ENHANCING EXPLAINABILITY IN AI-DRIVEN HEALTHCARE DECISION SUPPORT SYSTEMS: A DOMAIN-SPECIFIC AND STAKEHOLDER-CENTRIC APPROACH

Abstract

The use of artificial intelligence (AI) in healthcare decision-support systems has become increasingly popular in recent years. However, as these systems become more prevalent, concerns have arisen regarding the transparency and interpretability of their decision-making processes. In response to these concerns, this study proposes a new approach for enhancing explainability in AI-based healthcare decision support systems. The proposed method is designed to address the diverse needs of different stakeholders, including physicians, nurses, and patients. By focusing on domain-specific interpretability, the method is able to generate explanations that are grounded in established medical knowledge and terminology. This is achieved through the integration of medical ontologies and expert-driven feature extraction.

In addition, the proposed method employs a flexible framework for tailoring explanations according to user preferences. This ensures that stakeholders receive information that is customized, understandable, and relevant to their needs. To address ethical and privacy concerns, the method also utilizes privacy-preserving mechanisms and ethical guideline adherence modules. To evaluate the effectiveness of the proposed method, it was tested on a real-world healthcare dataset containing anonymized patient records related to a specific medical condition. The results demonstrated that the integration of explainability did not significantly affect the performance of the decision support system. Furthermore, domain experts and

Authors

Dr. M. V Viajaya Saradhi
Professor, Dept of CSE
ACE Engineering College
Hyderabad, Telangana, India
Meduri.vsd@gmail.com

M. S. V Aditya
Senior System Engineer
Infosys
Hyderabad, Telangana, India
medurisaivenkataaditya@gmail.com

Dr. M Nagabhushana Rao
Professor- AI&IT Dept
Vidya Jyothi Institute of Technology
Hyderabad, Telangana, India
mnraoit@vjit.ac.in

stakeholders rated the explanations generated by the proposed method as understandable, relevant, and consistent with medical knowledge. The proposed method was also found to outperform existing explainable AI methods in generating domain-specific, tailored explanations while maintaining comparable performance levels.

Overall, this study represents an important step towards the development of more transparent, understandable, and ethically responsible AI systems in healthcare decision support. By fostering greater trust and adoption among healthcare professionals and patients, this approach has the potential to transform the way that healthcare decisions are made. Moving forward, future research can extend the method to other domains and refine the privacy-preserving and ethical guideline adherence mechanisms to ensure robust protection of sensitive patient data and ethical decision-making.

Keywords: AI, Stakeholders, TreeSHAP, Deep learning

I. INTRODUCTION

The impact of Artificial Intelligence (AI) has been growing considerably in various industries, with healthcare being no exception. The potential for AI in healthcare is substantial, as it could significantly alter patient care and medical decision-making, thereby improving outcomes and optimizing resource utilization. One of the key applications of AI in healthcare is decision support systems, which help healthcare professionals make informed, evidence-based decisions pertaining to patient care management, diagnosis, and treatment.

However, as the utilization of AI systems in healthcare increases, there is a growing concern regarding the opaque nature of these algorithms, commonly referred to as the "black box" problem. The decision-making mechanisms of AI systems can be intricate and challenging to decipher, resulting in potential mistrust and ethical dilemmas among healthcare professionals and patients. As a result, there is an urgent need to develop more transparent, explainable, and interpretable AI systems, particularly for decision support in healthcare.

The primary objective of this study is to tackle the need for interpretability in AI-based healthcare decision support systems by proposing a novel method that incorporates transparency into these systems. The research aims to demonstrate the feasibility and efficacy of the proposed approach and its ability to enhance trust and understanding among healthcare professionals and patients. By providing insight into the underlying mechanisms of AI algorithms and facilitating informed decision-making, this research endeavors to contribute to the development of more ethical and accountable AI systems in the realm of healthcare.

In this chapter, we will first provide a comprehensive literature review of the current state of AI techniques in healthcare decision support systems and explainable AI (XAI) methods. Following the identification of a gap in the existing research, we will present a problem statement, outlining the specific issue to be addressed and listing the research objectives. Next, we will propose a novel methodology to incorporate explainability in AI-based healthcare decision support systems, discussing the advantages and potential challenges of this approach. Finally, we will present the results of our method applied to a healthcare decision-making task, analyzing the implications of our findings and suggesting future research directions in the field of explainable AI in healthcare decision support systems.

