Health, Empathy, and AI Learning (HEAL)
Vol. 1, No. 2, May, 2026, pp. 186−197
E-ISSN: 3163-267X | P-ISSN: 3163-2661

Interpretable Deep Learning Models for Transparent Health Decision Support Systems

Deni Sunaryo[1], Antonius Ary Setyawan[2], Po Abas Sunarya[3], Sondang Visiana Sihotang[4], Putri Anahera Oganda[5*]
[1]Faculty of Economics and Business, Serang Raya University, Indonesia
[2]Department of Information Systems, Sekolah Tinggi Ilmu Komputer Yos Sudarso, Indonesia
[3]APTISI, Indonesia
[4]Casindo Group, Indonesia
[5]Mfinitee Incorporation, South Africa
[1]denisunaryomm@gmail.com, [2]arysetpr@stikomyos.ac.id, [3]abas@raharja.info, [4]sondang@raharja.info [5]ahera@mfinitee.co.za
***Corresponding Author**

Article Info

Article history:

Submission February 22, 2026
Revised March 26, 2026
Accepted April 17, 2026
Published May 31, 2026

Keywords:

Healthcare AI
Interpretable Deep Learning
Clinical Decision Support
Explainable AI
Patient Safety

ABSTRACT

The increasing adoption of deep learning in healthcare has created significant opportunities for improving medical diagnosis, patient risk prediction, and clinical decision support. However, many deep learning models remain difficult to interpret, limiting their acceptance in health systems where transparency, accountability, and patient safety are essential. **This study aims** to develop and evaluate interpretable deep learning models for transparent health decision support systems. **The proposed approach** integrates convolutional neural networks, deep neural networks, attention mechanisms, SHAP values, and Grad-CAM to explain model predictions across medical imaging and clinical tabular datasets. Public healthcare datasets, including chest X-ray images, electronic health records, and diabetes patient records, are used to represent real clinical decision-making contexts. Model performance is evaluated using accuracy, precision, recall, F1-score, AUC, sensitivity, and specificity, while interpretability is assessed through explanation clarity, feature relevance, decision traceability, and clinical usefulness. The **findings** indicate that interpretable deep learning models can provide reliable predictive performance while offering clearer explanations for medical decisions. Visual heatmaps and feature attribution outputs help identify relevant disease indicators, patient risk factors, and clinical variables that influence model predictions. **This study contributes** to the development of trustworthy AI in healthcare by supporting transparent, ethical, and human-centered clinical decision-making. The research is also aligned with SDG 3 by promoting better health outcomes, patient safety, and responsible digital health innovation.

1. INTRODUCTION

Artificial Intelligence (AI) has become an important technology in the transformation of modern healthcare systems [1]. The use of AI enables healthcare institutions to process complex medical data, support clinical diagnosis, predict patient risks, and improve the quality of medical decision-making [2]. Among various AI approaches, deep learning has gained significant attention because of its ability to analyze high-

dimensional health data, including medical images, electronic health records, laboratory results, and patient clinical histories. Deep learning models have been widely applied in healthcare tasks such as disease detection, chest X-ray classification, hospital readmission prediction, intensive care unit risk assessment, and clinical decision support. These capabilities show that deep learning has strong potential to improve the efficiency, accuracy, and responsiveness of healthcare services.

However, the application of deep learning in healthcare also raises serious concerns related to transparency, accountability, and patient safety [3]. Many deep learning models operate as black-box systems, where the internal reasoning behind predictions is difficult to understand [4]. In medical practice, this limitation can create significant risks because doctors and healthcare professionals need to understand why a model produces a certain diagnosis or risk prediction [5, 6]. A prediction result without explanation may reduce clinical trust and make it difficult for medical professionals to validate whether the model focuses on relevant health indicators [7]. Therefore, interpretability is not only a technical requirement but also a clinical necessity in AI-based healthcare systems.

The lack of interpretability in healthcare AI can affect the quality and safety of medical decisions [8]. If an AI model provides an incorrect prediction without a clear explanation, it may lead to inappropriate diagnosis, delayed treatment, or unsafe clinical recommendations [9, 10]. In addition, healthcare datasets may contain bias, missing values, and imbalance between patient groups, which can influence model performance and fairness [11]. Without transparent explanations, these problems may remain hidden and reduce the reliability of AI systems in real clinical environments. For this reason, explainable and interpretable deep learning models are needed to ensure that AI-assisted decisions can be reviewed, trusted, and responsibly used by healthcare professionals [12].

Interpretable deep learning provides several approaches to improve transparency in healthcare decision support [13]. For medical image analysis, methods such as Grad-CAM can generate visual heatmaps that highlight important regions in images, helping clinicians verify whether the model focuses on medically relevant areas [14]. For tabular clinical data, SHAP values can identify the contribution of patient features such as age, laboratory results, medication history, comorbidities, and previous hospital admissions [15]. Attention-based models can also help explain temporal patterns in electronic health records by showing which clinical events influence prediction results [16]. These methods allow healthcare professionals to better understand model behavior and strengthen human oversight in clinical decision-making.

