Explainable AI for Medical Image Analysis in Medical Cyber-Physical Systems: Enhancing Transparency and Trustworthiness of IoMT

Wei Liu, Feng Zhao, Achyut Shankar, Carsten Maple, James Dinesh Peter, Byung-Gyu Kim, *Senior Member, IEEE*, Adam Slowik, Bidare Divakarachari Parameshachari, and Jianhui Lv

***Abstract*—This study explores the application of explainable artificial intelligence (XAI) in the context of medical image analysis within medical cyber-physical systems (MCPS) to enhance transparency and trustworthiness. Meanwhile, this study proposes an explainable framework that integrates machine learning and knowledge reasoning. The explainability of the model is realized when the framework evolution target feature results and reasoning results are the same and are relatively reliable. However, using these technologies also presents new challenges, including the need to ensure the security and privacy of patient data from Internet of Medical Things (IoMT). Therefore, attack detection is an essential aspect of MCPS security. For the MCPS model with only sensor attacks, the necessary and sufficient conditions for detecting attacks are given based on the definition of sparse observability. The corresponding attack detector and state estimator are designed by assuming that some IoMT sensors are under protection. It is expounded that the IoMT sensors under protection play an important role in improving the efficiency of attack detection and state estimation. The experimental results show that the XAI in the context of medical image analysis within MCPS improves the accuracy of lesion classification, effectively removes low-quality medical images, and realizes the explainability of recognition results. This helps doctors understand the logic of the system's decision-making and can choose whether to trust the results based on the explanation given by the framework.**

***Index Terms*—Explainable artificial intelligence, medical image, medical cyber-physical systems, IoMT, sparse observability.**

Received 4 August 2023; revised 28 September 2023; accepted 14 November 2023. Date of publication 27 November 2023; date of current version 7 April 2025. This paper was supported by National Natural Science Foundation of China under Grant 62202247. *(Corresponding author: Feng Zhao.)*

Wei Liu and Feng Zhao are with the Department of Emergency Medicine, Shengjing Hospital of China Medical University, Shenyang 110055, China (e-mail: lw2487551906@163.com; zhaojz120@163.com).

Achyut Shankar is with the Department of Cyber Systems Engineering, WMG, University of Warwick, CV74AL Coventry, U.K., and also with the Centre of Research Impact and Outreach, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab 140417, India (e-mail: ashankar2711@gmail.com).

Carsten Maple is with the Secure Cyber Systems Research Group, WMG, University of Warwick, CV74AL Coventry, U.K. (e-mail: cm@warwick.ac.uk).

James Dinesh Peter is with the Karunya Institute of Technology and Sciences, Coimbatore 641114, India (e-mail: dineshpeter@karunya.edu).

Byung-Gyu Kim is with the Department of IT Engineering, Sookmyung Women's University, Seoul 140-742, South Korea (e-mail: bg.kim@sookmyung.ac.kr).

Adam Slowik is with the Department of Computer Science and Engineering, Koszalin University of Technology, 75-453 Koszalin, Poland (e-mail: adam.slowik@tu.koszalin.pl).

Bidare Divakarachari Parameshachari is with the Department of Electronics and Communication Engineering, Nitte Meenakshi Institute of Technology, Bengaluru 560064, India (e-mail: paramesh@nmit.ac.in).

Jianhui Lv is with the Department of Networks, Peng Cheng Laboratory, Shenzhen 518057, China (e-mail: lvjh@pcl.ac.cn).