II. LITERATURE REVIEW

In this section, our objective is to examine the present body of literature concerning AI techniques in decision support systems for healthcare and explainable AI methods. This will establish a basis for comprehending the current status of the field and pinpointing avenues for future research.

1. **AI Techniques in Healthcare Decision Support Systems:** Due to their capacity to process large amounts of data, detect patterns, and provide accurate predictions, AI-based decision support systems have become increasingly popular in healthcare. Healthcare decision support has been enhanced by the application of several AI techniques, including machine learning. For tasks such as disease diagnosis, prognosis prediction, and treatment

planning, supervised and unsupervised learning algorithms have been utilized, with popular algorithms including neural networks, decision trees, and support vector machines.

2. **Deep Learning**, which is a branch of machine learning, has demonstrated significant achievements in healthcare applications, particularly in medical imaging and natural language processing assignments. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are common deep learning architectures used in these applications, as noted by Litjens et al. (2017) and Rajkomar et al. (2018).

   Deep learning is a subset of machine learning that has shown remarkable achievements in various healthcare applications. In particular, deep learning has demonstrated excellent performance in the fields of medical imaging and natural language processing assignments. The success of deep learning in these applications is due to its ability to learn complex features and patterns in large datasets, which can lead to more accurate predictions.

   Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are two of the most commonly used deep learning architectures in medical imaging and natural language processing applications. CNNs have shown promising results in image classification, segmentation, and detection tasks, while RNNs have been successful in natural language processing tasks such as speech recognition and machine translation. As noted by Litjens et al. (2017) and Rajkomar et al. (2018), these deep learning architectures have significantly improved the accuracy and efficiency of various healthcare applications.

   In deep learning has revolutionized healthcare applications by enabling the development of accurate and efficient models for medical imaging and natural language processing assignments. CNNs and RNNs are two popular deep learning architectures used in these applications, and their success has been noted by researchers such as Litjens et al. (2017) and Rajkomar et al. (2018). These achievements demonstrate the enormous potential of deep learning to improve healthcare outcomes and advance the field of medical research.

   In healthcare decision support systems, reinforcement learning has been utilized for personalized treatment recommendations. This technique can optimize treatment policies based on individual patient data, as demonstrated by Zhang et al. (2019) and Komorowski et al. (2018). Despite the success of these AI techniques in healthcare decision support, their adoption on a large scale has been impeded by concerns over their transparency and interpretability.

III. IMPORTANCE OF EXPLAINABILITY

Explainability refers to the ability of AI systems to provide clear and understandable explanations for their decisions or recommendations. It is important because it promotes transparency, accountability, and trust between healthcare providers and AI systems.

One of the key reasons why explainability is important in healthcare is that it allows healthcare providers to understand how an AI system is arriving at its recommendations. This understanding can help healthcare providers to make better-informed decisions, improve patient outcomes, and avoid potentially harmful or inappropriate interventions. Additionally, when AI systems are transparent in their decision-making process, it can help to reduce the potential for bias, as well as promote equity and fairness in healthcare.

Explainability also promotes trust and understanding between healthcare providers and AI systems. When healthcare providers understand the reasoning behind an AI system's recommendation, they are more likely to trust and accept its recommendations. This can lead to greater adoption of AI systems in healthcare, and ultimately, better healthcare outcomes for patients.

There are several examples of situations where explainability would be critical in healthcare decision-making. For instance, imagine a patient who is diagnosed with a rare disease that requires a specific treatment plan. An AI system may recommend a treatment plan that deviates from the established clinical guidelines. In this situation, explainability is critical because it allows healthcare providers to understand how the AI system arrived at its recommendation. This understanding can help healthcare providers to make an informed decision about whether to follow the AI system's recommendation or not.

Another example is in the case of clinical trials. AI systems can be used to analyze large amounts of data from clinical trials and provide recommendations for treatment plans. In this scenario, explainability is critical because it allows healthcare providers to understand the rationale behind the AI system's recommendations. This understanding can help healthcare providers to identify potential biases in the clinical trial data and make informed decisions about the treatment plan[1].

The importance is , explainability is critical in AI-driven healthcare decision-support systems as it promotes transparency, accountability, and trust between healthcare providers and AI systems. It allows healthcare providers to understand how an AI system arrived at its recommendation, and in turn, make better-informed decisions for their patients. Finally, it is important to note that the implementation of explainability in AI systems requires a collaborative effort between healthcare providers, data scientists, and regulators to ensure domain-specific and stakeholder-centric approaches are considered.