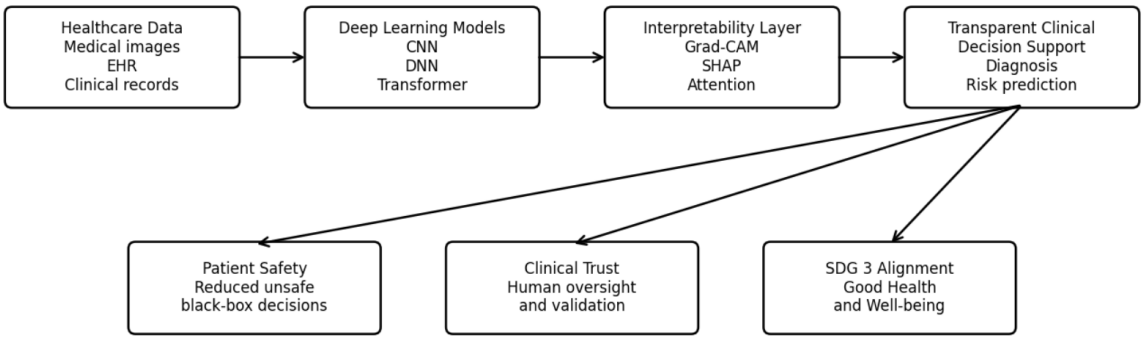

Figure 1. Interpretable Deep Learning Framework for Transparent Health Decision Support

Figure 1 illustrates the proposed flow of interpretable deep learning in healthcare decision support systems. The process begins with healthcare data, including medical images, electronic health records, and clinical records. These data are processed using deep learning models such as CNN, DNN, and Transformer. To improve transparency, interpretability methods such as Grad-CAM, SHAP, and attention mechanisms are applied. The explanation outputs support transparent clinical decision-making, including diagnosis and patient risk prediction. This framework emphasizes that healthcare AI should support patient safety, clinical trust, and SDG 3 on Good Health and Well-being [17].

This study focuses on the development and evaluation of interpretable deep learning models for transparent health decision support systems. The main objective is to examine how interpretability techniques can improve the transparency of deep learning models without significantly reducing predictive performance [4].

This research evaluates the relationship between model accuracy, explanation quality, decision traceability, and clinical usefulness. By integrating interpretability into healthcare AI, this study aims to support safer, more accountable, and more human-centered medical decision-making.

This research is also aligned with Sustainable Development Goal 3, which emphasizes good health and well-being [18]. Transparent AI systems can support early disease detection, patient risk prediction, and safer clinical decision support. In addition, the study contributes to responsible digital health innovation by ensuring that AI systems are not only accurate but also explainable, trustworthy, and clinically meaningful [19]. Therefore, interpretable deep learning is positioned as an important foundation for building ethical, reliable, and patient-centered healthcare AI systems.

2. LITERATURE REVIEW

The development of deep learning in healthcare has significantly transformed the way medical data are analyzed, interpreted, and applied in clinical decision-making [1, 20]. Deep learning models have been widely used for various health-related tasks, including disease detection, medical image classification, patient risk prediction, electronic health record analysis, and clinical decision support systems. These models are capable of identifying complex patterns from high-dimensional health data such as chest X-ray images, laboratory records, patient demographics, and clinical histories [21, 22]. However, although deep learning can provide high predictive accuracy, its black-box nature remains a major concern in healthcare because medical decisions directly affect patient safety, diagnosis accuracy, and treatment quality [23]. In clinical settings, a prediction result is not sufficient unless healthcare professionals can understand the reasoning behind the model's decision [24, 25]. Therefore, interpretability has become an essential requirement for the responsible implementation of artificial intelligence in healthcare [26].

Explainable Artificial Intelligence (XAI) has emerged as an important research area to address the transparency limitations of deep learning models in medical applications [27]. In healthcare, explainability allows clinicians to examine whether AI predictions are based on medically relevant indicators rather than irrelevant or biased patterns [28]. For example, in medical imaging, visual explanation methods such as Grad-CAM can highlight specific image regions that influence the model's prediction, helping radiologists verify whether the model focuses on abnormal lung areas, lesions, or other clinically meaningful signs [29]. Meanwhile, in tabular clinical data, SHAP values can be used to identify the contribution of patient-related features such as age, previous admissions, laboratory results, medication history, and comorbidities [30]. These explanation techniques support clinical validation, reduce uncertainty, and improve trust between healthcare professionals and AI-based decision support systems.

Existing studies classify interpretability methods into intrinsic and post-hoc approaches. Intrinsic interpretability refers to models that are designed to be understandable from the beginning, such as attention-based architectures that show which clinical variables or time-series patterns influence predictions [31]. This approach is useful in healthcare data because patient conditions often change over time and require temporal explanation. Post-hoc interpretability, on the other hand, explains the behavior of trained black-box models without changing their internal architecture. Methods such as Grad-CAM, SHAP, saliency maps, and surrogate models are commonly used to explain deep learning outputs after prediction. In healthcare, these techniques are valuable because they help doctors assess whether AI-generated predictions are clinically logical, ethically acceptable, and safe to use in patient care [32].

Despite the benefits of interpretable deep learning, several challenges remain in its application to healthcare. First, explanation results may not always be stable across similar patient cases, which can reduce clinical confidence [33]. Second, some explanations are technically understandable for data scientists but difficult for doctors, nurses, or healthcare administrators to interpret [34]. Third, healthcare datasets often contain sensitive patient information, missing values, class imbalance, and potential bias that may influence prediction outcomes [35]. If these issues are not addressed, AI models may produce inaccurate or unfair recommendations for certain patient groups. Therefore, interpretability in healthcare should not only focus on technical explanation but also on clinical usefulness, fairness, privacy, and human-centered decision-making.

