHAP, LIME, Grad-CAM, XAI in healthcare

1. INTRODUCTION

Artificial Intelligence (AI) is transforming healthcare by enabling faster, more accurate diagnoses, personalized treatments, and predictive analytics. AI systems analyze vast amounts of data, such as medical images, electronic health records, and genomic information, to identify patterns and provide actionable insights[18]. For example, AI-powered tools can detect abnormalities in radiology scans or predict disease progression, improving clinical efficiency and patient outcomes. As healthcare systems face challenges like aging populations and rising costs, AI offers scalable solutions to optimize resources and enhance care delivery. However, the adoption of AI in healthcare raises significant concerns, particularly around the "black box" nature of many algorithms. In healthcare, where decisions impact lives, transparency is critical. Explainable Artificial Intelligence (XAI) was created to address this issue, with the goal of making AI processes clearer and more trustworthy for users, regardless of their technical background [2]. Clinicians need to trust and validate AI recommendations, ensuring they align with medical expertise. Explainability also helps address ethical and legal concerns, such as bias and compliance with regulations like GDPR and HIPAA. FUTURE-AI is a structured framework that provides guiding principles and step-by-step recommendations for operationalising trustworthy and ethical AI in healthcare[1]. Additionally, patients have the right to understand AI-driven diagnoses and treatments, fostering trust and shared decision-making.

1.1 Importance of Explainable AI (XAI)

The increasing use of AI in healthcare has highlighted significant challenges associated with black-box AI models, which offer predictions and decisions without clear, understandable explanations for their processes. These opaque models create barriers to understanding how decisions are made, leading to potential mistrust and reluctance among healthcare professionals to rely on AI systems [3][17]. In healthcare, decisions directly impact patient lives, making transparency, interpretability, and trust essential components of any technology, including artificial intelligence (AI). Transparency ensures that the processes and data used by AI systems are clear and accessible, allowing clinicians to understand how conclusions are reached. Without transparency, healthcare providers may hesitate to rely on AI, fearing errors or biases. Interpretability, a key aspect of explainable AI (XAI), ensures that AI outputs are understandable to humans. For instance, if an AI system recommends a specific treatment, clinicians must comprehend the reasoning behind it to make informed decisions. This is especially critical in high-stakes scenarios like diagnosing diseases or predicting patient outcomes.

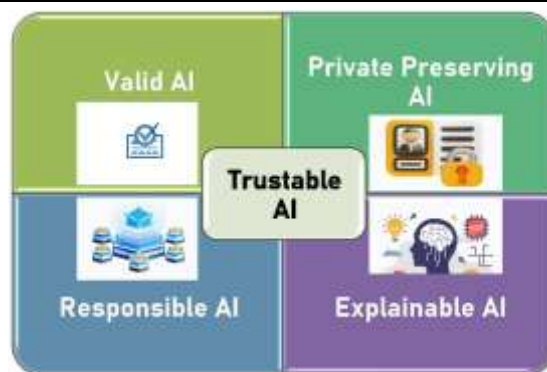


Figure 1- Attributes of trustable AI [5]

Figure 1 represents the attributes of trustable AI. AI can no longer be regarded as a "black box" that receives input and generates output without a clear grasp of what is happening within. People must understand how an AI system arrives at its conclusions and suggestions to trust its decisions, whether ethical or not [5]. Trust is the foundation of the clinician-patient relationship, and AI systems must earn this trust to be effectively integrated into healthcare. Patients and providers need assurance that AI tools are reliable, unbiased, and aligned with ethical standards. Transparent and interpretable AI systems foster this trust, enabling clinicians to confidently use AI while ensuring patients feel secure in their care. Figure 2 timelines the important events in EXAI adoption in healthcare verticals to subsequent integration with EXAI along with a shift from fourth-generation (4G) to 5G wireless networks [5].

1.2 Objectives of the Review

The primary objective of the review is to evaluate the various XAI techniques applied in healthcare. Understanding these methodologies helps in identifying their strengths and limitations within medical contexts.