Digital Object Identifier 10.1109/JBHI.2023.3336721

I. Introduction

SINCE 2016, the new generation of artificial intelligence (AI) technology represented by machine learning, especially deep learning, has been developing in a more advanced, complex, and autonomous direction, bringing new transformative opportunities to economic and social development [1]. AI applications are ushering in a "species explosion," increasingly penetrating all walks of life and all aspects of human life, and are expected to shape a new economic and social form. At the same time, science and technology ethics has increasingly become a "must-option" in the current development of AI technology and industrial applications. All walks of life have explored AI ethical principles, frameworks, and governance mechanisms [2]. Although not all AI models have black-box nature and are not more inexplicable than non-AI technology, traditional software, or human programs, for now, machine learning models, especially deep learning models, are often non-transparent and challenging for humans to understand [3]. In the future, the continuous progress of AI is expected to bring about autonomous systems with independent perception, learning, decision-making, and action. However, the practical utility of these systems is limited by the machine's ability to explain its thoughts and actions to human users adequately. The transparency and explainability of AI systems are critical if users understand, trust, and effectively manage the new generation of AI companions. Therefore, explainable AI (XAI) has become an emerging field of AI research recently, and academia and industry have been exploring methods and tools for understanding the behavior of AI systems [4], [5].

Medical AI applications must be transparent to increase trust with doctors [6]. Transparency in AI-driven medical decision-making is essential for several reasons. Humans often use perception and reasoning to make decisions together [7]. Consider how doctors make and interpret decisions during the diagnostic process when recognizing medical images. For example, when a doctor determines whether a cell is diseased, he/she can first perceive the overall characteristics of the cell shown by the cell image based on his/her own screening experience and draw a perceptual conclusion. Then, based on the medical knowledge about cytopathies, the doctor observes the morphological characteristics of each cell (such as the size of the nucleus, the level of the karyoplasmic-cytoplasm ratio, etc.) and combines these characteristics and knowledge to reach a reasoning conclusion. The doctor will combine and compare the two results to come up with a final diagnosis and use relevant medical knowledge to explain the reason for the diagnosis. The rapidly increasing number of digitized medical images gradually exceeds the range of appropriate numbers that can be processed manually. However, these processes have now been mostly replaced by machine learning, and there are already many deep learning models for medical image recognition. However, as neural network models become more complex and accurate, we must compromise accuracy and explainability. Especially with the further advancement of deep learning models, millions or even billions of parameters are often involved. As a result, researchers sometimes make a model that works without fully understanding why. Therefore, XAI is critical for medical AI applications to be accepted and integrated into practice. It can even map out patients' most critical clinical features based on the analysis of medical images.

The convergence of medical technology, AI, and the Internet of medical things (IoMT) has given rise to medical cyber-physical systems (MCPS) that revolutionize healthcare delivery. MCPS integrates physical medical devices, sensors, communication networks, and computational algorithms to create interconnected systems that enhance medical imaging, diagnosis, and treatment [8]. Integrating MCPS with XAI techniques adds transparency and interpretability to AI-driven medical image analysis [9]. XAI methods aim to provide understandable and interpretable explanations for the decision-making processes of complex AI models. By incorporating XAI into MCPS for medical image analysis, healthcare professionals gain insights into the AI algorithms' reasoning and outputs, enhancing their understanding and trust in the results [10].

Medical images are an essential source of clinical information for healthcare providers, allowing them to visualize the internal structures and functions of the human body [11]. MCPS is designed to integrate physical components with digital technologies to create a more efficient and effective healthcare environment. IoMT sensors play a critical role in MCPS, providing the physical interface between the patient and the digital system [12]. In the context of medical imaging, IoMT sensors can be used to capture physiological data while the patient is undergoing an imaging procedure [13]. However, using these technologies also presents new challenges, including the need to ensure the security and privacy of patient data. As MCPS integrate physical components with digital technologies, they are vulnerable to various cyber threats, including unauthorized access, data breaches, malware, and cyberattacks [14]. These threats can compromise patient data and disrupt medical services, leading to severe consequences. Therefore, attack detection is an essential aspect of MCPS security. The proposed framework recognizes the paramount importance of data security and privacy, especially in the context of IoMT, where patient data is transmitted across various devices and platforms. To address these challenges, the framework incorporates robust security protocols and mechanisms to protect patient data. Additionally, the necessity of attack detection is an essential aspect of MCPS security. By implementing attack detectors and state estimators, the framework ensures that any malicious activities or breaches are promptly identified and mitigated. Furthermore, the framework emphasizes the role of protected IoMT sensors in enhancing the efficiency of attack detection and state estimation. By prioritizing security at every data processing and transmission stage, the framework ensures the confidentiality, integrity, and availability of medical data. The main novelties of this study are:

