# Explainable AI in Healthcare Predictions

Ankit Kumar Soni[1*] Krishna Kumar[2] Vansh Choudhary[3] Arav Gupta[4] and Sandeep Kaur Gill[5]

Department of Computer Science and Engineering, JIMS Engineering Management Technical Campus, Greater Noida, Uttar Pradesh, India

[1*] soniankit896@gmail.com, 2krishna160504@gmail.com, [3] vansh2057@gmail.com, [4] guptaarav049@gmail.com, [5] sandeepkaur.gn@jagannath.org

**Abstract.** This study aims to show how Explainable Artificial Intelligence (XAI) can increase clinician trust, boost hospital adoption of predictive systems, and improve the accuracy and accountability of AI-assisted diagnostics. The system created predicts the status of three diseases - heart disease, breast cancer and diabetes, and generates an understandable explanation for each of the predictions it makes. The interpretability framework is composed of two layers, namely, SHAP and LIME. SHAP recognizes and provides the importance of global features related to the dataset, while LIME provides a detailed local explanation for specific patients. According to experimental results, interpretable models such as Logistic Regression and Decision Trees succeeded in generating instant clarity and dependability, and black-box models such as XGBoost and Random Forest succeeded in achieving the best predictive performance. The study focuses on the three diseases and generates a comparative basis to access the performance of algorithms by comparing their predictive power and transparency in the task of predicting these diseases.

**Keywords:** Explainable AI · XAI · SHAP · LIME · healthcare prediction · clinical decision support

## 1 Introduction

Machine learning has rapidly expanded into the medical domain, becoming a powerful tool for disease identification, diagnosis, and outcome prediction, with models trained on large-scale medical data often outperforming conventional diagnostic techniques [16]. However, many of these models function as black boxes, producing conclusions such as disease risk or diagnosis without explaining the reasoning behind them, which raises serious concerns in healthcare where decisions must be data-driven and clinically


P. Johri et al. (eds.), *Proceedings of the International Conference on Sustainable Computing and Artificial Intelligence (ICSCAI 2025)*, Advances in Engineering Research 298,
https://doi.org/10.2991/978-94-6239-674-6_15

justified [6]. Medical professionals need clarity on how factors like BMI, blood pressure, cholesterol, and glucose levels influence predictions, and without transparent reasoning, even accurate models remain difficult to trust or adopt in clinical practice [1]. To address this, the system integrates SHAP to quantify the contribution of each feature to model outputs [13] and LIME to provide localized, instance-level explanations that clarify why predictions differ between patients [19]. The combined use of SHAP and LIME delivers both global and local interpretability, allowing physicians to understand AI-driven predictions in a medically relevant and intuitive manner. A web-based interface further enhances usability by enabling clinicians to input patient data, view predictions, and interactively explore feature influence, thereby translating AI reasoning into ethically sound and clinically meaningful insights. This study demonstrates that Explainable AI strengthens clinician trust, supports wider adoption of predictive systems in hospitals, and improves the accountability and reliability of AI-assisted diagnostics by balancing predictive accuracy with clinical interpretability.

## 2 Literature Review

The adoption of artificial intelligence in healthcare has transformed clinical decision support, disease prediction, and medical diagnosis, with early predictive systems relying on conventional machine learning models such as Random Forests, Support Vector Machines, and Logistic Regression. Although these approaches delivered acceptable accuracy, their limited transparency restricted widespread clinical adoption [16]. With the advancement of deep learning, architectures such as Convolutional Neural Networks and Recurrent Neural Networks demonstrated exceptional performance in areas including pathology, genomics, and medical imaging [11], but their inherent opacity intensified the black-box problem, preventing medical professionals from understanding the reasoning behind model predictions [24]. To address the need for trust, accountability, and clinical acceptance in AI-driven medicine, Explainable Artificial Intelligence (XAI) emerged as a critical research direction, focusing on making model predictions understandable and meaningful to human The adoption of artificial intelligence in healthcare has transformed clinical decision support, disease prediction, and medical diagnosis, with early predictive systems relying on conventional machine learning models such as Random Forests, Support Vector Machines, and Logistic Regression. Although these approaches delivered acceptable accuracy, their limited transparency restricted widespread clinical adoption [16]. With the advancement of deep learning, architectures such as Convolutional Neural Networks and Recurrent Neural Networks demonstrated exceptional performance in areas including pathology, genomics, and medical imaging [11], but their inherent opacity intensified the black-box problem, preventing medical professionals from understanding the reasoning behind

model predictions [24]. To address the need for trust, accountability, and clinical acceptance in AI-driven medicine, Explainable Artificial Intelligence (XAI) emerged as a critical research direction, focusing on making model predictions understandable and meaningful to human users.