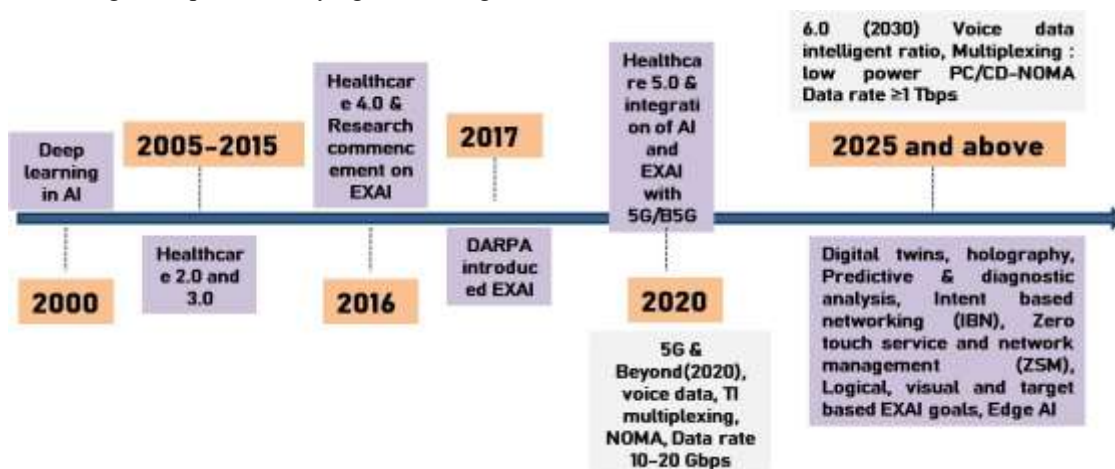


Figure 2- Timeline of events in EXAI [5]

Another key goal is to explore the diverse applications of XAI in healthcare, encompassing areas such as medical imaging, patient care, drug discovery, and administrative tasks. By reviewing these applications, researchers can determine how XAI enhances clinical workflows and patient outcomes. A critical objective is to identify the challenges associated with implementing XAI in healthcare, including technical issues like model complexity and data quality, as well as ethical concerns such as patient privacy and algorithmic bias. Addressing these challenges is vital for the successful integration of XAI in medical practice. The review aims to assess the impact of XAI on healthcare outcomes, analyzing how explainable AI models influence clinical decision-making, patient trust, and overall healthcare quality. Understanding this impact is crucial for justifying the adoption of XAI technologies in medical settings. Lastly, the review seeks to provide insights that can guide future research directions in XAI for healthcare, synthesizing current knowledge and identifying gaps to focus on areas that require further exploration to advance the field.

2. EXPLAINABLE AI ALGORITHMS

Some of the popular XAI algorithms are described briefly:

2.1 LIME (Local Interpretable Model-agnostic Explanations)

LIME is designed to generate localized explanations for individual predictions rather than interpreting the entire model. It operates by analyzing specific instances, requiring users to specify the number of instances to evaluate. The algorithm excels with inherently interpretable models, such as decision trees. LIME employs two primary techniques: SP-

LIME (Selective Pick LIME) and **RP-LIME** (Random Pick LIME). SP-LIME strategically selects a representative subset of instances for interpretation, while RP-LIME randomly samples instances, both aiming to minimize the number of cases needed to explain model behavior effectively [21].

2.2 SHAP (SHapley Additive exPlanations)

SHAP is a versatile XAI method capable of providing both instance-specific (local) and holistic (global) explanations. It establishes a baseline prediction (mean value) and quantifies how each feature contributes to deviations from this baseline. Inspired by cooperative game theory, SHAP assigns an "importance score" to features, where higher scores reflect greater influence on the prediction. Additionally, it distinguishes whether each feature's impact is positive or negative relative to the base value. This additive approach ensures transparency in understanding feature contributions [20].

2.3 Anchor

Anchor identifies minimal sets of feature conditions (called "anchors") that guarantee high-confidence predictions for specific instances. These conditions act as decision rules, ensuring that any input meeting the anchor criteria will yield the same output with high precision. The algorithm is model-agnostic, making it applicable to diverse data types, including text, images, and tabular data, for classification tasks. By isolating critical feature thresholds, Anchor simplifies complex model behavior into human-understandable rules [20].