1) An explainable framework that integrates machine learning and knowledge reasoning for medical image analysis within MCPS was proposed. The framework aims to enhance transparency and trustworthiness of IoMT in the decision-making process. It achieves explainability by ensuring that the framework's evolution target feature results and reasoning results are the same and relatively reliable.
2) The security and privacy challenges were addressed in MCPS, particularly in the context of medical image analysis. It highlights the importance of attack detection to ensure the integrity and reliability of MCPS. This study presents necessary and sufficient conditions for detecting attacks in MCPS with IoMT sensor-based attacks based on the definition of sparse observability. By assuming the protection of specific IoMT sensors, the study designs corresponding attack detectors and state estimators. It emphasizes the role of protected IoMT sensors in improving the efficiency of attack detection and state estimation. These security measures contribute to safeguarding the integrity and trustworthiness of MCPS in medical image analysis.

The rest of this study is organized as follows. Section II reviews the related works. Section III proposes an XAI framework integrating machine learning and knowledge reasoning for medical image classification. Section IV studies medical image analysis in MCPS. The experimental results are provided in Section V. Lastly, Section VI concludes this paper.

II. Related Work

Explainability allows humans to understand the logic of a system's decision-making and better understand why an outcome might fail. Machine learning explainability can be classified into intrinsic and post-explainability [15]. Intrinsic explainability

means that the machine learning model is explainable, while post-explainability is the use of explainability to explain complex machine learning models [16], [17]. XAI has richer content in the medical field, and explainability must consider the disease and the algorithm itself [18], [19], [20]. There are many kinds of diseases, and the appearance and character of diseases are different, so it is necessary to build a model for specific diseases and to explain the corresponding work. So far, researchers have proposed some explainable works for different model structures and conditions, including Alzheimer's, lung cancer, heart disease, and more [21], [22]. Additionally, many visualization methods only provide the image description of the model's focus area. This description method is fuzzy, and doctors need to observe the area further to determine the semantics of the image area. Tjoa and Guan extensively evaluated interpretability methodologies proposed in various research studies and subsequently classified them into distinct categories [23]. These categories encompass diverse dimensions within interpretability research, ranging from methods that yield intuitively interpretable insights to those that explore intricate patterns. Pneumonia, renowned as a highly transmissible infection, poses a significant threat to many of the population, particularly individuals with compromised immune systems. To address this issue, Sheu et al. [24] successfully employed an XAI approach to develop an interpretable classification system for pneumonia infection. There needs to be more consensus regarding integrating computational pathology systems into pathologists' workflow. In light of this, Tosun et al. [25] provided a comprehensive overview of XAI's applications in the anatomic pathology field. These applications were found to enhance the efficiency and accuracy of pathology practice. Siddiqui and Doyle [26] devised a framework to investigate the concept of trustworthiness as it pertains to the utilization of AI in the medical field, specifically focusing on key stakeholders involved in developing medical devices. Within this framework, particular attention was given to the element of explainability in AI models, which was thoroughly examined by evaluating a collection of XAI methods.