From an ethical and sustainability perspective, interpretable AI in healthcare is related to patient safety, accountability, and responsible digital health innovation [36]. Transparent AI systems allow healthcare professionals to validate predictions, detect errors, and maintain human oversight in decision-making. This is important because clinical decisions require responsibility and cannot be fully delegated to automated systems. In

addition, interpretable deep learning supports Sustainable Development Goal 3, which focuses on good health and well-being, by improving diagnosis support, early risk detection, and safer healthcare services. Therefore, integrating interpretability into deep learning models is not only a technical improvement but also a strategic requirement for building trustworthy, ethical, and patient-centered healthcare AI systems.

Based on [37, 38], there is still a research gap in how interpretable deep learning models can balance predictive performance and explanation quality in healthcare decision support systems. Many prior studies focus mainly on model accuracy, while fewer studies examine whether the generated explanations are clinically meaningful, understandable, and useful for healthcare stakeholders. In addition, limited attention has been given to comparing different explanation techniques across medical image data, electronic health records, and tabular clinical datasets. This study addresses these gaps by evaluating interpretable deep learning models in the context of healthcare decision support, with emphasis on transparency, patient safety, clinical trust, and responsible AI adoption.

3. METHODOLOGY

3.1. Research Approach

This study adopts a mixed-methods research approach to evaluate interpretable deep learning models for transparent health decision support systems. Quantitative analysis is used to measure model performance in healthcare prediction tasks, while qualitative analysis is used to assess the clarity and usefulness of explanation outputs. This approach is suitable for healthcare AI because clinical decision support systems require not only accurate predictions but also understandable explanations that can support patient safety, clinical trust, and responsible medical decision-making.

The research focuses on comparing conventional black-box deep learning models with interpretable deep learning models. The comparison is conducted to determine whether interpretability techniques can improve transparency without significantly reducing predictive performance. In the healthcare context, this evaluation is important because doctors and healthcare professionals need to understand the reasons behind AI-generated predictions before using them in diagnosis, treatment planning, or patient risk assessment.

3.2. Methodology Workflow

The research process begins with the identification of healthcare problems, particularly medical diagnosis and patient risk prediction. After that, suitable healthcare datasets are selected, including medical image datasets, electronic health records, and tabular clinical records. The collected data are processed through cleaning, normalization, feature preparation, and train-test splitting. Deep learning models such as CNN, DNN, and Transformer are then developed as baseline models. To improve transparency, interpretability techniques such as Grad-CAM, SHAP values, and attention visualization are integrated into the model evaluation process.

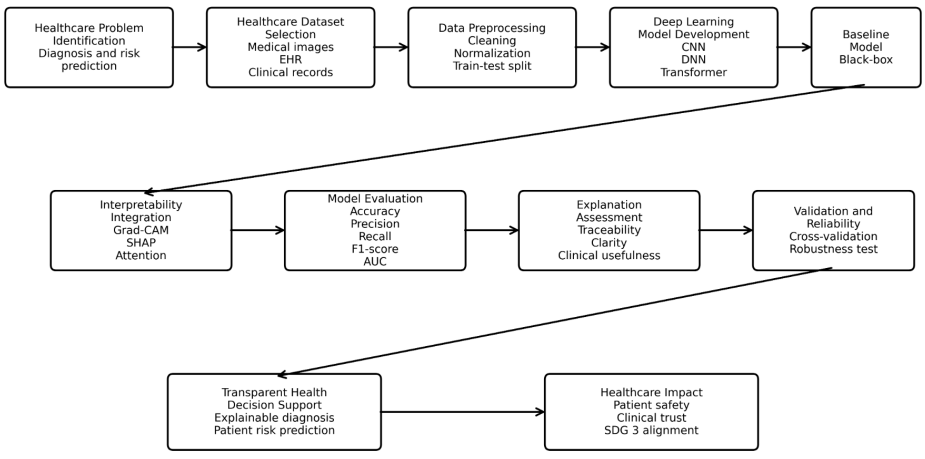

Figure 2. Methodology Workflow for Interpretable Deep Learning in Healthcare

Figure 2 presents the methodological workflow used to develop and evaluate interpretable deep learning models for healthcare decision support. The process begins with the identification of healthcare problems,

such as disease diagnosis and patient risk prediction, followed by the selection of relevant healthcare datasets, including medical images, electronic health records, and clinical records. After data preprocessing, deep learning models such as CNN, DNN, and Transformer are developed as baseline models. Interpretability techniques, including Grad-CAM, SHAP, and attention mechanisms, are then integrated to explain model predictions and improve transparency. The model is evaluated using predictive performance metrics and explanation quality indicators to ensure that the system is not only accurate but also clinically understandable. Finally, validation and reliability testing are conducted to support transparent health decision-making, patient safety, clinical trust, and alignment with SDG 3 on Good Health and Well-being.

3.3. Data Collection Methods

This study uses publicly available healthcare datasets that represent clinical decision-making environments. The datasets are selected based on their relevance to health diagnosis, patient risk prediction, and clinical decision support. Three types of healthcare data are considered: medical images, electronic health records, and tabular clinical data. These datasets are suitable for evaluating interpretable deep learning because healthcare decisions require high predictive accuracy and clear explanations that can be understood by medical professionals.