2.4 Grad- CAM

Grad-CAM is an explainability technique that visually highlights the regions in an image that are most important for a deep neural network's classification decision. Grad-CAM works by computing the gradients of the model's output with respect to the feature maps in the final convolutional layer, effectively revealing which parts of the image the model 'looks at' when making a prediction.[22]

3. APPLICATIONS OF EXPLAINABLE AI IN HEALTHCARE

Table 1 highlights the applications of AI in healthcare including the ML and XAI methods.

3.1 Image Classification in Radiology

XAI enhances the interpretation of AI-based image classification models, such as those detecting tumors or abnormalities in X-rays and MRIs[23]. Techniques like LIME, SHAP, Grad-CAM and LRP are used to highlight regions in images contributing to the AI model's decision, allowing radiologists to see exactly what the model is focusing on [27].

3.2 Diagnostic Decision Support

XAI can make diagnostic decision support systems more transparent by showing how AI models arrive at specific conclusions. Decision trees or rule-based systems can provide step -by-step explanations for how a model arrived at a specific diagnosis based on patients symptoms, results and medical history[26].

3.3 Transparent AI-assisted Surgery

In AI-assisted surgery, transparency is critical for ensuring safety and precision. XAI techniques like Saliency Maps can be used to visualize the regions of the body that the AI is focusing on during surgery, providing real-time feedback to surgeons[25].

3.4 Clinical Trial Outcome Prediction

XAI helps improve the design and interpretation of clinical trials by providing insights into the factors influencing trial outcomes. SHAP values can be used to rank features based on their importance in predicting trial success, while decision trees or rule-based models offer interpretable results on how specific variables impact trial predictions[28].

3.5 Drug Discovery and Development

XAI algorithms can help in the analysis of large amounts of complex and diverse data, including chemical, biological, and clinical data, to identify potential drug targets, predict drug efficacy and toxicity, and optimize drug designs[29]. Decision trees are used for drug discovery and Rule based systems are used to determine the appropriate method for drug delivery[30].

3.6 Resource Allocation and Healthcare Planning

Explainable AI (XAI) empowers healthcare administrators and policymakers to optimize resource allocation by evaluating medical data to pinpoint high-risk populations, anticipate disease outbreaks, and streamline healthcare services. Transparent AI systems generate actionable insights, enabling evidence-based decision-making and strengthening strategic healthcare planning efforts[31].

Table 1: Summary of various XAI methods in healthcare

Disease	ML method	XAI method	Year	Accuracy	Observations	Ref
Alzheimer's disease	XGBoost, LightGBM, and GB	SHAP	2025	Accuracy of 98%	Efficient output with an accuracy of more than 94% using minimal features for the detection procedure	8
Parkinson's Disease	XRFILR model	SHAP	2025	Accuracy of 96.46%	Effective in harmonizing predictive accuracy, feature selection and interpretability	10
Breast Cancer	KNN, SVM, XG-Boost, RF and ANN	SHAP	2025	Accuracy of 97.9-98.6 %	For WBC dataset, KNN and for WDBC dataset, ANN obtains the best performance	7
Brain cancer	DenseNet121, DenseNet169, DenseNet201, ResNet50, ResNet101, ResNet152, MobileNetV3, Xception, and InceptionV3	GradCAM, GradCAM++, ScoreCAM and LayerCAM	2025	DenseNet169 achieved the highest accuracy of 0.9983	Enhanced the visualization of important characteristics in microscopic images, helping to differentiate between malignant and benign tissue	9
Gastrointestinal disease	DenseNet-121, Greedy soup	SHAP and Grad-CAM	2025	Precision of 85.2 macro average	Notable improvement in diagnostic accuracy, as evidenced by a 2.4% points increase in Macro Precision over existing baselines	11
Glaucoma	KNN, SVM, RF	SHAP	2025	Accuracy of 92%	Outperformed clinicians' performance in glaucoma staging and overall glaucoma diagnosis across multiple trials of clinicians	12
Heart arrhythmia	CNN, LSTM	SHAP	2024	Mean diagnostic accuracy of 98.24 %	Accurate detection of cardiac arrhythmias	13
Lung Cancer	CNN and XGBoost	SHAP	2024	Accuracy of 97.43 %	Minimize the risk of error or bias in the results	14
Epilepsy	Bagged tree-based classifier (BTBC), DT, RF, XGB, LR, NB	SHAP	2024	Mean Accuracy 99.50 %	Patient-independent multimodel data using the proposed framework	15
Pneumonia identification	VGG16	Grad-CAM	2022	Highest accuracy of VGG16 reaches 95.6 %	Traditional deep learning methods for pneumonia identification take less account of the influence of the lung X-ray image background on the model's testing effect	16