MCPS represents a revolutionary integration of medical devices, computational algorithms, and communication networks. These systems are designed to enhance healthcare delivery by combining physical components with digital intelligence and connectivity. MCPS is vital in improving patient care, monitoring, diagnosis, treatment, and overall healthcare management. At their core, MCPS leverages cyber-physical systems (CPS) principles, which involve the tight integration of computational and physical elements [27]. CPS combines real-time sensing, data processing, and control to create intelligent systems interacting with the physical world. When applied to the medical field, these systems create a new paradigm of healthcare where the physical and digital realms converge. To address this gap, Wang et al. [28] conducted a study in which they developed a comprehensive threat model and put forth potential attack techniques targeted explicitly at medical imaging devices. Li et al. [29] conducted an in-depth investigation into the security and safety risks prevalent in the networks of MCPS. Kocabas et al. [30] presented a comprehensive overview of the general architecture of MCPS, which consists of four distinct layers: data acquisition, data aggregation, cloud processing, and action. Priya et al. [31] introduced energy-conscious security strategies explicitly designed for MCPS. Almohri et al. [32] conducted a comprehensive analysis focusing on the roles of stakeholders and system components within the context of MCPS.

The existing research on XAI in medical image analysis highlights the importance of explainability in machine learning models and distinguishes between intrinsic and post-explainability. However, the current works need a comprehensive focus on the medical domain, where disease-specific considerations are necessary. The paper addresses this gap by emphasizing the need for disease-specific models and explanations. Additionally, the integration of MCPS into healthcare has shown significant potential for improving patient care and overall management. However, the security and privacy concerns of MCPS, particularly related to medical imaging devices, have been overlooked in some instances. The paper recognizes this gap and highlights the importance of attack detection, threat modeling, and access control to ensure the security and integrity of MCPS. By addressing these weaknesses, the paper aims to contribute to developing effective security strategies and encryption schemes within MCPS.

III. XAI Framework Integrating Machine Learning and Knowledge Reasoning

We represent the overall features as target features, multiple fine-grained features as sub-features, and outcomes (e.g., lesion or not) as decision targets. Among them, the target feature covers many sub-features: the correlation features between expert knowledge and data related to the decisive goal. In contrast, the decisive goal refers to the common goal of a system or model. Based on the above ideas, this study proposes an XAI model integrating machine learning and knowledge reasoning for medical image recognition. The integration of machine learning and knowledge reasoning in the framework is designed to provide a holistic and comprehensive approach to medical image analysis. In cases where there might be discrepancies or conflicts between the machine learning predictions and knowledge reasoning insights, the framework employs a reconciliation mechanism. Additionally, the framework might flag such discrepancies for manual review by medical professionals, ensuring that decisions are made with the utmost care and consideration. The structure of the model is shown in Fig. 1.

The model recognizes the importance of catering to diverse patient populations and has algorithms that can adapt to demographic variations, genetics, and environmental factors. The model ensures that its analysis is inclusive and representative by continuously learning from a diverse dataset that encompasses various patient profiles. Furthermore, the model's emphasis on knowledge reasoning allows for integrating domain-specific knowledge, ensuring that a comprehensive understanding of diverse patient populations informs its decisions.

1) Knowledge Reasoning Module: The knowledge reasoning module provides domain knowledge and business rules for reasoning decisions, that is, ontology library O and rule library K

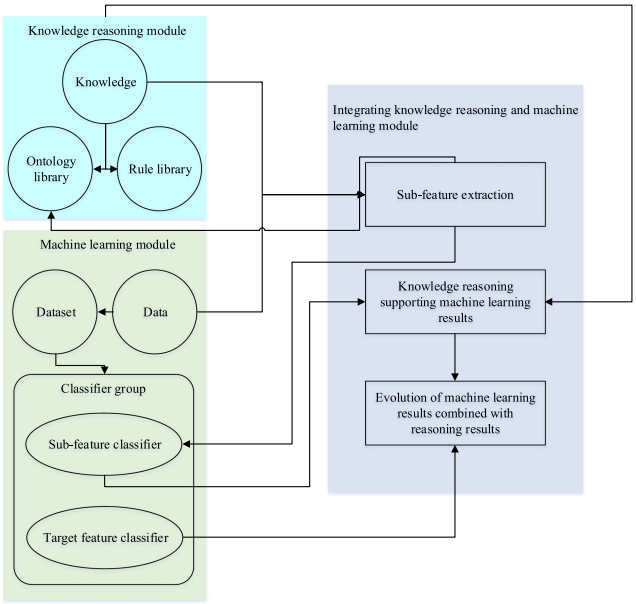


Fig. 1. XAI model integrating machine learning and knowledge reasoning for medical image recognition.

