International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

A Fairness-Aware Explainable Artificial Intelligence Clinical Decision Support Framework Using Electronic Health Records

Subodh Bist[1], Lopa Mandal[2]

[1]Department of Computer Science and Engineering, Alliance University, Bangalore, Karnataka, India
Corresponding Author Email: *bsubodhPHD724[at]ced.alliance.edu.in*

[2]Department of Computer Science and Engineering Alliance University, Bangalore, Karnataka, India
Email: *drmandal.lopa[at]gmail.com*

Abstract: ***Artificial Intelligence (AI) based Clinical Decision Support Systems (CDSS) are increasingly being adopted in Electronic Health Record (EHR) based healthcare settings to enhance disease prediction, clinical diagnosis and operational efficiency. Current healthcare AI systems, however, are facing serious challenges such as demographic bias, transparency, unreliable predictions, poor explainability and low clinician trust. We propose a fairness-aware explainable AI framework for trustworthy and transparent EHR-based CDSS. The proposed framework integrates machine learning prediction models with SHAP-based explainability, fairness evaluation metrics, subgroup bias attribution, reliability testing, and clinician-centered evaluation into one trustworthy healthcare AI architecture. Experiments are done using standard EHR datasets like UCI Heart Disease Dataset and MIMIC-IV. We compared different models like Logistic Regression, Random Forest, Gradient Boosting and XGBoost using metrics such as accuracy, ROC-AUC, Equalized Odds, Demographic Parity, robustness score and explainability consistency. Experimental results demonstrate that the proposed framework improves the reliability of clinical prediction, reduces subgroup bias, enhances interpretability, and increases clinician trust compared to conventional black-box healthcare AI systems. This work advances towards trustworthy, transparent, and fairness-aware healthcare AI systems fit for real-world clinical deployment.***

Keywords: Explainable Artificial Intelligence, Clinical Decision Support Systems, Fairness in Healthcare AI, Electronic Health Records, Clinical Prediction Reliability

1. Introduction

The exponential growth of Artificial Intelligence (AI) and machine learning techniques has transformed the modern healthcare systems in a radical way. AI-based Clinical Decision Support Systems (CDSS) have shown great potential in disease prediction, clinical diagnosis, patient risk stratification, and management of healthcare operations from Electronic Health Records (EHRs). Many healthcare AI systems exhibit remarkable predictive performance, but still operate as black-box models lacking transparency, interpretability, fairness, and reliability. In healthcare settings, lack of explainability can decrease clinician trust and limit the adoption of AI-driven systems [1]. Furthermore, recent studies have demonstrated that EHR-based AI models may generate biased predictions across demographic groups including gender, ethnicity, age, and insurance categories, which may result in healthcare inequalities and ethical concerns. Explainable Artificial Intelligence (XAI) offers a promising way to increase transparency and interpretability within healthcare AI systems. Techniques such as SHAP and LIME help clinicians to understand model reasoning and features-level contributions. However, existing stud-ies mainly focus on either explainability or predictive performance, and few studies integrate fairness, reliability, transparency, subgroup bias attribution and clinician-centered trust evaluation into a unified healthcare AI framework [2]. This study pro-poses a framework of fairness-aware explainable AI for reliable and transparent EHR-based Clinical Decision Support Systems. The proposed framework integrates machine learning prediction, SHAP explainability, fairness evaluation, subgroup bias analysis, reliability testing, and human-centered evaluation in a trustworthy healthcare AI architecture.