4. CHALLENGES OF EXPLAINABLE AI IN HEALTHCARE

Transparent and interpretable AI models are vital for safeguarding patient safety, adhering to regulatory standards, and supporting clinical decision-making [6][32]. While XAI holds immense potential in healthcare by fostering trust in AI systems, it also introduces several challenges that must be addressed, as outlined below:

4.1 Model Complexity

State-of-the-art AI algorithms often rely on intricate architectures with millions of parameters, making their decision logic difficult to decipher [33]. Interpreting these models demands advanced techniques to extract meaningful insights from their complex data representations[19].

4.2 Accuracy vs. Interpretability Trade-off

Simpler models, while easier to interpret, often lag in predictive performance compared to highly accurate but opaque counterparts like deep learning systems [34]. Striking a balance between clarity and precision remains a critical hurdle in XAI development.

4.3 Black-Box Nature of Advanced Models

Deep learning (DL) models, in particular, operate as "black boxes," with internal mechanisms that are not inherently transparent [4]. Specialized post-hoc interpretation methods are required to approximate their decision pathways without accessing core model logic.

4.4 High-Dimensional Data Complexity

AI applications in healthcare often process datasets with vast numbers of features, exacerbating the "curse of dimensionality." This complexity obscures feature relevance and complicates model interpretation [35].

4.5 Context-Dependent Interpretations

Explanations deemed useful in one clinical scenario may lose relevance in another. Ensuring contextually appropriate, actionable insights across diverse healthcare settings remains a persistent challenge for XAI systems.

4.6 User-Centric Explanation Design

Effective XAI requires tailoring outputs to both medical professionals and non-experts [37]. Developing intuitive interfaces that translate technical model behaviors into practical guidance is essential for real-world adoption.

4.7 Standardized Evaluation Metrics

The lack of universal benchmarks for assessing explanation quality—such as fidelity, completeness, and usability—hinders objective comparisons between XAI methodologies.

4.8 Ethical and Societal Risks

Poorly designed explanations risk amplifying biases, misinformation, or discriminatory practices [24]. Proactive governance frameworks are needed to ensure ethical XAI deployment, prioritizing fairness, accountability, and transparency.

5. CONCLUSION

The incorporation of Explainable Artificial Intelligence (XAI) within the healthcare domain signifies a critical progression in harmonizing technological advancements with clinical responsibility. As artificial intelligence systems progressively assist in diagnostics, tailoring treatment strategies, and optimizing resource allocation, the imperative for transparency becomes indispensable within high-stakes medical settings.

This review highlights the transformative significance of XAI methodologies—such as LIME, SHAP, and Grad-CAM—in elucidating AI decision-making processes, thereby enabling healthcare professionals to corroborate recommendations with their clinical acumen and promoting patient trust through interpretable insights.

By enhancing radiological workflows, facilitating surgical accuracy, and expediting pharmaceutical discovery, XAI effectively reconciles the complexity of algorithms with pragmatic clinical reasoning.

The paper also highlights the challenges of implementing XAI in healthcare and emphasizes the importance of model explainability in this sector.

It underscores the need for a human-centered approach in designing and developing XAI methods, facilitating the interpretation of AI models by both patients and medical professionals. While XAI has made significant inroads in healthcare, particularly in diagnosis and surgery, the integration of comprehensive explainability tools with ML and DL methods remains a challenge, especially in critical areas like surgery.

Explainable artificial intelligence (XAI) possesses the potential to transform healthcare delivery, enhance patient outcomes, and elevate both quality and equity by addressing these challenges and capitalizing on advancement opportunities.