related to decision objectives. According to the domain knowledge related to the decisive goal, through the steps of knowledge extraction, integration, and processing, the ontology library O is constructed for the decisive goal, which expresses the classes related to the decisive goal and the relationship between the classes. The ontology library O supports ontology web language (OWL) [33], where the target feature class contains the child feature class. The obtained expert knowledge about decisive goals is transformed into business rules, and the rule library K is formed, which supports semantic web rule language [34].

2) Machine Learning Module: The machine learning module provides a classifier group including a target feature classifier C and multiple sub-feature classifiers $C_1, C_2, \ldots, C_n$, the results of which are used for reasoning and evolution. The classifier group is obtained by training the neural network group and datasets D, D_1, $D_2, \ldots, D_n$. The neural network group consists of a target feature classification neural network N and n sub-feature classification neural network $N_1, N_2, \ldots, N_n$. Dataset D is used for training N, and the data annotation of D takes the decisive goal as the classification standard. Dataset $D_1, D_2, \ldots, D_n$ are used for training $N_1, N_2, \ldots, N_n$respectively, and data labeling $D_1, D_2, \ldots, D_n$ is based on their corresponding sub-features.

3) Integrating Knowledge Reasoning and Machine Learning Module: Based on the extracted sub-features, the ontology library O in the knowledge reasoning module constructs sub-feature classes, and the machine learning module constructs sub-feature classifiers. Sub-features are the associated features between expert knowledge and data related to decisive goals. Knowledge and data features are associated and correspond, and the overlapping features are sub-features. According to the knowledge features and data features related to the decisive goal, the framework extracts n sub-features $f_1, f_2, \ldots, f_n$. Then, the sub-feature classes of ontology library O are constructed according to f_1, $f_2, \ldots, f_n$. The annotation categories of the datasets D_1, $D_2, \ldots, D_n$ use f_1, $f_2, \ldots, f_n$ as the classification standard, respectively, and the constructed $n$sub-feature classifiers C_1, $C_2, \ldots, C_n$ also use f_1, $f_2, \ldots, f_n$. as the standard to classify the data to be classified.

Medical image t to be classified passes through the classifier group to obtain the classification results R_c of target feature classifier C and R_1, $R_2, \ldots, R_n$ of sub-feature classifier C_1, $C_2, \ldots, C_n$. R_1, $R_2, \ldots, R_n$ are mapped to the entity data of their corresponding sub-feature classes in ontology library$O$. Knowledge reasoning is carried out based on ontology library O and rule library K, and the reasoning result R_r is obtained. Results R_c and R_r are target feature results. That is, the framework decides that data t is R_c and R_r. The two target feature results, R_c and R_r, will evolve later to achieve explainability.

Combining the target feature result R_c (machine learning result) and the target feature result R_r (reasoning result) for evolution, the framework makes corresponding decisions according to whether R_c and R_r are the same and whetherR_c and R_r are reliable. To measure whether the results are reliable, this study introduces a metric to evaluate the quality of the results—credibility. The reliability $A_{R_c}(R_c)$and the reliability $A_{R_r}(R_r)$ of the target feature result are calculated, respectively, and then the two results are combined with evolving. The results are deemed reliable if both components produce consistent results that meet the accuracy and robustness thresholds. Additionally, the model may employ statistical methods and validation techniques to assess the results' reliability further. The framework is modular and scalable, allowing for easy integration of new advancements and techniques in AI and machine learning. Individual framework components can be updated or replaced by a modular architecture without affecting the overall system, ensuring that as new methods and algorithms emerge in medical image analysis, the framework can seamlessly incorporate them, maintaining its relevance and effectiveness. The framework's emphasis on explainability directly caters to the need for better communication between healthcare providers and patients. By providing clear, understandable explanations for its decisions, the framework enables healthcare providers to convey AI-driven decisions to patients in a comprehensible and transparent manner, empowers patients with knowledge about their diagnosis or treatment, and fosters trust in the AI-driven processes, ensuring that patients feel involved and informed throughout their care journey.