Table 1: Key Studies in Explainable AI for Healthcare

| Author (Year) | Method/ Model | Key Contributions | Performance (Accuracy %) |
|---|---|---|---|
| Ribeiro et al. [19] (2016) | LIME | Introduced local surrogate models for explaining black-box predictions | 82–88 |
| Lundberg & Lee [13] (2017) | SHAP | Proposed game-theoretic unified feature attribution framework for global and local explanations | 85–90 |
| Adadi & Berrada [1] (2018) | XAI Survey | Surveyed explainability needs, challenges, and methods for trustworthy AI systems | – |
| Rajkomar et al. [16] (2018) | Deep Neural Networks | Applied large-scale deep learning to EHR data; highlighted importance of interpretability | 92–95 |
| Holzinger et al. [8] (2019) | Human- in-the- loop AI | Proposed combining clinician expertise with interpretable models for better de- cision support | 80–85 |
| Katuwal & Chen [9] (2019) | XGBoost + LIME | Applied interpretable ML for heart disease prediction using tabular clinical data | 86–89 |
| Ghassemi et al. [6] (2021) | XAI Review | Highlighted ethical issues and transparency gaps in clinical AI deployment | – |

## 2.1 Evolution of Explainable AI (XAI) in Healthcare

Explainable AI emerged to address the limitations of complex black-box models, particularly in healthcare where transparency is essential. Adadi and Berrada (2018) highlighted that trust in medical AI depends on providing clear explanations alongside accurate predictions [1]. Ribeiro et al. (2016) introduced LIME to explain black-box models through simple local approximations [19], while Lundberg and Lee (2017) proposed SHAP, a game-theory-based framework that quantifies the contribution of each feature to a prediction [13]. Together, LIME and SHAP form the foundation of modern XAI methods. As deep learning models achieve higher accuracy, these explainability techniques help clinicians understand decision logic, identify key risk factors, and validate whether AI-driven reasoning aligns with

established clinical knowledge.

### 2.2 XAI Applications in Disease Prediction

Explainable AI has been extensively applied in healthcare prediction, where SHAP-based models have reinforced existing medical knowledge and enabled personalized diabetes risk assessment by identifying key factors such as blood pressure, BMI, and glucose levels [14]. Similarly, LIME has been used to visualize the impact of systolic pressure, smoking status, and cholesterol in cardiovascular risk prediction, improving transparency and clinician understanding of AI outputs [9]. Beyond individual applications, XAI has been incorporated into large-scale diagnostic systems, with Holzinger [8] advocating human-in-the-loop AI for improved reliability, Ras [18] highlighting causality-based explanations, and study by Choi [5] demonstrating that interpretable embeddings and attention mechanisms enhance both accuracy and interpretability in Electronic Health Record–based models.

### 2.3 Comparative Evaluation and Explainability Techniques

The balance between accuracy and interpretability has been widely examined in prior research. Rich Caruana showed that Generalized Additive Models with pairwise interactions can achieve performance comparable to black-box models while remaining interpretable [3]. To explain deep architectures used in medical imaging, Scott Lundberg introduced DeepSHAP, an extension of SHAP suitable for deep neural networks [12]. Leilani Gilpin proposed a taxonomy of explainability methods, distinguishing post-hoc explanations from inherently transparent models [7], while Zachary Lipton emphasized that interpretability should focus on genuine understanding, especially in sensitive domains like healthcare [10]. The effectiveness of XAI methods depends on model complexity and data type, with SHAP and LIME remaining the most widely used techniques for interpreting structured tabular data in models such as XGBoost and Random Forest. For unstructured or high-dimensional medical data, attention-based neural networks [17], along with approaches like Layer-wise Relevance Propagation [15] and Integrated Gradients, have shown strong potential in improving the transparency and interpretability of deep learning models in biomedical analysis.