Table 1. Healthcare Dataset and Use Case Overview

Health Dataset	Data Type	Task Type	Health Relevance	Justification
Chest X-ray Dataset	Medical Image	Disease Classification	Very High	Used to detect thoracic diseases and support radiology-based diagnosis
Electronic Health Record Dataset	Clinical Time-Series / EHR	Patient Risk Prediction	Very High	Used to predict ICU risk, mortality risk, or patient deterioration
Diabetes Clinical Dataset	Tabular Clinical Data	Readmission Prediction	High	Used to predict hospital readmission risk among diabetic patients

The information in Table 1 shows that each dataset represents a different healthcare decision-making context. Medical image data are relevant for disease detection, electronic health records are useful for monitoring patient conditions over time, and tabular clinical data are important for identifying patient risk factors. These datasets provide a strong foundation for evaluating whether interpretable deep learning models can support transparent and reliable healthcare decisions.

3.4. Analytical Techniques

The analytical process begins with the development of baseline deep learning models. Convolutional Neural Networks are used for medical image classification, Deep Neural Networks are used for tabular clinical prediction, and Transformer-based models are applied to sequential or electronic health record data. These models represent common deep learning architectures used in healthcare AI.

After developing the baseline models, interpretability methods are applied to explain the prediction results. Grad-CAM is used to generate visual heatmaps for medical image classification, allowing clinicians to identify which image regions influence the model prediction. SHAP values are used to explain tabular clinical predictions by showing the contribution of each patient feature. Attention visualization is used to interpret temporal clinical patterns in electronic health records.

Table 2. Model Architecture and Interpretability Techniques

Health Task	Data Modality	Baseline Model	Interpretability Method	Explanation Output
Chest disease detection	Medical Image	CNN	Grad-CAM	Heatmap showing disease-related image regions
ICU risk prediction	EHR / Time-Series	Transformer / LSTM	Attention Visualization	Important clinical events and temporal patterns
Diabetes readmission prediction	Tabular Clinical Data	DNN	SHAP Values	Feature contribution scores for patient risk factors

Table 2 explains the relationship between healthcare tasks, data modality, baseline deep learning models, interpretability methods, and the explanation outputs produced by each model. For chest disease detection,

medical image data are processed using CNN because this architecture is effective for image classification tasks, while Grad-CAM is applied to generate heatmaps that show important image regions related to disease prediction. For ICU risk prediction, EHR or time-series data are analyzed using Transformer or LSTM models because both are suitable for capturing temporal patterns in patient conditions, and attention visualization is used to identify important clinical events that influence the prediction. Meanwhile, diabetes readmission prediction uses tabular clinical data with a DNN model, supported by SHAP values to explain the contribution of each patient risk factor. Overall, this table shows that each interpretability technique is selected based on the type of healthcare data and model architecture, so the explanation results can be more clinically relevant, transparent, and useful for health decision support.

3.5. Evaluation Metrics

Model evaluation is conducted using both predictive performance metrics and interpretability metrics. Predictive performance is assessed using accuracy, precision, recall, F1-score, AUC, sensitivity, and specificity. These metrics are important in healthcare because false predictions may affect patient safety and clinical decision quality. Interpretability is assessed through explanation clarity, decision traceability, explanation consistency, and clinical usefulness.

Table 3. Healthcare Performance and Explainability Evaluation Metrics

Evaluation Aspect	Metric	Description
Predictive Performance	Accuracy	Measures the overall correctness of model predictions in healthcare classification or prediction tasks
Predictive Performance	Precision	Measures the reliability of positive predictions generated by the healthcare AI model
Predictive Performance	Recall / Sensitivity	Measures the model ability to detect relevant patient cases, such as disease-positive or high-risk patients
Predictive Performance	Specificity	Measures the model ability to correctly identify non-risk or non-disease patient cases
Predictive Performance	AUC	Measures the model discrimination ability in distinguishing between different patient outcome classes
Interpretability	Explanation Clarity	Measures how understandable the explanation output is for healthcare professionals
Interpretability	Decision Traceability	Measures whether the model decision process can be traced and reviewed by clinical stakeholders
Interpretability	Clinical Usefulness	Measures whether the explanation output is useful for supporting clinical decision-making and patient safety

Table 3 presents the evaluation metrics used to assess the performance and explainability of the proposed healthcare AI model. The predictive performance metrics, including accuracy, precision, recall or sensitivity, specificity, and AUC, are used to measure how well the model identifies patient conditions, disease risks, or clinical outcomes. These metrics are important in healthcare because incorrect predictions may affect diagnosis accuracy, treatment decisions, and patient safety. In addition, the table also includes interpretability metrics such as explanation clarity, decision traceability, and clinical usefulness. These metrics evaluate whether the model's prediction results can be understood, reviewed, and applied by healthcare professionals.

3.6. Validation and Reliability

Validation is conducted using k-fold cross-validation to ensure that model performance is stable across different training and testing subsets. Robustness testing is also applied to evaluate whether the model produces consistent predictions when exposed to similar or slightly modified patient data. In healthcare AI, this step is important because clinical models must remain reliable across different patient conditions and data variations.

Reliability is strengthened by documenting preprocessing steps, model parameters, training procedures, and evaluation settings. This supports reproducibility and allows other researchers to validate the findings. In addition, explanation outputs are reviewed based on their clinical relevance to ensure that the model does not rely on irrelevant or misleading patterns. Through this validation process, the study aims to develop interpretable deep learning models that are accurate, transparent, and safe for healthcare decision support.

4. RESULT AND DISCUSSION

This section presents the evaluation results of black-box deep learning models and interpretable deep learning models in healthcare decision support. The evaluation focuses not only on predictive performance but also on explanation quality, decision traceability, clinical usefulness, and patient safety. In healthcare applications, model accuracy alone is not sufficient because medical predictions must be understandable and clinically valid before being used by healthcare professionals. Therefore, the analysis compares conventional deep learning models with interpretable models that integrate Grad-CAM, SHAP values, and attention visualization.