IV. Medical Image Analysis in MCPS

To ensure the safety of the medical image of patients and the safety of system model parameters and improve the accuracy of lesion classification, this study proposes an MCPS-based system model for collaborative analysis of medical images.

A. MCPS Model

This Section will analyze and construct necessary and sufficient conditions for attack detectability and the security estimation problem by exploiting s-sparse observability for the MCPS model where IoMT sensors are attacked [35]. Based on

these conditions, the corresponding attack detection observer is designed.

Assuming an MCPS model where IoMT sensors are attacked, i.e.,

$$\sum\nolimits_a = \begin{cases} x(t+1) = Mx(t) + Nu(t) \\ y_a(t) = Px(t) + D_a a(t) \end{cases} \tag{1}$$

where $x(t) \in R^n$ represents the internal state of the system, which is generally unknown, while $u(t) \in R^m$ represents the control input of the system, $a(t) \in R^p$ represents the malicious attack vector injected by the attacker into the actuator network and sensor network, $y(t) \in R^p$ represents the measured output after the malicious attack, and $t \in N$ represents the discrete time. The relevant parameters M, N, and P are known system matrices with suitable dimensions. When the attack vector $a(t)$ exists, the purpose of the system is to accurately detect the attack's existence and estimate the system's true state by designing an appropriate attack detection observer and security state estimator.

Without loss of generality, assuming that measurements are collected from time 0, all outputs of the ith IoMT sensor at time t can be denoted as follows.

$$Y_i^a(t) = \Psi_i x(0) + E_i(t) + F_i U(t) \tag{2}$$

where $Y_i^a(t) = [y_i(0) \quad \dots \quad y_i(t)]^T$, $A_i(t) = [a_i(0) \quad \dots \quad a_i(t)]^T$, $U(t) = [u(0) \quad \dots \quad u(t)]^T$, $\Psi_i = [C_i \quad \dots \quad C_i M^t]^T$, and $F_i = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ P_i N & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ P_i M^{t-1} N & P_i M^{t-2} N & P_i N & 0 \end{bmatrix}$. Since the control input $U(t)$ is known data, (1) can be simplified as follows.

$$\sum\nolimits_a \begin{cases} x(t+1) = Mx(t) \\ y_a(t) = Px(t) + D_a a(t) \end{cases} \tag{3}$$

Correspondingly, (2) can be simplified as follows.

$$Y_i^a(t) = \Psi x(0) + A_i(t) \tag{4}$$

Then, the matrix form of all the measured data is as follows.

$$Y^a(t) = \Psi x(0) + A(t) \tag{5}$$

where $Y^a(t) = [Y_1(t) \quad \cdots \quad Y_p(t)]^T$, $A(t) = [A_1(t) \quad \cdots \quad A_p(t)]^T$, and $\Psi = [\Psi_1 \quad \cdots \quad \Psi_p]^T$.

Furthermore, (5) can be integrated as follows.

$$Y^a(t) = \Psi x[0] + A(t) = \begin{bmatrix} \Psi & I \end{bmatrix} \begin{bmatrix} x(0) \\ A(t) \end{bmatrix} = Qz(t) \tag{6}$$

where $Q = [\Psi \quad I]$, and $z(t) = [x(0) \quad A(t)]^T$. If the attack vector $A(t)$ is a linear combination of the column vectors of the matrix Ψ, then the attack vector $A(t)$ is undetectable at this time; that is, the attack is undetectable at this time.