IV. CHALLENGES AND LIMITATIONS IN ACHIEVING EXPLAINABILITY

While explainability is critical in promoting trust and transparency in AI systems, achieving it can be challenging due to the complexity of AI algorithms, the lack of standardized methods for measuring explainability, and the need for balancing explainability with accuracy and efficiency.

One of the main challenges in achieving explainability is the complexity of AI algorithms. Many AI algorithms used in healthcare decision support systems are black-box models, which means that it can be difficult to understand how the model arrived at its decision or recommendation. Additionally, the use of multiple algorithms and models in AI systems can further complicate the process of achieving explainability.

Another challenge is the lack of standardized methods for measuring explainability. There is currently no agreed-upon method for measuring the explainability of AI systems, which can make it difficult to compare different systems and ensure that they are meeting the required standards of explainability[2].

Balancing explainability with accuracy and efficiency in AI systems is another challenge. In some cases, increasing explainability may come at the expense of accuracy or efficiency, which can limit the effectiveness of the AI system. For example, adding additional layers of explainability to an AI system may increase the time required to generate a recommendation, which could limit its use in time-sensitive healthcare decision-making.

Inadequate explainability in healthcare decision-making can have several potential consequences. One consequence is that healthcare providers may be hesitant to trust or adopt AI systems, which could limit their potential benefits. Additionally, inadequate explainability can lead to mistrust and lack of transparency, which can harm patient-provider relationships and potentially compromise patient outcomes.

To address these challenges, several approaches can be taken. One approach is to use interpretable AI algorithms that allow healthcare providers to understand how the model arrived at its decision. Another approach is to develop standardized methods for measuring explainability, which can help to ensure that AI systems meet the required standards of explainability. Additionally, the development of domain-specific and stakeholder-centric approaches to explainability can help to ensure that AI systems are transparent and understandable to the relevant stakeholders.

The challenges be, achieving explainability in AI-driven healthcare decision-support systems is a critical but challenging task. The complexity of AI algorithms, the lack of standardized methods for measuring explainability, and the need for balancing explainability with accuracy and efficiency are among the main challenges. However, addressing these challenges can help to promote trust and transparency in AI systems, and ultimately improve healthcare outcomes. It is important to note that the development of domain-specific and stakeholder-centric approaches to explainability requires a collaborative effort between healthcare providers, data scientists, and regulators.

V. EXPLAINABLE AI METHODS

Explainable AI (XAI) aims to address the "black box" nature of AI systems by providing human-understandable explanations for their decision-making processes. Numerous XAI (explainable AI) approaches have been suggested in the existing body of literature, which includes:

1. **Model-Agnostic Methods:** The goal of these approaches is to furnish justifications for any machine learning model. Some illustrations comprise Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016) and Shapley Additive Explanations (SHAP) (Lundberg & Lee, 2017), which provide explanations at a local level for individual predictions. These techniques approximate the behavior of the model by utilizing more straightforward, understandable models.

2. **Model-Specific Methods:** These methods are tailored to specific types of AI models. For instance, TreeSHAP (Lundberg et al., 2018) extends the SHAP framework for tree-based models, and Layer-wise Relevance Propagation (LRP) (Bach et al., 2015) provides explanations for neural networks by attributing relevance scores to input features. Interpreting the results of artificial intelligence models can be challenging due to the complexity of the models. Various methods are available, but not all of them are suitable for every type of model. In this article, we will focus on model-specific methods, which are designed to be tailored to specific AI models. We will explore two popular model-specific methods: TreeSHAP and Layer-wise Relevance Propagation (LRP) and discuss how they provide valuable insights into model behavior and performance.

   TreeSHAP is a method that extends the SHAP framework, which assigns a value to each input feature to indicate its contribution to the model prediction. However, the SHAP framework is not optimized for tree-based models, which can have complex interaction effects between features. TreeSHAP addresses this limitation by introducing an algorithm that efficiently computes the Shapley values for tree-based models. The main advantage of TreeSHAP is that it can handle complex interaction effects between features, making it an important tool for analyzing tree-based models in domains such as finance, healthcare, and marketing. LRP is a model-specific method that provides explanations for neural networks by attributing relevance scores to input features. Neural networks are powerful models commonly used in image and speech recognition, natural language processing, and other domains. However, they are often seen as "black boxes" due to the difficulty in understanding how they make predictions. LRP addresses this issue by propagating the prediction back through the network and assigning relevance scores to each input feature based on its contribution to the prediction. LRP enables analysts and data scientists to identify which input features are most relevant for prediction accuracy, making it a valuable tool for understanding and interpreting neural network predictions.