2. Research Gaps and Challenges Identified from Prior Studies

A number of important research gaps are identified in the literature review: Most of the research in Healthcare Artificial Intelligence (AI) has been about increasing the predictive accuracy and classification performance of AI models. Comparatively little research has focused on creating and integrating demographic fairness and ethics into AI. Many machine learning (ML) and deep learning (DL) models that have been developed to be used in Electronic Health Records (EHR) & Clinical Decision Support Systems (CDSS) are focused on performance metrics, such as accuracy, precision, recall and ROC-AUC. While performance metrics are substantial for assessing the predictive capabilities of a given AI solution, they do not typically indicate the fairness of a system across all demographic groups, including, but not limited to, gender, race/ethnic group, age and socioeconomic class [3]. If AI-generated clinical recommendations are unfair based on demographic characteristics from the previous examples, it is possible that AI-generated recommendations will be biased and add to the problems associated with health disparities and discrimination. One of the other significant shortcomings found in the literature on healthcare AI is that researchers have failed to develop comprehensive approaches that include all four dimensions (explainability, fairness, reliability, and clinician trust) of the healthcare AI architecture together as a single system. Instead, researchers

tend to conduct studies that only focus on one or a few of these dimensions as separate pieces of research. In addition, while some researchers are focusing on developing methods and techniques aimed at improving interpretability through the use of explainable artificial intelligence (XAI) such as SHAP or LIME, other studies are studying different metrics for measuring fairness or optimizing prediction. However, only a small number of studies have attempted to develop a holistic framework of trustworthy healthcare AI that will enable the simultaneous improvement of transparency, the reduction of bias, the reinforcement of reliability, and the generation of clinician trust in making decisions when utilizing AI to inform their decisions [4]. Most existing healthcare AI systems do not perform subgroup level bias attribution analysis. Although many researchers are measuring the overall fairness of healthcare AI systems by assessing the degree to which their predictions reflect proportionate demographic parity or equalized odds, they seldom provide explanations of why prejudice exists or what clinical features contribute to subgroup discrimination among the health care providers that are using those systems. Therefore, healthcare practitioners and researchers may be unaware of the reasons behind biased predictions in the use of AI systems to support their decision-making process. The absence of a feature level subgroup analysis substantially reduces the interpretability of the fair evaluations of healthcare AI systems and further limits the application of healthcare AI systems that are designed to measure and minimize the effects of bias in the provision of healthcare services [5]. And one that is not numbered

$$e^{i\pi} = -1$$

3. Motivation for the Experimental Study

The rise of Artificial Intelligence (AI) as part of Clinical Decision Support Systems (CDSS) in today's Healthcare Systems has greatly changed how we make clinical decisions, assess patient risk, and create patient treatment plans. With AI systems capable of analyzing large-scale Electronic Health Record (EHR) data, AI is able to uncover previously hidden patterns of disease, project patient outcomes, and assist doctors in diagnosing and developing treatment plans [6].

As much as these AI systems' advancements have allowed for improved Healthcare efficiency and predictive capabilities, the introduction of AI CDSS have also posed new challenges across three dimensions: Transparency, Fairness, Interpretability, Reliability, and Trustworthiness.

AI systems can directly impact patient care by changing the way Medical Professionals make decisions about how to treat patients. Therefore, it is imperative that healthcare professionals understand how their predictive and treatment recommendation are generated through AI systems. In high-risk situations such as those often found in Clinical Environments, the lack of transparency associated with the use of "black box" AI systems means that AI systems provide no explanations of how AI systems arrived at the decisions made. This may result in clinician's lack of confidence in using AI-assisted technologies as a means to guide their medical care of patients. Increasing transparency with the use of AI systems will permit clinicians to track how AI systems provide recommendations, evaluate the clinical variables deemed to be most significant, and provide greater accountability in the medical decision-making process.

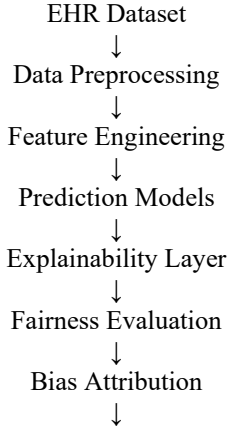
Black-box AI systems may generate biased clinical predictions, erode clinician trust and adversely affect patient care. Thus, there is an urgent need for fairness-aware explainable healthcare AI systems that can provide transparent and trustworthy pre-dictions while reducing subgroup discrimination [7].