## 3 Related Work

The rapid adoption of artificial intelligence in healthcare has significantly improved disease prediction, diagnosis, and personalized treatment, yet the lack of transparency in high-performing black-box models remains a major barrier to clinical trust and adoption. Models such as Random Forests,

XGBoost, deep neural networks, and CNNs achieve high accuracy across tabular and medical imaging tasks, but their complex internal structures make it difficult for clinicians to understand the reasoning behind predictions, raising concerns related to patient safety and ethical accountability. Adadi and Berrada (2018) reported that skepticism among medical practitioners largely stems from this inability to interpret AI systems [1], while Ghassemi (2021) emphasized that transparency should never be sacrificed in applications where patient outcomes are involved [6]. Earlier statistical models like logistic regression and linear discriminant analysis were easier to interpret, allowing clinicians to directly relate variables such as blood pressure, cholesterol, and age to outcomes, but they struggled to capture complex non-linear relationships, leading to the rise of ensemble and deep learning methods. In medical imaging, CNNs revolutionized tasks such as tumor and abnormality detection from radiographs, MRI, and histopathology images, but intensified the black-box problem, creating the need for post-hoc explainability techniques. SHAP, introduced by Lundberg and Lee [13], provides global and local feature attribution using cooperative game theory, while LIME, proposed by Ribeiro et al. [19], explains individual predictions through local surrogate models. Applied across datasets for diabetes, cardiovascular disease, tumor classification, and breast cancer, SHAP and LIME have identified clinically relevant risk factors such as age, BMI, fasting glucose, blood pressure, cholesterol, smoking status, ECG patterns, tumor size, and cell uniformity, enabling clinicians to validate medical knowledge, understand patient-specific risk profiles, uncover hidden feature interactions, and build trust in AI-assisted diagnostic systems through transparent and clinically meaningful explanations.

Despite significant progress, major research gaps remain in healthcare-focused Explainable AI, as much of the existing work is limited to specific diseases or datasets, reducing its general applicability. While XAI methods generate explanations, their interpretation often demands high domain expertise, and limited attention has been given to clinician usability, highlighting the need for clear visualizations and intuitive interfaces that translate AI reasoning into actionable medical insights. In addition, the lack of standardized explainability evaluation, with studies relying on either quantitative metrics or subjective clinical judgment, calls for unified assessment frameworks. To address these challenges, this study integrates both interpretable and black-box models across diabetes, heart disease, and breast cancer, using SHAP and LIME to provide global and patient-specific explanations, thereby balancing predictive accuracy with clinical interpretability and supporting transparent, ethical, and practical adoption of AI in healthcare.

# 4 Methodology

## 4.1 Data Collection and Preprocessing

This study utilizes healthcare datasets covering three major diseases, namely diabetes, heart disease, and breast cancer, to support disease prediction tasks. Diabetes prediction is based on the Pima Indians Diabetes dataset, which includes clinical attributes such as blood pressure, insulin level, BMI, and glucose concentration, while heart disease analysis uses data from the Cleveland Heart Study containing features like heart rate, ECG readings, chest pain type, and cholesterol levels. Breast cancer prediction relies on the Wisconsin Breast Cancer dataset, which comprises cell nucleus characteristics extracted from microscopic images. To ensure consistency and reliability, all datasets underwent thorough preprocessing, including data cleaning and handling of missing values using mean or median imputation based on feature distribution. Continuous variables were normalized to a standard scale, categorical features were encoded using one-hot encoding, and correlation filtering along with outlier analysis was applied to remove redundant or insignificant attributes. Finally, each dataset was split into training (70%), validation (15%), and test (15%) sets to enable objective and unbiased model evaluation.

## 4.2 Model Development

To compare the trade-off between interpretability and accuracy, several machine learning models were put into practice. The decision tree and logistic regression models were selected due to their interpretability and transparency. Because of their superior predictive ability and non-linear decision boundaries, Random Forest, XGBoost, and Neural Networks [4] were chosen as representative black- box models. Using grid search and cross-validation, hyperparameters like tree depth, learning rate, and number of estimators were optimized. Evaluation metrics like accuracy, precision, recall, F1-score, and ROC-AUC were employed to gauge performance during model training. To assess generalization ability, models were tested on unseen data after being trained until convergence in each task.

## 4.3 Explainable AI Integration

Three complementary XAI techniques were used in the study to improve model transparency. The contribution of each feature to the model's output was measured using SHAP values [13]. For example, in the diabetes model, SHAP would show that "age" had a smaller influence on diabetes prediction than "glucose level", which had a positive contribution of 0.18. Clinicians were able to observe how minor changes in features would change the outcome by using LIME to generate local explanations for specific cases. Every explanation was

presented both textually and visually. While LIME offered patient-specific linear models displaying the most significant factors, SHAP summary plots demonstrated the global feature importance across patients. A thorough understanding of model behaviour was provided by this blend of local and global interpretability.