   In model-specific methods, such as TreeSHAP and LRP, provide valuable insights into model behavior and performance. By using these methods, analysts and data scientists can gain a better understanding of how AI models make predictions and identify which input features are most important for prediction accuracy. These insights are essential for building trustworthy and interpretable AI systems in various domains[4].

3. **Rule Extraction Methods:** These techniques extract human-readable rules from complex AI models to facilitate interpretation. One example is the TREPAN algorithm (Craven & Shavlik, 1996), which extracts decision trees from trained neural networks. with the increasing prevalence of artificial intelligence (AI) in various industries, ensuring the accuracy and interpretability of AI models has become essential. One of the challenges of using complex AI models is that they can be difficult to understand and interpret, which makes it challenging to base decisions on their outputs. This is where rule extraction methods come in. In this article, we will discuss the benefits of rule extraction methods and explore one popular example, the TREPAN algorithm.

   Rule extraction methods are techniques that extract human-readable rules from complex AI models, making it easier for analysts and data scientists to interpret their outputs. These methods aim to provide a transparent and interpretable representation of a complex model by identifying the most important features and relationships used in the model's predictions. The extracted rules can then be used to gain insights into how the model works and to validate its performance [5].

There are several benefits to using rule extraction methods. Firstly, they can provide a more transparent and interpretable representation of a complex AI model. This can be useful for stakeholders who may not have expertise in AI, allowing them to understand the model's predictions and how they were derived. Secondly, rule extraction methods can help to validate the performance of a model by extracting rules that match with domain knowledge or prior research. Finally, extracted rules can be used to generate hypotheses for further investigation or to develop simpler models that may be easier to implement [6].

In Rule extraction methods are valuable tools for ensuring the accuracy and interpretability of complex AI models. By extracting human-readable rules, these methods enable analysts and data scientists to gain insights into how models make predictions, validate model performance, and generate hypotheses for further investigation. One popular example of rule extraction methods is the TREPAN algorithm, which has been demonstrated to be effective in several applications, as noted by Craven and Shavlik (1996).

4. **Trepan Algorithm-Extracting Decision Trees From Trained Neural Networks:** One popular example of a rule extraction method is the TREPAN algorithm. TREPAN is an algorithm that extracts decision trees from trained neural networks. The algorithm works by analyzing the neural network's internal structure to identify hidden nodes that can be represented as decision trees. These decision trees can then be used to provide a more interpretable representation of the neural network's behavior.

The TREPAN algorithm is particularly useful because it can extract decision trees that match human intuition or prior domain knowledge. This can be used to validate the neural network's performance and to provide explanations for its predictions. Additionally, the decision trees extracted by TREPAN can be used to generate hypotheses for further investigation or to develop simpler models that may be easier to implement.

Rule extraction methods like the TREPAN algorithm can provide a more transparent and interpretable representation of complex AI models. By extracting human-readable rules from these models, analysts and data scientists can gain insights into how they make predictions and validate their performance. The TREPAN algorithm is just one example of a rule extraction method that can be used to extract decision trees from trained neural networks. Other methods, such as rule-based models, decision tables, or logical expressions, can also be used to extract rules from complex AI models.

If you are working with complex AI models, you should consider using rule extraction methods to facilitate interpretation. These methods can provide a more transparent and interpretable representation of your models, making it easier to understand how they make predictions and validate their performance. The TREPAN algorithm is just one example of a rule extraction method that can be used to extract decision trees from trained neural networks, but there are many other methods available depending on your needs and expertise [7].

VI. GAP IDENTIFICATION

While there has been significant progress in the development of XAI methods, their application in healthcare decision-support systems remains limited. There is a need for more research on how to effectively incorporate explainability into these systems without compromising their performance. In addition, many existing XAI methods lack a focus on domain-specific interpretability, which is particularly important in healthcare, where domain expertise plays a crucial role in understanding and evaluating the decision-making process[8].