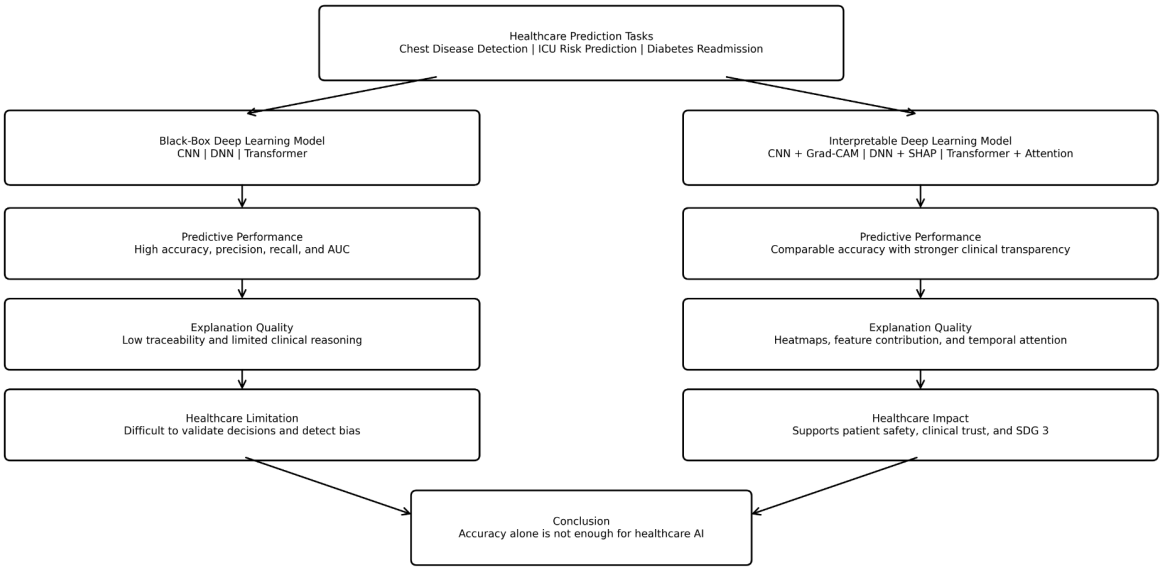

Figure 3. Comparative Results of Black-Box and Interpretable Deep Learning in Healthcare

Figure 3 presents the comparison between black-box deep learning models and interpretable deep learning models in healthcare decision support. The black-box model can produce high predictive performance through accuracy, precision, recall, and AUC, but it has limitations in explanation quality because the decision process is difficult to trace and validate clinically. In contrast, interpretable deep learning models provide comparable predictive performance while offering stronger transparency through Grad-CAM, SHAP, and attention mechanisms. These explanation methods help healthcare professionals understand whether model predictions are based on relevant medical indicators. Therefore, the figure emphasizes that accuracy alone is not enough in healthcare AI, because clinical trust, patient safety, bias detection, and transparent decision-making must also be considered.

In addition, Figure 3 shows that interpretable deep learning models are more suitable for healthcare because they support performance, transparency, and clinical accountability. Explanation methods help clinicians review AI predictions before making decisions. Thus, healthcare AI should support, not replace, professional judgment. Interpretable models can improve clinician confidence, reduce uncertainty, and support safer AI use in diagnosis, risk prediction, and patient care.

4.1. Healthcare Prediction Performance Results

The predictive performance evaluation shows that both black-box and interpretable deep learning models are capable of producing reliable results for healthcare prediction tasks. Black-box models generally achieve slightly higher accuracy because they focus mainly on optimizing prediction outcomes. However, interpretable models show only a marginal decrease in predictive performance while providing clearer explanation outputs. This finding indicates that transparency can be improved without significantly reducing the reliability of healthcare AI models.

Table 4 shows that interpretable deep learning models maintain competitive predictive performance across different healthcare tasks. Although the black-box models obtain slightly higher accuracy and AUC values, the difference is relatively small. In contrast, interpretable models provide higher transparency, making them more suitable for clinical settings where healthcare professionals need to understand the basis of

Table 4. Healthcare Prediction Performance Comparison

Health Task	Model Type	Accuracy	Precision	Recall	F1-score	AUC	Interpretability Level
Chest disease detection	Black-box CNN	89.2%	0.88	0.86	0.87	0.91	Low
Chest disease detection	Interpretable CNN + Grad-CAM	88.6%	0.87	0.86	0.86	0.90	High
ICU risk prediction	Black-box Transformer	87.8%	0.86	0.84	0.85	0.89	Low
ICU risk prediction	Interpretable Transformer + Attention	87.1%	0.85	0.84	0.84	0.88	High
Diabetes readmission prediction	Black-box DNN	84.5%	0.82	0.80	0.81	0.86	Low
Diabetes readmission prediction	Interpretable DNN + SHAP	83.9%	0.81	0.80	0.80	0.85	High

AI-generated predictions. For example, Grad-CAM helps identify important regions in chest X-ray images, attention visualization highlights important clinical events in patient records, and SHAP values explain the influence of patient risk factors in readmission prediction.

4.2. Clinical Robustness and Explanation Quality Analysis

In healthcare AI, robustness is essential because clinical data may vary across patient groups, hospital environments, and medical conditions. The comparison shows that interpretable models provide more stable decision behavior because their prediction process can be examined through explanation outputs. Black-box models may still perform well numerically, but their lack of explanation makes it difficult to determine whether the model prediction is based on clinically relevant features or misleading patterns.