Lemma 1: If the attack vector $A(t)$ is a linear combination of the column vectors of the matrix Ψ, then the attack vector $A(t)$ is undetectable at this time; that is, the attack is undetectable at this time.

Therefore, from the above exceptional cases, it is necessary to study the necessary and sufficient conditions for an attack to be detectable. Until then, the following assumptions are made about the system and the attacker.

Assumption 1: Among the p IoMT sensors, r IoMT sensors are protected by the system; that is, the attacker cannot attack these r selected protected IoMT sensors.

Assumption 2: Among the p IoMT sensors, the attacker can attack s IoMT sensors at most.

Apart from Assumption 2, no additional assumptions are made to the attacker; that is, the attacker can master the system parameters, control commands $u(t)$, and IoMT sensor measurements $y(t)$, and even know the true state $x(t)$ of the system at any time.

It should be noted that from the above exceptional case if the attacker can control and invade all IoMT sensor signals, the obtained measured signal $y(t)$ does not contain any accurate information about the system's internal state $x(t)$. The system cannot perform attack detection and security state estimation now.

The rationality of the above assumption lies in the following aspects. On the one hand, the attacker can only control part of the IoMT sensors because of the available resources and attack capabilities. On the other hand, the degree of protection of IoMT sensors in different locations differs, and some IoMT sensors are strictly protected.

For ease of exposition, this study introduces s-sparse observability [36].

Assume that the linear control system is s-sparse observable that is, for any set $\Gamma \subseteq \{1, \dots, p\}$, $|\Gamma| = p - s$, and (M, P_Γ) is observable.

In other words, the system is s-sparse observable if the remaining $p - s$ dimensional system is still observable after removing any s IoMT sensors.

For the convenience of subsequent matrix representation and operation, when it comes to the definition of s-sparse observability, P_Γ means that the corresponding s rows in matrix P are set to zero vectors; that is, matrix P_Γ is still $p \times n$ dimensional. It is essentially consistent with the observable after deleting s rows.

From the above s-sparse observability definition, if the system is s-sparse observable, then for any set $S \subset \{1, 2, \dots p\}$ and $|S| = s$, the following system is completely observable.

$$\sum\nolimits_{\overline{S}} = \begin{cases} x(t+1) = Mx(t) + Nu(t) \\ y_{\overline{S}}(t) = P_{\overline{S}} x(t) \end{cases} \tag{7}$$

where $\overline{S}$ is the complement of the set S and represents the system formed by the remaining $p - s$ IoMT sensors after removing all IoMT sensors in the set S.

For a given MCPS model, an exhaustive method can be used to find the most prominent exponent s that satisfies the sparse observability of the system, as follows.

Step 1: Let $s = 1$;

Step 2: For any set S that satisfies $\overline{S} = s$, determine whether the system $(M, P_{\overline{S}})$ is entirely observable. If not, go to Step 4; if yes, continue to execute;

Step 3: Let $s = s + 1$, skip to Step 2;

Step 4: Output the system's maximum sparse observable index $s-1$.

Using the above-proposed definition of attack detection as well as the sparse observability definition, under Assumptions 1 and 2, this study focuses on the attack detection problem and the state estimation problem. Aiming at the MCPS system under IoMT sensor attack to solve the problem of attack detection, on the one hand, it is necessary to establish the detectability condition of the attack; on the other hand, it is necessary to design the corresponding attack detector under the premise of the attack detection. In this process, the superiority embodied in protecting part of the IoMT sensors in the system needs to be deeply studied. Aiming at the MCPS system under IoMT sensor attack to solve the problem of state estimation, on the one hand, it is necessary to establish a sufficient and necessary condition for the state to be safely estimated; on the other hand, under the premise of the state can be safely estimated, the corresponding state estimator is designed to reconstruct the state of the system. Similarly, the superiority of protecting some IoMT sensors must be explored in depth.