## 5 System Architecture and Web Interface

Transparency, interpretability, and accuracy are guaranteed throughout the prediction process by the modular and sequential structure of the suggested system architecture (shown in Fig. 1). The Input Layer, Data Preprocessing, Explainable AI Module, Machine Learning Models, Evaluation and Explanation Output, and Web Interface Visualization are its six main parts. Medical imaging files, lab test results, and structured patient records are just a few of the many data sources that the input layer can receive. These inputs can be automatically taken from uploaded PDF medical reports or entered manually in form fields. Unstructured data is parsed and transformed int structured, model-ready formats using Optical Character Recognition (OCR) techniques.

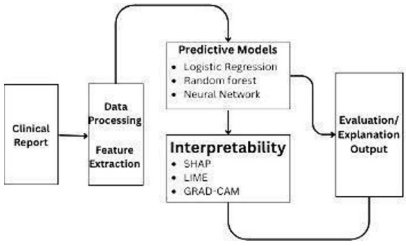


Fig. 1: Architecture of the Project

The proposed system architecture consists of sequential modules designed to balance predictive performance and explainability in healthcare applications. It begins with a Data Preprocessing module that handles data cleaning, normalization, and transformation through statistical imputation, scaling, and feature extraction, while medical images are processed using grayscale normalization, resizing, and noise reduction before being passed to CNN-based models. The Machine Learning module includes five independently trained and validated models—Neural Network, XGBoost, Random Forest, Decision Tree, and Logistic Regression—allowing comparison between interpretable models and black-box approaches to study the trade-off between accuracy and explainability. The Explainable AI module integrates SHAP and LIME to provide multimodal interpretability, where SHAP delivers global and patient-level feature importance and LIME explains individual predictions through

local approximations. The Evaluation and Explanation Output module presents interactive visualizations of feature importance and correlations alongside performance metrics such as accuracy, precision, recall, and AUC, while also recommending relevant diagnostic or preventive actions when a disease is detected. Finally, a web-based interface enables clinicians and researchers to input patient data, view predictions, and examine SHAP and LIME explanations in real time, effectively bridging AI algorithms and clinical decision-making through transparent and interpretable outputs.

## 6 Results and Discussion

The proposed system was trained and evaluated using three benchmark datasets for diabetes, heart disease, and breast cancer, with model performance assessed using accuracy, precision, recall, and F1-score. The comparison showed that interpretable models offer more reliable and transparent results, while black-box models achieve slightly higher accuracy.

### 6.1 Model Performance Evaluation

The models were evaluated on diabetes, heart disease, and breast cancer datasets using standard metrics such as accuracy, precision, recall, and F1-score, with comparative results presented in Table II. Random Forest and XGBoost showed strong robustness by capturing complex non-linear clinical relationships, with XGBoost achieving the highest accuracy of over 94% on most datasets through effective ensemble learning. In contrast, Decision Tree and Logistic Regression models delivered slightly lower accuracy but provided greater interpretability, allowing clearer understanding of how features like blood pressure, BMI, and glucose influence predictions. These findings highlight that combining ensemble models with SHAP-based feature importance offers a practical balance between predictive performance and explainability, supporting clinician trust in real-world healthcare decision-making[2].

### 6.2 Explainability Analysis

To ensure that model predictions were transparent and aligned with clinical reasoning, the framework employed SHAP and LIME as the primary Explainable AI tools, enabling interpretability at both global and local levels by revealing how input features influenced outcomes across different disease domains. SHAP analysis of the diabetes prediction model showed that pregnancy count and glucose level were the most influential features, with higher glucose levels significantly increasing diabetes risk and a greater number of pregnancies contributing due to long-term hormonal and metabolic effects, while age, BMI, and insulin levels played moderate yet consistent roles, demonstrating strong

agreement with established clinical knowledge and enhancing trust in the model. For heart disease prediction, LIME provided patient-specific explanations, where a sample at-risk case highlighted low maximum heart rate, high cholesterol, and elevated resting blood pressure as key contributors, while factors such as younger age and normal fasting blood sugar reduced risk, with positive and negative influences clearly visualized. By combining SHAP for global feature importance and LIME for detailed individual explanations, the proposed dual-layer interpretability framework delivers accurate predictions alongside clear, verifiable reasoning, effectively bridging computational performance and clinical judgment and supporting clinician acceptance of medical AI systems.