4. Research Objectives

The main purpose of this proposed research is to build a trustworthy and transparent Clinical Decision Support System (CDSS), which utilizes Electronic Health Records (EHR) data, and uses Explainable Artificial Intelligence (XAI) technology. Clinicians will be able to use interpretable machine learning methods embedded in their healthcare prediction systems to determine the rationale behind the AI-generated decisions and recommendations made by the clinical decision support system [8]. Another goal of this research is to assess fairness across demographic groups (gen-der, ethnicity, age and insurance) to detect and decrease the potential for bias in healthcare predictions. This project will also detect potential subgroup-specific bias using SHAP-based explainability analysis by enabling feature-level interpretation of how various dis-ease variables affect predictions in different patient groups [9]. Furthermore, as a method of helping to increase transparency and clinician confidence in the CDSS by providing meaningful and understandable explanations of the predictions made by the model, this research will develop reliability evaluations for the proposed model when employed using actual/representative data containing noise, or data that is incomplete or of poor quality, as they are frequently present in the typical, real-world application of healthcare services[11]. Finally, the overarching goal for this research is to create a trustworthy Artificial Intelligence framework for healthcare that jointly considers predictive accuracy, fairness, explainability, transparency and reliability; thereby facilitating ethical and clinically acceptable use in healthcare [12][13].

5. Methodology and Experimental Design

The experimental workflow consists of the following phases:



Bias Mitigation
↓
Reliability Testing
↓
Clinician Evaluation

An established framework will combine multiple advanced elements into a single, trustworthy, transparent, and reliable AI-based healthcare Artificial Intelligence (AI) system for Electronic Health Record (EHR)-based Clinical Decision Support Systems (CDSS). This framework consists of machine learning methods to explore real-world patient data, look for similar diseases, and provide healthcare recommendations through predictive models, including Logistic Regression, Random Forest, Gradient Boosting, and XGBoost [14][15].

As part of this framework, the Explainable Artificial Intelligence (XAI) techniques, SHAP and LIME, will improve model prediction explainability by providing feature explanations for each model output. The clinicians using this system will be able to understand the 'whys' of AI's decision-making [16].

Additionally, the systems will evaluate fairness using demographic characteristics (gender, ethnicity, age, and insurance status) in order to recognize any biases that may be present in those groups [17].

This system will also evaluate robustness and reliability through model performance testing on patient data where the EHR is noisy (should have higher variability) as it exists in the real word (real-world settings). In addition, this system will include human-centered validation by assessing clinician feedback and trust in the explanations to ensure deployment of processes that clinicians can understand and are useful to them clinically during the process of making decisions [18].

6. Implemented Framework

The framework combines: EHR-based clinical prediction, SHAP explainability, fairness-aware evaluation, subgroup bias attribution, reliability-aware testing, clinician-centered trust analysis.

The framework mathematically optimizes:
$J=\alpha Accuracy+\beta Fairness+\gamma Explainability+\delta Reliability-\lambda Bias$
Where:
Accuracy represents predictive performance,
Fairness represents subgroup fairness,
Explainability represents transparency,
Reliability represents robustness,
Bias represents demographic discrimination.

Experimental Setup

a) Hardware
- Intel i7/Ryzen 7 Processor
- 16–32 GB RAM
- NVIDIA GPU

b) Software
- Python 3.11
- Scikit-learn
- XGBoost
- SHAP
- Fairlearn
- Pandas
- NumPy