6. REFERENCES

- [1] Karim Lekadir Et all., “FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare”, RESEARCH METHODS AND REPORTING, BMJ 2025;388:e081554 | doi: 10.1136/bmj-2024-081554
- [2] D. Kaur, S. Uslu, K. J. Rittichier, and A. Durresi, “Trustworthy Artificial Intelligence: A Review,” ACM Comput. Surv., vol. 55, no. 2, Jan. 2022, doi: 10.1145/3491209.
- [3] Adewale Abayomi Adeniran, Amaka Peace Onebunneand Paul William, “Explainable AI (XAI) in healthcare: Enhancing trust and transparency in critical decision-making”, Journal of Advanced Research and Reviews, vol. 23, no. 3, September. 2024, doi: 10.30574/wjarr,2647–2658.
- [4] V. Hassijaet al., “Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence,” Cognit Comput., vol.16, no. 1, pp. 45–74, Jan. 2024, doi: 10.1007/S12559-023-10179-8/FIGURES/14
- [5] 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: Opportunitiesand Challenges,” IEEE Access, August 2022, doi: 10.1109/ACCESS.2022.3197671
- [6] Ahmad Chaddad, Jihao Peng, Jian Xu and Ahmed Bouridane, “Survey of Explainable AI Techniques in Healthcare,”Sensors, January 2023, 23, 634. <https://doi.org/10.3390/s23020634>
- [7] TAREK KHATER Et all, An Explainable Artificial Intelligence Model for the Classification of Breast Cancer, IEEE Access, Vol. 13, 2025, doi 10.1109/ACCESS.2023.3308446
- [8] Bala Krishnan Raghupathy, Harnessing Explainable Artificial Intelligence(XAI) based SHAPLEY Values and Ensemble Techniques for Accurate Alzheimer's Disease Diagnosis, Engineering, Technology & Applied Science Research, Vol. 15, No. 2, 2025, 20743-20747
- [9] Md. Arafat Alam Khandaker Et all.,From Images to Insights: Transforming Brain Cancer Diagnosis with Explainable AI, ICCIT, arXiv:2501.05426v1 [eess.IV] 9 Jan 2025.
- [10] K. VELU AND N. JAISANKAR, Design of an Early Prediction Model for Parkinson’s Disease Using Machine Learning, IEEE Access, Jan 2025, Vol 13, 17457, doi: 10.1109/ACCESS.2025.3533703
- [11] Muhammad Fahad Et all., Deep insights into gastrointestinal health: A comprehensive analysis of GastroVision dataset using convolutional neural networks and explainable AI, Elsevier- Biomedical Signal Processing and Control, 102 (2025) 107260, <https://doi.org/10.1016/j.bspc.2024.107260>
- [12] Md Mahmudul Hasan, J.P, H.W, OCT-based diagnosis of glaucoma and glaucoma stages using explainable machine learning, Scientific Reports| (2025) 15:3592, <https://doi.org/10.1038/s41598-025-87219-w>
- [13] Alamatsaz N, Tabatabaei L, Yazdchi M, Payan H, Alamatsaz N, Nasimi F. A lightweight hybrid CNN-LSTM explainable model for ECG-based arrhythmia detection. Biomed Signal Process Control 2024;90:105884
- [14] Wani NA, Kumar R, Bedi J., DeepXplainer: an interpretable deep learning based approach for lung cancer detection using explainable artificial intelligence. Comput Method Programs Biomed 2024;243:107879.
- [15] Ahmad I, Yao C, Li L, Chen Y, Liu Z, Ullah I, Chen S. An efficient feature selection and explainable classification method for EEG-based epileptic seizure detection. J Inf Security Applications 2024;80:103654
- [16] Zhang, Z.; Chen, P.; Sapkota, M.; Yang, L. Tandemnet: Distilling knowledge from medical images using diagnostic reports asoptional semantic references. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Quebec City, QC, Canada, 10–14 September 2017; pp. 320–328.
- [17] Sheu RK, Pardeshi MS. A Survey on Medical Explainable AI (XAI): recent Progress, Explainability Approach, Human Interaction and Scoring System. Sensors 2022;22(20):8068.
- [18] Ribeiro, M.T.; Singh, S.; Guestrin, C. “Why should i trust you?” Explaining the predictions of any classifier. In Proceedings of the22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August2016; pp. 1135–1144
- [19] Das, A.; Rad, P. Opportunities and challenges in explainable artificial intelligence (xai): A survey. arXiv2020, arXiv:2006.11371.
- [20] Devam Dave Het Naik, Smriti Singhal Pankesh Patel, “Explainable AI meets Healthcare: A Study on Heart Disease Database”. In: arXiv:2011.03195, Nov. 2020,<https://doi.org/10.48550/arXiv.2011.03195>
- [21] Khedkar S, Gandhi P., “Deep Learning and Explainable AI in Healthcare Using EHR”, In book: Deep Learning Techniques for Biomedical and Health Informatics(pp.129-148), Jan. 2020, doi:10.1007/978-3-030-33966-1_7 https://xai-tutorials.readthedocs.io/en/latest/_model_specific_xai/Grad-CAM.html
- [22] Wenya Linda Bi Et. All, Artificial Intelligence in Cancer Imaging: Clinical Challenges and Applications, CA CANCER J CLIN MARCH/APRIL 2019,VOLUME 69, NUMBER 2, page no:127–157