Table 5. Clinical Robustness and Explanation Quality Results

Evaluation Aspect	Black-box Deep Learning	Interpretable Deep Learning
Decision stability under patient data variation	Moderate	High
Explanation availability	Not available	Available
Clinical traceability	Low	High
Bias detection support	Weak	Strong
Human oversight support	Limited	Effective
Patient safety contribution	Moderate	High

The results in Table 5 indicate that interpretable deep learning models provide stronger support for healthcare decision-making. The availability of explanation outputs allows clinicians to examine whether predictions are aligned with medical reasoning. This is important because AI systems in healthcare should not only classify diseases or predict risks but also support safe and accountable clinical judgment. Interpretable models improve decision traceability, support bias detection, and strengthen human oversight, which are essential for reducing unsafe black-box decisions.

4.3. Interpretability and Clinical Usefulness

The interpretability analysis demonstrates that explanation techniques improve the usability of deep learning models in healthcare contexts. Grad-CAM produces visual heatmaps that can help clinicians review whether image-based predictions focus on relevant disease regions. SHAP values provide feature contribution scores that explain how patient variables, such as laboratory results, medication history, comorbidities, or previous hospital admissions, influence prediction outcomes. Attention visualization supports the interpretation of sequential clinical data by showing which patient events contribute most to risk prediction.

Table 6. Healthcare Explanation Quality Comparison

Explanation Dimension	Black-box Models	Interpretable Models
Decision traceability	Absent	Present
Explanation clarity	Poor	High
Clinical usefulness	Limited	Strong
Support for clinician validation	Weak	Effective
Support for patient safety	Limited	Strong

Table 6 shows that interpretable models are more appropriate for healthcare decision support because they provide explanations that can be reviewed by healthcare professionals. In clinical practice, explanation clarity and decision traceability are important because doctors need to validate AI outputs before applying them to patient care. The ability to explain predictions also helps identify potential errors, detect bias, and improve confidence in AI-assisted diagnosis or risk prediction.

4.4. Comparative Discussion

The comparison between black-box and interpretable deep learning models shows that predictive accuracy should not be the only evaluation criterion in healthcare AI. While black-box models may achieve slightly better numerical performance, their limited transparency creates challenges for clinical validation and patient safety. In contrast, interpretable models provide a more balanced approach by combining acceptable predictive performance with clearer explanations and stronger clinical usefulness.

These findings support the argument that healthcare AI must be designed as a human-centered decision support tool rather than a fully autonomous decision-maker. Interpretable deep learning allows clinicians to remain involved in the decision-making process by reviewing model explanations and validating whether predictions are clinically reasonable. This strengthens trust between healthcare professionals and AI systems.

The results also show that interpretability contributes directly to responsible digital health innovation. By providing transparent explanations, interpretable deep learning supports patient safety, clinical accountability, and ethical AI implementation. This aligns with SDG 3 on Good Health and Well-being because the proposed approach contributes to safer diagnosis, early risk detection, and more reliable healthcare decision support.

5. MANAGERIAL IMPLICATIONS

The findings of this study provide practical implications for healthcare managers, hospital administrators, clinical technology leaders, and digital health practitioners. Interpretable deep learning models can improve clinical decision support by ensuring that AI predictions are accurate, transparent, and understandable. Healthcare managers should prioritize AI systems with explanation outputs such as Grad-CAM, SHAP values, and attention visualization to help clinicians review predictions before diagnosis, risk assessment, or treatment planning. In practice, healthcare institutions need clear AI governance policies, model validation procedures, regular audits, and training programs so medical professionals can interpret AI results, detect possible errors, and ensure clinical relevance. By integrating interpretable AI into healthcare workflows, hospitals can strengthen patient safety, ethical accountability, clinician trust, and responsible digital health innovation aligned with SDG 3 on Good Health and Well-being.

6. CONCLUSION

This study demonstrates that interpretable deep learning models can support transparent and reliable health decision support systems. By applying explainability techniques such as Grad-CAM, SHAP values, and attention visualization, the proposed approach enables healthcare stakeholders to understand how AI models generate predictions. The findings show that interpretable models can maintain competitive predictive performance while providing clearer decision explanations, which are essential for clinical trust, patient safety, and responsible healthcare decision-making.

From a healthcare perspective, the study highlights that prediction accuracy alone is not sufficient for AI implementation in clinical environments. Medical decisions require transparency, accountability, and human validation because incorrect or unexplained AI predictions may affect diagnosis quality, treatment planning,

and patient outcomes. Interpretable deep learning helps clinicians examine whether AI recommendations are based on relevant medical indicators, detect possible bias, and maintain professional oversight in AI-assisted healthcare.

Future research may expand this study by testing the proposed model on larger and more diverse healthcare datasets, including multimodal data that combine medical images, laboratory results, electronic health records, and clinical notes. Further studies should also involve doctors, nurses, and healthcare administrators in evaluating explanation quality to ensure that AI outputs are not only technically accurate but also clinically meaningful, ethical, and useful in real-world medical environments.