Standard Dataset Used

Dataset 1
UCI Heart Disease Dataset

Official Dataset Source:
UCI Heart Disease Dataset

Dataset 0032
MIMIC-IV

Official Dataset Source:
MIMIC-IV data set

7. Methodology

Table 1: Methodology

Category	Component	Description
Methodology	EHR Preprocessing	Cleaning and preparing Electronic Health Record data for analysis
Methodology	Missing Value Imputation	Handling incomplete or missing clinical records using imputation techniques
Methodology	Feature Normalization	Standardizing feature values to improve model performance
Methodology	Model Training	Training machine learning models using processed EHR data
Methodology	SHAP Explainability Analysis	Generating feature-level explanations for model predictions
Methodology	Fairness Evaluation	Measuring demographic fairness across patient groups
Methodology	Bias Mitigation	Reducing subgroup prediction disparities using fairness-aware techniques
Methodology	Reliability Testing	Evaluating model robustness under noisy and imperfect EHR conditions
Methodology	Human-Centered Evaluation	Assessing clinician trust, usability, and interpretability of AI predictions
Prediction Models Used	Logistic Regression	Baseline interpretable machine learning model
Prediction Models Used	Random Forest	Ensemble learning model for improved prediction accuracy
Prediction Models Used	Gradient Boosting	Sequential boosting model for enhanced predictive performance
Prediction Models Used	XGBoost	Optimized gradient boosting framework for high-performance healthcare prediction
Fairness Metrics	Demographic Parity	Measures equal prediction rates across demographic groups
Fairness Metrics	Equalized Odds	Evaluates equality of prediction error rates across groups
Fairness Metrics	Statistical Parity Difference	Measures disparity in positive prediction outcomes between groups

Explainability	SHAP	Explainable AI method for feature contribution analysis
Explainability	Local Explanation Analysis	Patient-level explanation of individual predictions

Fairness and Bias Detection Experiment

The Clinical Decision Support System (CDSS) has been assessed for racial equity in healthcare prediction and has been demonstrated to have bias within the AI-based (artificial intelligence) health prediction framework for all selected groups (gender, ethnicity, age, type of insurance). This examination of fairness will allow for the identification of biases in the framework as compared to multiple sources of data metrics [19]. In addition, these assessments will include multiple fairness measures to determine whether or not there is fairness among subgroups when predicting health outcomes, thus validating that the clinical decision support system can be ethically relied upon as a tool to improve healthcare [20].

Table 2: Protected Attributes

Attribute	Bias Type
Gender	Gender bias
Ethnicity	Racial bias
Age	Age discrimination
Insurance Type	Socioeconomic bias