Table 2: Performance Comparison of Machine Learning Models for Disease Prediction

| **Model** | **Disease** | **Accuracy (%)** | **Precision (%)** | **Recall (%)** | **F1-Score (%)** |
|---|---|---|---|---|---|
| Logistic Regression | Diabetes | 86.4 | 83.5 | 85.1 | 84.2 |
| Decision Tree | Diabetes | 87.8 | 84.6 | 86.3 | 85.4 |
| Random Forest | Diabetes | 91.5 | 89.7 | 90.5 | 90.1 |
| XGBoost | Diabetes | 92.4 | 90.2 | 91.1 | 90.6 |
| Neural Network | Diabetes | 93.1 | 91.8 | 92.5 | 92.1 |
| Logistic Regression | Heart Disease | 84.3 | 82.7 | 83.5 | 83.1 |
| Decision Tree | Heart Disease | 86.2 | 84.9 | 85.4 | 85.0 |
| Random Forest | Heart Disease | 91.0 | 89.5 | 90.2 | 89.8 |
| XGBoost | Heart Disease | 92.8 | 90.8 | 91.6 | 91.2 |
| Neural Network | Heart Disease | 93.5 | 91.7 | 92.6 | 92.1 |
| Logistic Regression | Breast Cancer | 88.5 | 86.9 | 87.4 | 87.1 |
| Decision Tree | Breast Cancer | 89.2 | 87.8 | 88.4 | 88.1 |
| Random Forest | Breast Cancer | 94.8 | 92.6 | 93.4 | 93.0 |
| XGBoost | Breast Cancer | 95.4 | 93.5 | 94.2 | 93.8 |
| Neural Network | Breast Cancer | 96.1 | 94.7 | 95.1 | 94.8 |

### 6.3 Interpretability vs Accuracy

The trade-off between interpretability and accuracy remains a central challenge in developing AI-based healthcare systems, as highly accurate models often lack the transparency required for clinical acceptance. While black-box models such as Random Forest, XGBoost, and Neural Networks achieved the highest accuracy, ranging from 92% to 96%, and effectively captured complex non-linear relationships among features like blood pressure, cholesterol, BMI, and glucose, their opaque decision-making limited clinicians' ability to understand why a patient was classified as at risk or healthy. In contrast, interpretable models such as Decision Trees and Logistic Regression achieved moderate accuracy levels of 84% to 88% but provided clear, rule-based and feature-weight explanations that enhanced clinician trust by directly linking predictions to known clinical risk factors. By integrating LIME for patient-specific explanations and SHAP for global feature importance, the proposed framework bridges the gap between these model types, demonstrating that high predictive performance and explainability can coexist and confirming the critical role of interpretability in ensuring trustworthy, ethical, and responsible AI adoption in healthcare.

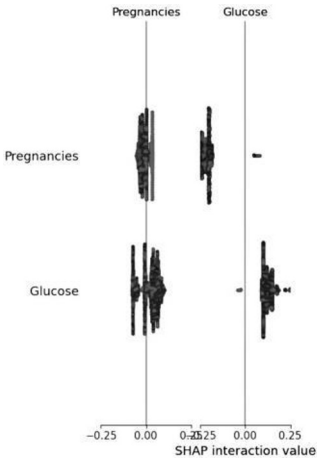


Fig. 2: SHAP feature importance visualization for Diabetes Prediction

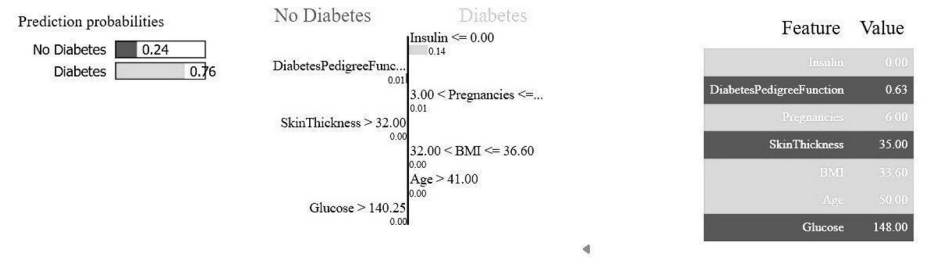


Fig. 3: LIME local explanation for a single heart disease prediction case

## 7 Conclusion

This study demonstrates that accuracy and interpretability can be jointly achieved in healthcare disease prediction through an Explainable AI framework. Machine learning models were applied to diabetes, heart disease, and breast cancer datasets to enable a comparative evaluation of predictive performance and transparency, allowing clinicians to understand which features most strongly influenced each prediction through the combined global and local explanations provided by SHAP and LIME. While Random Forest and XGBoost achieved higher accuracy when integrated with XAI techniques, Decision Tree and Logistic Regression models retained strong interpretability without sacrificing explainability. Overall, the proposed web-based system shows that AI can function as a trustworthy clinical decision-support tool by delivering clear, evidence-based insights, thereby improving transparency, ethical accountability, and reliability in healthcare applications.