7. DECLARATIONS

7.1. About Authors

Deni Sunaryo (DS) https://orcid.org/0000-0002-1897-7587

Antonius Ary Setyawan (AA) https://orcid.org/0009-0007-2575-1125

Po Abas Sunarya (AS) https://orcid.org/0000-0002-3869-2837

Sondang Visiana Sihotang (SV) https://orcid.org/0009-0008-4042-5798

Putri Anahera Oganda (PA) -

7.2. Author Contributions

Conceptualization: DS; Methodology: SV; Software: AS; Validation: AA and PA; Formal Analysis: DS and AS; Investigation: SV; Resources: PA; Data Curation: DS; Writing Original Draft Preparation: SV and PA; Writing Review and Editing: PA and SV; Visualization: AA; All authors, DS, AA, AS, SV, and PA, have read and agreed to the published version of the manuscript.

7.3. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

7.4. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

7.5. Declaration of Conflicting Interest

The authors declare that they have no conflicts of interest, known competing financial interests, or personal relationships that could have influenced the work reported in this paper.

REFERENCES

[1] S. A. Alowais, S. S. Alghamdi, N. Alsuhebany, T. Alqahtani, A. I. Alshaya, S. N. Almohareb, A. Aldairem, M. Alrashed, K. Bin Saleh, H. A. Badreldin *et al.*, "Revolutionizing healthcare: the role of artificial intelligence in clinical practice," *BMC medical education*, vol. 23, no. 1, p. 689, 2023.

[2] S. Maleki Varnosfaderani and M. Forouzanfar, "The role of ai in hospitals and clinics: transforming healthcare in the 21st century," *Bioengineering*, vol. 11, no. 4, p. 337, 2024.

[3] P. Goktas and A. Grzybowski, "Shaping the future of healthcare: ethical clinical challenges and pathways to trustworthy ai," *Journal of Clinical Medicine*, vol. 14, no. 5, p. 1605, 2025.

[4] E. ŞAHiN, N. N. Arslan, and D. Özdemir, "Unlocking the black box: an in-depth review on interpretability, explainability, and reliability in deep learning," *Neural computing and applications*, vol. 37, no. 2, pp. 859–965, 2025.

[5] P. Hager, F. Jungmann, R. Holland, K. Bhagat, I. Hubrecht, M. Knauer, J. Vielhauer, M. Makowski, R. Braren, G. Kaissis *et al.*, "Evaluation and mitigation of the limitations of large language models in clinical decision-making," *Nature medicine*, vol. 30, no. 9, pp. 2613–2622, 2024.

[6] J. Alderman, R. Riley, D. Parekh, C. Summers, X. Liu, and A. Denniston, "Hidden risks of predictive models in healthcare," *BMJ evidence-based medicine*, 2025.

[7] O. Wysocki, J. K. Davies, M. Vigo, A. C. Armstrong, D. Landers, R. Lee, and A. Freitas, "Assessing the communication gap between ai models and healthcare professionals: Explainability, utility and trust in ai-driven clinical decision-making," *Artificial Intelligence*, vol. 316, p. 103839, 2023.

[8] M. Ennab and H. Mcheick, "Enhancing interpretability and accuracy of ai models in healthcare: a comprehensive review on challenges and future directions," *Frontiers in Robotics and AI*, vol. 11, p. 1444763, 2024.

[9] R. R. Kothinti, "Artificial intelligence in healthcare: Revolutionizing precision medicine, predictive analytics, and ethical considerations in autonomous diagnostics," *World Journal of Advanced Research and Reviews*, vol. 19, no. 3, pp. 3395–3406, 2024.

[10] H. Xu and K. M. J. Shuttleworth, "Medical artificial intelligence and the black box problem: a view based on the ethical principle of "do no harm"," *Intelligent Medicine*, vol. 4, no. 1, pp. 52–57, 2024.

[11] D. Ueda, T. Kakinuma, S. Fujita, K. Kamagata, Y. Fushimi, R. Ito, Y. Matsui, T. Nozaki, T. Nakaura, N. Fujima *et al.*, "Fairness of artificial intelligence in healthcare: review and recommendations," *Japanese journal of radiology*, vol. 42, no. 1, pp. 3–15, 2024.

[12] P. Pandey, S. Chaudhary, and X. Nie, "Deep learning technique for interpretable diagnosis of polycystic ovary syndrome in ultrasound imaging," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 7, no. 3, pp. 779–792, 2025, https://doi.org/10.34306/att.v7i3.768.

[13] A. A. Adeniran, A. P. Onebunne, and P. William, "Explainable ai (xai) in healthcare: Enhancing trust and transparency in critical decision-making," *World Journal of Advanced Research and Reviews*, vol. 23, no. 3, pp. 2647–2658, 2024.

[14] S. Guluwadi *et al.*, "Enhancing brain tumor detection in mri images through explainable ai using grad-cam with resnet 50," *BMC medical imaging*, vol. 24, no. 1, pp. 1–19, 2024.

[15] H. Luo, C. Xiang, L. Zeng, S. Li, X. Mei, L. Xiong, Y. Liu, C. Wen, Y. Cui, L. Du *et al.*, "Shap based predictive modeling for 1 year all-cause readmission risk in elderly heart failure patients: feature selection and model interpretation," *Scientific reports*, vol. 14, no. 1, p. 17728, 2024.

[16] A. S. Anita, T. Kuusk, G. Nicola, M. Hardini, and U. Rahardja, "Advancements in artificial intelligence and their contributions to sustainable development goals: A multidisciplinary review," *Sundara Advanced Research on Artificial Intelligence*, vol. 2, no. 1, pp. 37–47, 2026.

[17] United Nations Department of Economic and Social Affairs, "The 17 goals," https://sdgs.un.org/goals, 2026, accessed: 2026-03-26.

[18] ——, "Goal 13: Take Urgent Action to Combat Climate Change and Its Impacts," https://sdgs.un.org/goals/goal13, 2025, accessed: 2026-03-26.