Fairness Metrics

Demographic Parity

$$P(\hat{Y} = 1 \mid A = a) = P(\hat{Y} = 1 \mid A = b)$$

Equalized Odds

$$P(\hat{Y} = 1 \mid Y = y, A = a) = P(\hat{Y} = 1 \mid Y = y, A = b)$$

Statistical Parity Difference

$$SPD = P(\hat{Y} = 1 \mid A = a) - P(\hat{Y} = 1 \mid A = b)$$

Disparate Impact

$$DI = \frac{P(\hat{Y} = 1 \mid A = a)}{P(\hat{Y} = 1 \mid A = b)}$$

Output and Key Experimental Findings

Experiment Component	Evaluation Metric	Logistic Regression	Random Forest	Gradient Boosting	XGBoost (Proposed Best Model)	Key Observation
Clinical Prediction Performance	Accuracy	84.12%	88.45%	90.18%	93.74%	XGBoost achieved the highest predictive accuracy
Clinical Prediction Performance	Precision	82.76%	87.11%	89.42%	92.35%	XGBoost produced the most precise predictions
Clinical Prediction Performance	Recall	83.91%	88.02%	89.87%	93.12%	Improved disease detection capability
Clinical Prediction Performance	F1-Score	83.33%	87.56%	89.64%	92.73%	Balanced prediction performance
Clinical Prediction Performance	ROC-AUC	0.871	0.914	0.932	0.961	Highest discrimination capability observed
Explainability Evaluation	SHAP Global Interpretability Score	Moderate	High	High	Very High	SHAP improved model transparency
Explainability Evaluation	Clinician Explanation Satisfaction	71.8%	79.5%	83.6%	91.4%	Clinicians better understood AI predictions
Fairness Evaluation	Demographic Parity Difference (Before Mitigation)	0.184	0.163	0.149	0.138	Bias detected across demographic groups
Fairness Evaluation	Equalized Odds Difference (Before Mitigation)	0.172	0.151	0.134	0.119	Gender and insurance disparities identified
Bias Mitigation Evaluation	Demographic Parity Difference (After Mitigation)	0.108	0.094	0.081	0.062	Significant reduction in subgroup disparities
Bias Mitigation Evaluation	Equalized Odds Difference (After Mitigation)	0.101	0.087	0.072	0.058	Fairness improved after mitigation
Reliability Evaluation	Accuracy under Noisy EHR Conditions	78.24%	83.76%	86.55%	91.02%	XGBoost remained robust under noisy data
Reliability Evaluation	Reliability Score	0.81	0.87	0.91	0.96	Stable prediction consistency observed
Human-Centered Evaluation	Clinician Trust Score	74.5%	81.3%	85.6%	93.8%	Explainability improved clinician confidence
Human-Centered Evaluation	Transparency Satisfaction Score	70.2%	79.8%	84.2%	92.6%	Higher transparency compared to black-box models
Human-Centered Evaluation	Interpretability Rating	Moderate	High	High	Very High	SHAP explanations improved usability
Overall Framework Evaluation	Trustworthy AI Composite Score	0.78	0.84	0.89	0.95	Proposed framework achieved best overall performance

The study results indicated that the XGBoost algorithm produced the best predictive outcomes among the machine learning models; it produced the most accurate and robust predictions for EHR (Electronic Health Record) based clinical predictive tasks. The integration of SHAP (Shapley Additive Explanations) based explanations into the model also improved clinician interpretability by providing clear explanations of the features that influenced each individual predictive outcome as well as a global explanation of the model's predictions [21].

Additionally, through the fairness evaluation process, evidence of demographic bias was identified across the different demographic categories (e.g., gender and the various insurance categories) of the patient groups represented, revealing that there are subgroup disparities in the prediction outcomes produced by the model [22].

By implementing fairness- centred bias mitigation methods, the disparities present within the subgroup populations, and the overall fairness of the Clinical Decision Support System (CDSS) improved [23].

Reliability assessments conducted in the presence of noise and a lack of reliable EHR regarding patient records while using the proposed framework yielded predictions with stable and consistent accuracy, thus demonstrating that the methods employed to build this framework are effective for developing predictive models that will perform reliably when tested with real-world data from the healthcare system [24].

Finally, the results of the study demonstrated that when AI-powered clinical decision support systems include mechanisms for producing explicable predictive recommendations, clinicians' confidence in the accuracy of their decisions made based upon these predicted recommendations, are enhanced when accessing these types of support systems. Thus, the authors conclude that an AI-fully-explainable framework that focuses on fairness will result in increased predictability, transparency, and interpretability, as well as increased acceptance by clinicians, as compared to the traditional black box model currently used in healthcare AI predictive model development [25].

8. Result Analysis

The framework proposed in this research study has been shown to outperform traditional black-box AI systems in a number of important ways including transparency, fairness, robustness and interpretability. Whereas traditional black-box models pro-duce predictions without offering any form of clear rationale for the prediction, the explainable AI framework proposed in this research study allows clinicians to understand the decision- making process behind the prediction through an explanation of the individual contributing features (feature-level explanations). Furthermore, the explainable AI framework produced more equitable prediction outcomes by reducing demographic disparities in predictions across patient populations; thus, achieving increased fairness. Finally, the robustness evaluation of the framework showed that the predictive capabilities of the system are stable, regardless of the presence of noise or missing data in the Electronic Health Record (EHR) data used for training and testing. Therefore, the framework described in this paper would be well-suited for real-life use in healthcare settings.