- [23] L. Weidinger et al., "Ethical and social risks of harm from Language Models," Dec. 2021, Accessed: Aug. 13, 2024.[Online]. Available: <https://arxiv.org/abs/2112.04359v1>
- [24] Katarzyna Borys and Et. All, Explainable AI in medical imaging: An overview for clinical practitioners – Saliency-based XAI approaches, European Journal of Radiology, march 2023,162, 110787, <https://doi.org/10.1016/j.ejrad.2023.110787>
- [25] Al Amin Biswas, A comprehensive review of explainable AI for disease diagnosis, Array- Elsevier, 22 (2024) 100345, <https://doi.org/10.1016/j.array.2024.100345>
- [26] Dost Muhammad, Malika Bendeche, Unveiling the black box: A systematic review of Explainable Artificial Intelligence in medical image analysis, Elsevier- Computational and Structural Biotechnology Journal, 24 (2024) 542–560, <https://doi.org/10.1016/j.csbj.2024.08.005>
- [27] Zahra Sadeghi Et all, A review of Explainable Artificial Intelligence in healthcare, Elsevier- Computers and Electrical Engineering, 118 (2024) 109370, <https://doi.org/10.1016/j.compeleceng.2024.109370>
- [28] Tobias Harren, Hans Matter, Gerhard Hessler, Matthias Rarey, and Christoph Grebner. Interpretation of structure–activity relationships in real-world drug design data sets using explainable artificial intelligence. Journal of Chemical Information and Modeling, 62(3):447– 462, 2022.
- [29] RoohallahAlizadehsania Et all., Explainable Artificial Intelligence for Drug Discovery and Development - A Comprehensive Survey, arXiv:2309.12177v2 [cs.AI] 2 Nov 2023
- [30] M. T. Rashid, N. Wei, and D. Wang, "A survey on social-physical sensing: An emerging sensing paradigm that explores the collective intelligence of humans and machines," Collective Intelligence, vol. 2, no. 2, p. 263391372311708, Apr. 2023, doi: 10.1177/26339137231170825
- [31] N. A. Farooqui, A. K. Mishra, and R. Mehra, "Automatic crop disease recognition by improved abnormality segmentation along with heuristic-based concatenated deep learning model," Intelligent Decision Technologies, vol. 16, no. 2, pp. 407–429, Jan. 2022, doi: 10.3233/IDT-210182.
- [32] Y. Xu et al., "Artificial intelligence: A powerful paradigm for scientific research," The Innovation, vol. 2, no. 4, Nov. 2021, doi: 10.1016/J.XINN.2021.100179.
- [33] T. Wang and Q. Lin, "Hybrid Predictive Models: When an Interpretable Model Collaborates with a Black-box Model," Journal of Machine Learning Research, vol. 22, no. 137, pp. 1–38, 2021, Accessed: Aug. 13, 2024. [Online]. Available: <http://jmlr.org/papers/v22/19-325.html>
- [34] M. A. Bessa et al., "A framework for data-driven analysis of materials under uncertainty: Countering the curse of dimensionality," Comput Methods Appl Mech Eng, vol. 320, pp. 633–667, Jun. 2017, doi: 10.1016/J.CMA.2017.03.037.