Moreover, most current XAI methods do not take into account the diverse needs and preferences of different stakeholders in healthcare, such as physicians, nurses, and patients, who may require varying levels of explanation and interpretability. This gap highlights the necessity for a more tailored approach to explainability that caters to the specific requirements of each stakeholder group.

Furthermore, ethical considerations and patient privacy concerns are often not adequately addressed in existing XAI approaches. As AI-based healthcare decision support systems deal with sensitive personal information and have significant implications for patients' well-being, it is crucial that any explainable AI method developed for this domain takes into account the ethical and privacy concerns associated with the use of patient data.

The identified gaps in the current research landscape include:

1. Limited application of XAI methods in healthcare decision support systems
2. Insufficient focus on domain-specific interpretability
3. Lack of consideration for diverse stakeholder needs and preferences
4. Inadequate attention to ethical and privacy concerns

Addressing these gaps in research will contribute to the development of more robust, transparent, and ethically responsible AI systems in healthcare decision support, ultimately leading to better patient outcomes and increased trust in AI-driven decision-making processes.

VII. PROBLEM STATEMENT

Given the identified gaps in the current research landscape, the primary problem that this study seeks to address is the lack of effective explainability in AI-based healthcare decision support systems. The complexity and opacity of these systems can lead to mistrust among healthcare professionals and patients, hinder the adoption of AI in healthcare, and raise ethical concerns. Therefore, there is a pressing need to develop a novel method for incorporating explainability into healthcare decision support systems that caters to the specific requirements of the healthcare domain and its diverse stakeholders.

The **Research Objectives** of this study are as follows:

1. To propose a novel method for integrating explainability into AI-based healthcare decision support systems, with a focus on domain-specific interpretability.
2. To design an approach that takes into consideration the diverse needs and preferences of different stakeholder groups, such as physicians, nurses, and patients, ensuring that the explanations provided are tailored to their requirements.
3. To ensure that the proposed method addresses the ethical and privacy concerns associated with the use of sensitive patient data in AI-driven healthcare decision-making processes.
4. To assess the efficacy of the suggested approach, in terms of its ability to provide meaningful, interpretable explanations without compromising the performance of the decision support system.
5. To explore the potential impact of the proposed method on trust, understanding, and adoption of AI-based healthcare decision support systems among healthcare professionals and patients.

By addressing the problem of explainability in AI-based healthcare decision support systems, this study aims to contribute to the development of more transparent, understandable, and ethically responsible AI systems in healthcare, ultimately enhancing trust and promoting better patient outcomes.

VIII. METHODOLOGY

To address the problem of explainability in AI-based healthcare decision support systems, this study proposes a novel method that integrates domain-specific interpretability, caters to diverse stakeholder needs, and addresses ethical and privacy concerns. The following subsections describe the key components of the proposed method.

1. **Domain-specific Interpretability:** The proposed method aims to make AI systems in healthcare more interpretable by leveraging domain-specific knowledge and concepts. This ensures that the explanations provided by the AI system are not only accurate but also meaningful and easily understandable by healthcare professionals.

   To achieve this, the proposed method incorporates medical ontologies and expert-driven feature extraction. Medical ontologies provide a structured representation of medical knowledge and terminology, which enables the AI system to generate explanations that are grounded in established medical concepts. Expert-driven feature extraction involves identifying the most relevant features for a specific task and extracting them in a way that is consistent with medical knowledge and terminology.

   By incorporating medical ontologies and expert-driven feature extraction, the proposed method generates accurate and interpretable explanations. Healthcare professionals can easily understand and use these explanations to make informed decisions about patient care. This approach has the potential to improve patient outcomes and advance medical research by providing a more transparent and trustworthy AI system.

2. **Tailored Explanations for Stakeholders:** Recognizing that different stakeholder groups have varying needs and preferences when it comes to explanations, the proposed method incorporates a flexible framework for generating explanations tailored to individual stakeholders. This is achieved by allowing users to specify their desired level of detail, complexity, and focus of the explanations, enabling the AI system to generate customized explanations that cater to the specific requirements of each user.

3. **Ethical and Privacy Considerations:** To address ethical and privacy concerns, the proposed method implements a privacy-preserving explanation mechanism that operates on anonymized and aggregated patient data. This ensures that sensitive personal information is protected while still enabling the AI system to provide meaningful explanations. Additionally, the method incorporates an ethical guideline adherence module that evaluates the generated explanations against a set of predefined ethical criteria, ensuring that the AI system's decision-making process aligns with established ethical standards.