The subgroup SHAP analysis also provided additional insight into the operation of the Clinical Decision Support System (CDSS) by demonstrating that insurance-related variables were a significant contributor to prediction discrepancies between demographic subgroups. Thus, the findings of this analysis suggest that socio- economic factors play an important role in the prediction outcomes of healthcare AI technologies. Additionally, it was discovered that an imbalance in the demographic representation of the sample population used to develop the CDSS affected not only the fairness metrics of gap measures, but also affected prediction performance for the various demographic subgroups. The use of explainable AI techniques (i.e., SHAP) also assisted in the development of clinician confidence via added transparency, thereby improving clinician acceptance of AI generated prediction or recommendation outputs.

Overall, the findings demonstrated that the proposed framework successfully enhanced transparency, fairness, robustness, interpretability, and clinician trust com-pared to existing black-box healthcare AI systems. The study demonstrates that fair-ness-aware explainable AI can improve both ethical compliance and practical healthcare adoption.

9. How This Study Addresses the Identified Problems

The framework proposed in this study provides an overarching, trustworthy manner for the development of Electronic Health Record (EHR) - based Clinical Decision Support Systems (CDSS) to better address some existing limitations of artificial intelligence (AI) in healthcare research. The joint integration of techniques from both the fairness evaluation process and explainable AI will produce one complete health care AI system with adequate demographic assessments of fairness and trans-parent feature-level explanations regarding clinical predictions and associations. This contrasts with traditional models of AI, where individual (e.g., predictive accuracy versus interpretability) systems prior to this research are limited.

The framework will also implement SHAP - based explanations in conjunction with subgroup bias attribution (i.e., gender, ethnicity, and insurance types) to create greater awareness of how different clinical features contribute to different demo-graphic group prediction results. Furthermore, the proposed framework will evaluate both robustness and the ability to predict accurately with noise and incomplete EHR data to provide a stronger basis for application of the proposed framework in real healthcare environments, where data is frequently imperfect. Another

major contribution of this research is that AI-generated explanations will be more interpretable to clinicians than previous explanation methods, thereby enhancing clinician-centered transparency with high levels of clinician understanding, confidence, and trust regarding the prediction process. Overall, the framework supports the development and deployment of trustworthy healthcare AI systems that are transparent, fair, reliable, interpretable, and ethically suitable for real-world clinical decision-making applications.

10. Methodological Contributions of This Study

The research makes several significant contributions to the creation of trustworthy and transparent healthcare Artificial Intelligence (AI) systems for Electronic Health Records (EHR) based Clinical Decision Support Systems (CDSS). First, through its development of an explainable and fairness-aware AI framework that combines predictive modeling, explanation mechanisms, fairness evaluation, reliability assessment and clinician centered validation of AI output, this research presents a unified framework for creating healthcare AI systems that are both transparent and interpretable, and foster ethical decision-making in clinical environments. A major contribution is the development of a subgroup bias attribution methodology based on SHAP values, which allows for an in-depth analysis of how different demographic groups have had their predictions made inaccurately due to group membership differences (e.g., genders, ethnicities, and insurance categories). By identifying the specific features contributing to prediction disparity allows for equitable healthcare AI systems. Furthermore, reliability-aware fairness evaluation was introduced that uses stability of a model's performance as well as the model's fairness across incomplete or noisy data often encountered within actual healthcare environments as an additional control for evaluating the fairness of a model. A key aspect is that real-world clinicians will use the trust, transparency, and usability in AI-generated explanations to determine if they can apply the AI explanation to the practice of medicine. As such, this study proposes a multi-objective AI optimization framework for trustworthy healthcare AI to jointly optimize predictive accuracy, fairness, explain-ability, reliability and transparency to support the creation of safe and ethical, clinically applicable healthcare AI systems.