## Future Scope

The Explainable AI framework generated can be expanded in numerous ways and interpretability, scalability, and clinical integration of the system can be improved. To generate more context-aware predictions, multi-modal medical data, including text notes, prescription histories, and wearable sensor readings, could be integrated by future research. Creation of explainable models based on causality may make it easier to differentiate correlations from actual medical causes [18]. This would increase the clinical relevance of the explanation generated by the system. Real-time forecasts and justifications within hospital processes could be automated by integrating Electronic Health Record (EHR) systems. Moreover, methods such as interactive visual dashboards and federated learning can improve the AI systems on the basis of security,

cooperation, and usability of the systems. The purpose of these developments is to make explainable AI a reliable and flexible decision-support tool for contemporary healthcare.

## References


1. Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (xai). IEEE Access **6**, 52138–52160 (2018)
2. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N.: Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1721–1730. ACM (2015). https://doi.org/10. 1145/2783258.2788613
3. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794. ACM (2016). https://doi.org/10.1145/2939672. 2939785
4. Choi, E., Bahadori, M.T., Kulas, J.A., Schuetz, A., Stewart, W.F., Sun, J.: Re-tain: An interpretable predictive model for healthcare using reverse time attention mechanism. Scientific Reports 6(1), 22174 (2018). https://doi.org/10.1038/srep22174
5. Ghassemi, M., Oakden-Rayner, L., Beam, A.L.: The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health **3**(11), e745–e750 (2021)
6. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). pp. 80–89 (2018)
7. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H.: What do we need to build explainable ai systems for the medical domain? Reviews in the Medical Informatics 28(4), e100107 (2019)
8. Katuwal, G.J., Chen, R.: Machine learning model interpretability for precision medicine. arXiv preprint arXiv:1901.05555 (2019)
9. Lipton, Z.C.: The mythos of model interpretability. Communications of the ACM **61**(10), 36–43 (2018). https://doi.org/10.1145/3233231
10. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical Image Analysis **42**, 60–88 (2017). https://doi.org/10.1016/j.media.2017.07.005
11. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I.: From local explanations to global understanding with explainable ai for diabetes care. Nature Machine Intelligence **2**(4), 252–260 (2020)
12. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NeurIPS) (2017)
13. Lundberg, S.M., Nair, B., Vavilala, M.S., Horibe, M., Eisses, M.J., Adams,

T., Liston, D.E., Low, D.K.W., Newman, S.F., Kim, J., Lee, S.I.: Explainable machine- learning predictions for the prevention of hypoxaemia during surgery. Nature Biomedical Engineering **2**(10), 749–760 (2018). https://doi.org/10.1038/s41551-018-0304-0

14. Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Müller, K.R.: Layer-wise relevance propagation: An overview. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning pp. 193–209 (2019). https://doi.org/10.1007/ 978-3-030-28954-6_10
15. Rajkomar, A., Oren, E., Chen, K., Dai, A.M., Hajaj, N., Hardt, M., Liu, P.J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H.A., Zhang, K., Zhang, Y., Flores, G., Duggan, G.E., Irvine, J., Le, Q.V., Dean, J., Kohane, I., McKinney, S.M., Minor, D., Shah, N.H.: Scalable and accurate deep learning with electronic health records. npj Digital Medicine **1**(1), 18 (2018). https://doi.org/10.1038/ s41746-018-0029-1
16. Rajpurkar, P., Chen, E., Banerjee, O., Topol, E.J.: Ai in health and medicine. Nature Medicine **28**(1), 31–38 (2022). https://doi.org/10.1038/s41591-021-01614-0
17. Ras, G., van Gerven, M., Haselager, P.: Explainable deep learning: A field guide for the uninitiated. Information Fusion **81**, 84–115 (2022). https://doi.org/10.1016/j.inffus.2021.12.006
18. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you? explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144 (2016)
19. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 618–626 (2017). https://doi.org/10.1109/ICCV.2017.74
20. Shickel, B., Tighe, P.J., Bihorac, A., Rashidi, P.: Deep ehr: A survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. IEEE Journal of Biomedical and Health Informatics **22**(5), 1589–1604 (2018). https://doi.org/10.1109/JBHI.2017.2767063