[19] M. Nurkanti, U. Rahardja, M. Lubis, and A. A. M. Shukri, "Deep learning-based biology learning with ethnopedagogy and local wisdom to support sustainable development goals," *Jurnal Pendidikan IPA Indonesia*, vol. 15, no. 1, 2026.

[20] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.

[21] T. H. Kim, R. Chinthaginjala, A. Srinivasulu, S. P. Tera, and S. O. Rab, "Covid-19 health data prediction: a critical evaluation of cnn-based approaches," *Scientific Reports*, vol. 15, no. 1, p. 9121, 2025.

[22] F. Liu, H. Zhou, K. Wang, Y. Yu, Y. Gao, Z. Sun, S. Liu, S. Sun, Z. Zou, Z. Li *et al.*, "Metagp: A generative foundation model integrating electronic health records and multimodal imaging for addressing unmet clinical needs," *Cell Reports Medicine*, vol. 6, no. 4, 2025.

[23] N. Azizah, P. A. Sunarya, U. Rahardja, A. B. Mutiara, P. Prihandoko, and C. Pasha, "Improving smear-negative tuberculosis detection using data augmentation and faster r-cnn," *International Journal of Cyber and IT Service Management (IJCITSM)*, vol. 6, no. 1, pp. 65–77, 2026.

[24] C. C. Yang, "Explainable artificial intelligence for predictive modeling in healthcare," *Journal of healthcare informatics research*, vol. 6, no. 2, pp. 228–239, 2022.

[25] N. Hassan, R. Slight, G. Morgan, D. W. Bates, S. Gallier, E. Sapey, and S. Slight, "Road map for clinicians to develop and evaluate ai predictive models to inform clinical decision-making," *BMJ Health & Care Informatics*, vol. 30, no. 1, p. e100784, 2023.

[26] M. Ennab and H. Mcheick, "Designing an interpretability-based model to explain the artificial intelligence algorithms in healthcare," *Diagnostics*, vol. 12, no. 7, p. 1557, 2022.

[27] D. Juliastuti, Y. F. C. P. Meilani, I. N. Hikam, and J. Nathalie, "Integrating clinical intelligence and emotional responsiveness to build patient trust in healthcare delivery," *Health, Empathy, and AI Learning*

(HEAL), vol. 1, no. 1, pp. 17–25, 2025.

[28] H. Wang, G. Shi, X. Liu, J. Zhang, S. Chen, J. Shi, Y.-a. Li, G. Yue, and Y. Wu, "Explainable ai in medicine: A comprehensive narrative review of methods, applications, and future directions," *Expert Systems*, vol. 43, no. 6, p. e70259, 2026.

[29] V. Agarwal, M. Lohani, A. S. Bist, and D. Julianingsih, "Application of voting based approach on deep learning algorithm for lung disease classification," in *2022 International Conference on Science and Technology (ICOSTECH)*. IEEE, 2022, pp. 01–07.

[30] J. H. Jeong, K.-S. Lee, S.-M. Park, S. R. Kim, M.-N. Kim, S. C. Chae, S.-H. Hur, I. W. Seong, S. K. Oh, T. H. Ahn *et al.*, "Prediction of longitudinal clinical outcomes after acute myocardial infarction using a dynamic machine learning algorithm," *Frontiers in Cardiovascular Medicine*, vol. 11, p. 1340022, 2024.

[31] D. Jin, E. Sergeeva, W.-H. Weng, G. Chauhan, and P. Szolovits, "Explainable deep learning in healthcare: A methodological survey from an attribution view," *WIREs Mechanisms of Disease*, vol. 14, no. 3, p. e1548, 2022.

[32] K. Karnawati, D. N. Ramadhan, T. L. Anita, R. Nurmala, and L. Maria, "Designing inclusive companion robots to mitigate bias and enhance empathy in ai-driven care systems," *Journal of Orange Technology*, vol. 2, no. 2, pp. 83–92, 2026.

[33] F. Tang, F. Yao, and M. Zhou, "Innovating diagnostic learning: How generative ai explanation styles influence cognitive load and confidence calibration in medical students," *Computers & Education*, p. 105619, 2026.

[34] J. Seringa, A. Hirata, A. R. Pedro, R. Santana, and T. Magalhães, "Health care professionals and data scientists' perspectives on a machine learning system to anticipate and manage the risk of decompensation from patients with heart failure: qualitative interview study," *Journal of Medical Internet Research*, vol. 27, p. e54990, 2025.

[35] B. Draghi, Z. Wang, P. Myles, and A. Tucker, "Identifying and handling data bias within primary healthcare data using synthetic data generators," *Heliyon*, vol. 10, no. 2, 2024.

[36] C. Trocin, P. Mikalef, Z. Papamitsiou, and K. Conboy, "Responsible ai for digital health: a synthesis and a research agenda," *Information Systems Frontiers*, vol. 25, no. 6, pp. 2139–2157, 2023.

[37] Q. Abbas, W. Jeong, and S. W. Lee, "Explainable ai in clinical decision support systems: a meta-analysis of methods, applications, and usability challenges," in *Healthcare*, vol. 13, no. 17. MDPI, 2025, p. 2154.

[38] M. M. R. Sweet, M. P. Ahmed, S. Akter, and S. A. Tisha, "Integrating deep learning and interpretable regression models for transparent decision support in healthcare diagnostics," *Journal of Medical and Health Studies*, vol. 6, no. 5, pp. 17–38, 2025.