   To address ethical and privacy concerns, the proposed method implements a privacy-preserving explanation mechanism that operates on anonymized and aggregated patient data. This approach ensures that sensitive personal information is protected while still allowing the AI system to provide meaningful explanations. The privacy-preserving mechanism involves removing any personal identifiers from the data, such as names and addresses, and aggregating the data to ensure that individual patients cannot be identified. This mechanism helps to protect patient privacy while still allowing for the development of accurate and interpretable AI systems.

   Additionally, the proposed method incorporates an ethical guideline adherence module that evaluates the generated explanations against a set of predefined ethical criteria. This module ensures that the AI system's decision-making process aligns with established ethical standards, such as respect for patient autonomy and non-maleficence. By evaluating the generated explanations against these ethical criteria, healthcare professionals can be confident that the AI system is making decisions in an ethical and responsible manner.

   In the proposed method aims to enhance the interpretability of AI systems in healthcare by leveraging domain-specific knowledge and concepts. The method incorporates medical ontologies and expert-driven feature extraction to generate accurate and interpretable explanations that are grounded in established medical knowledge and terminology. To address ethical and privacy concerns, the method implements a privacy-preserving explanation mechanism that operates on anonymized and aggregated patient data and an ethical guideline adherence module that ensures the AI system's decision-making process aligns with established ethical standards. This approach has the potential to improve patient outcomes and advance medical research by providing a more transparent and trustworthy AI system.

4. **Performance Evaluation and Explainability-Performance Trade-off:** To gauge the effectiveness of the suggested approach, an extensive range of performance metrics will be utilized, comprising accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC). These metrics will be used to assess

the performance of the AI system both with and without the integration of the proposed explainability method, allowing for an analysis of any potential trade-offs between explainability and performance.

5. **Experimental Setup and Dataset:** The proposed method will be evaluated on a real-world healthcare dataset containing anonymized patient records related to a specific medical condition. The dataset will be split into training, validation, and testing subsets to ensure a rigorous evaluation of the AI system's performance and the effectiveness of the proposed explainability method.

   By integrating domain-specific interpretability, providing tailored explanations for stakeholders, and addressing ethical and privacy concerns, the proposed method aims to enhance the explainability of AI-based healthcare decision support systems, ultimately leading to greater trust and adoption among healthcare professionals and patients.

IX. RESULTS

In this section, we present the results of the proposed method applied to the real-world healthcare dataset, demonstrating its effectiveness in providing meaningful, interpretable explanations without compromising the performance of the decision support system. We also compare the results with existing methods to highlight the advantages of our approach.

1. **Performance Evaluation:** The performance of the AI system, both with and without the integration of the proposed explainability method, was assessed using the aforementioned performance metrics. The findings demonstrate that upon integration of the explainability technique, the AI system upheld elevated degrees of accuracy, precision, recall, F1-score, and AUC-ROC. These outcomes suggest that the inclusion of interpretability does not significantly impact the overall functioning of the decision support system.

2. **Explainability Assessment:** To evaluate the quality of the explanations generated by the proposed method, a set of domain experts were asked to rate the explanations in terms of their understandability, relevance, and consistency with established medical knowledge. The results show that the explanations provided by our method were rated highly by the experts, indicating that the proposed method successfully generates meaningful and interpretable explanations that align with domain-specific knowledge.

3. **Stakeholder Satisfaction:** A user study was conducted with healthcare professionals and patients to assess their satisfaction with the tailored explanations provided by the proposed method. The results revealed that both groups found the customized explanations helpful and easily understandable, demonstrating the effectiveness of our approach in catering to the diverse needs and preferences of different stakeholder groups.

4. **Comparison with Existing Methods:** Finally, the proposed method was compared with existing XAI methods in terms of performance, explainability, and stakeholder satisfaction. Our method outperformed the existing methods in generating more understandable and domain-specific explanations while maintaining a comparable level of performance in the decision support system. describes how the proposed method was compared with existing XAI methods in terms of performance, explainability, and

stakeholder satisfaction. The results showed that the proposed method outperformed the existing methods in generating more understandable and domain-specific explanations while maintaining a comparable level of performance in the decision support system.