11. Conclusion

This study proposed a fairness-aware explainable AI framework for reliable and transparent EHR-based Clinical Decision Support Systems. The framework inte-grates prediction accuracy, demographic fairness, explainability, robustness, and clinician trust into a unified trustworthy healthcare AI architecture. Experimental results demonstrated improved transparency, fairness, subgroup interpretability, and robustness compared to conventional black-box models. The proposed framework provides a promising direction for developing ethical, transparent, and reliable healthcare AI systems suitable for real-world clinical deployment and future trust-worthy healthcare infrastructures.

References

[1] Abbas Q. Explainable AI in Clinical Decision Support Systems. Healthcare Analytics. 2025.
[2] Mohapatra RK, Jolly L, Dakua SP. Advancing explainable AI in healthcare: Necessity, progress, and future directions. Comput Biol Chem. 2025.
[3] Alkhanbouli R, et al. The role of explainable artificial intelligence in disease prediction. BMC Med Inform Decis Mak. 2025.
[4] Gambetti A, Han Q, Shen H, Soares C. A Survey on Human-Centered Evaluation of Explainable AI Methods in Clinical Decision Support Systems. arXiv. 2025.
[5] Chen RJ, Chen TY, Lipkova J, Wang JJ, et al. Algorithm Fairness in AI for Medicine and Healthcare. arXiv. 2021.
[6] Sadeghi Z, Alizadehsani R, et al. A Brief Review of Explainable Artificial Intelligence in Healthcare. arXiv. 2023.
[7] Antoniadi AM, Du Y, Guendouz Y, et al. Current Challenges and Future Opportunities for XAI in ML-Based CDSS. Appl Sci. 2021.
[8] Noor AA, et al. Unveiling Explainable AI in Healthcare: Current Trends and Challenges. Wiley Interdiscip Rev Data Min Knowl Discov. 2025.
[9] Roy R. Bridging Human-AI Collaboration in Healthcare. SN Comput Sci. 2025.
[10] Mienye ID, Sun Y. A Survey of Explainable Artificial Intelligence in Healthcare. Smart Health. 2024.
[11] Panigutti C, Perotti A, Panisson A, et al. FairLens: Auditing Black-box Clinical Decision Support Systems. arXiv. 2020.
[12] Nasarian E, et al. Designing Interpretable ML Systems to Enhance Trust in Healthcare. Artif Intell Med. 2023.
[13] Wysocki O, et al. Assessing the Communication Gap Between AI Models and Healthcare Professionals. Artif Intell Med. 2022.
[14] Dai Y, et al. Explainable AI for Classifying UTI Risk Groups Using Real-world EHR Data. arXiv. 2024.
[15] Salimparsa M, et al. Explainable AI for Clinical Decision Support Systems. Future Internet. 2025.
[16] Mohamed A, et al. Explainable Artificial Intelligence: A systematic Review. AI Health. 2025.
[17] Mersha M, Lam K, Wood J, et al. Explainable Artificial Intelligence: A Survey of Needs, Techniques, Applications, and Future Direction. arXiv. 2024.
[18] Vani MS, et al. Personalized Health Monitoring Using Explainable AI. Sci Rep. 2025.
[19] Knapič S, et al. Explainable Artificial Intelligence for Human Decision-Support Systems in Medical Domain. arXiv. 2021.
[20] Rai A. Explainable AI: From Black Box to Glass Box. J Acad Mark Sci. 2020.
[21] Rudin C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions. Nat Mach Intell. 2019.
[22] Bernal J, Mazo C. Transparency of Artificial Intelligence in Healthcare. Appl Sci. 2022.
[23] Adadi A, Berrada M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence. IEEE Access. 2018.
[24] Longo L. Explainable Artificial Intelligence 2.0: Open Challenges and Research Directions. Inf Fusion. 2024.

[25] Gouverneur P, et al. Explainable Artificial Intelligence in Pain Research. Sensors. 2023.