In terms of performance, the proposed method was found to be just as effective as existing XAI methods in supporting healthcare decision-making. This was demonstrated by the fact that the integration of explainability did not significantly affect the performance of the decision support system. This means that healthcare professionals can use the decision support system with confidence, knowing that it is both accurate and transparent.

In terms of explainability, the proposed method was found to be superior to existing XAI methods in generating more understandable and domain-specific explanations. This is important because healthcare decisions are complex and can involve a wide range of factors. By providing tailored explanations that are grounded in established medical knowledge and terminology, the proposed method can help to ensure that healthcare professionals and patients are better equipped to make informed decisions.

Finally, in terms of stakeholder satisfaction, the proposed method was found to be more effective than existing XAI methods in meeting the diverse needs of different stakeholders, including physicians, nurses, and patients. This was demonstrated by the fact that domain experts and stakeholders rated the explanations generated by the proposed method as understandable, relevant, and consistent with medical knowledge. By tailoring explanations according to user preferences, the proposed method can help to ensure that stakeholders receive information that is customized, understandable, and relevant to their needs.

Overall, the comparison with existing XAI methods demonstrates that the proposed method represents a significant step forward in enhancing explainability in AI-driven healthcare decision support systems. By providing more transparent, understandable, and ethically responsible AI systems, the proposed method has the potential to transform the way that healthcare decisions are made, ultimately fostering greater trust and adoption among healthcare professionals and patients.

5. **Summary of Results:** The results of our study demonstrate that the proposed method effectively incorporates explainability into AI-based healthcare decision support systems without compromising performance. The method generates meaningful, domain-specific explanations tailored to the needs of different stakeholders, ultimately fostering greater trust and understanding among healthcare professionals and patients.

X. DISCUSSION

In this section, we analyze the results and their implications, discuss the strengths and limitations of the proposed method, and address potential ethical considerations that may arise in the context of explainable AI in healthcare decision support systems.

1. **Implications of Results:** The results of our study have several important implications. First, they demonstrate the feasibility of incorporating explainability into AI-based healthcare decision support systems without significantly affecting their performance (Ribeiro et al., 2016). This finding is crucial in addressing concerns about the potential trade-offs between explainability and system performance. Second, the generation of domain-specific, tailored explanations can enhance trust and understanding among healthcare professionals and patients, fostering greater acceptance and adoption of AI-driven decision-making processes in healthcare (Holzinger et al., 2017).

2. **Strengths and Limitations:** The main strengths of the proposed method lie in its ability to generate domain-specific, understandable explanations and cater to the diverse needs of different stakeholder groups. However, there are also limitations to our approach. For instance, the reliance on medical ontologies and expert-driven feature extraction may limit the generalizability of the method to other domains or medical conditions that lack well-defined ontologies. Moreover, the user study conducted to assess stakeholder satisfaction was limited in terms of sample size and diversity, which may affect the generalizability of the findings.

3. **Ethical Considerations:** As AI-based healthcare decision support systems deal with sensitive patient data and have significant implications for patient outcomes, it is crucial to address ethical considerations (Mittelstadt et al., 2016). Our method incorporates privacy-preserving mechanisms and ethical guideline adherence modules to address these concerns. However, ongoing research is necessary to ensure that the evolving ethical landscape of AI in healthcare is adequately considered and incorporated into explainable AI methods.

4. **Future Research Directions:** Future research can focus on extending the proposed method to other domains and medical conditions, as well as refining the privacy-preserving and ethical guideline adherence mechanisms to ensure robust protection of sensitive patient data and ethical decision-making. Additionally, further investigation into the impact of explainable AI on trust, understanding, and adoption among different stakeholder groups can help to inform the development of more effective and user-centered explainability methods in healthcare decision support systems.

REFERENCES

[1] Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain? Arxiv preprint arxiv:1712.09923.

[2] Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, l. (2016). The ethics of algorithms: mapping the debate. Big data & society, 3(2), 205395171667967.

[3] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "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).

[4] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: visual explanations from deep networks via gradient-based localization. In proceedings of the ieee international conference on computer vision (pp. 618-626).

[5] Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. Arxiv preprint arxiv:1708.08296.

[6] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). 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).

[7] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: visual explanations from deep networks via gradient-based localization. In proceedings of the ieee international conference on computer vision (pp. 618-626).

[8] Samek, W., Wiegand, T., & Müller, k. R. (2017). Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. Arxiv preprint arxiv:1708.08296.