B. Attack Detector Design

According to the above, the necessary and sufficient conditions for the detection of attacks can be obtained.

Theorem 1: The following two descriptions are equivalent [36].

i) It is detectable for an attacker to attack any s IoMT sensors.

ii) The system is s-sparsely observable.

Assume that the system protects r IoMT sensors. From the perspective of algebraic theory, the rows of matrix P corresponding to these r IoMT sensors should be linearly independent as far as possible. It may be possible to set it as the first r rows of P and write $rank(P)=m$, for $r \geq m$, $rank(P_1,\ldots,r)=m$. Suppose that the first m of P is a maximal linearly independent group of P, then any row of P can be linearly represented by the first m rows, that is, there exist $k_1^i,\ldots,k_m^i$, such that $P_i = k_1^i P_1 + \ldots + k_m^i P_m$. Then, the attack detector can be designed according to Theorem 2.

Theorem 2: For $r \geq m$, if $y_i(t) \neq k_1^i y_1(t) + \ldots + k_m^i y_m(t)$, $i = r+1,\ldots,p$, then IoMT sensor i is attacked; otherwise it is not attacked.

From the above analysis, it can be seen that the system needs to protect at most m IoMT sensors, and the linear representation relationship between IoMT sensor measurements can be used to detect the presence of attacks and identify the specific location of the attack. Suppose $r \leq m$ and r rows of the corresponding matrix P are linearly independent. In that case, it may be set as the first r rows of P to be linearly independent, and $m-r$ rows are found to form a maximum linearly independent group of row vectors of P, which may be set as the first m rows. In this case, the attack detector can be constructed as follows.

$$\sum O = \begin{cases} z(t+1) = Mz(t) + Nu(t) + L\left[y_a(t) - y'(t)\right] \\ y(t) = P_{\{1,\ldots,m\}} z(t) \\ r(k) = y_a(t) - y(t) \end{cases} \tag{8}$$

where the matrix L makes $M - LP_{\{1,\ldots,m\}}$ a strictly stable matrix, all its characteristic roots are in the unit circle. Then there is the following Theorem 3.

Theorem 3: For $r < m$, if the above attack detector has $r(t) \to 0$, then the m IoMT sensors are not attacked. Currently, whether or not the remaining $p-m$ IoMT sensors are attacked is determined according to Theorem 2. Otherwise, the m IoMT sensors are attacked.

Therefore, the attack detector can be designed as follows.

Step 1: Parameters initialization. $rank(P)=m$;

Step 2: If $r \geq m$, go to Step 3; Otherwise, go to Step 4;

Step 3: Determine whether the measurement data meets the relevant linear representation relationship. If so, there is no attack; otherwise, there is an attack;

Step 4: Take the $m-r$ rows of C and the protected r rows to form a P maximal linear independent group matrix $P_{\{1,\ldots,m\}}$, and use the detector to determine whether it is attacked. If attacked, the output system is attacked; Otherwise, use the linear expression to judge whether the remaining $p-m$ IoMT sensors are under attack.

C. State Estimator Design

This subsection will study the relationship between indistinguishable attacks and secure state estimation, analyze the necessary and sufficient conditions under which secure state estimation can be performed, and design the corresponding state estimator.

The necessary and sufficient conditions for attack discriminability can be obtained, like attack detectability.

Theorem 4: The following two descriptions are equivalent.

i) The attacker attacking any s IoMT sensors is distinguishable.

ii) The system is $2s$-sparsely observable.

Due to the $2s$-sparse observability of the system, any $p-2s$ IoMT sensor data in the system can be used to estimate the initial state $x(t)$) if there is no attack. However, if there is an attack, some $p-2s$ IoMT sensors must not be attacked because the attack is s-distinguishable. The purpose of security estimation is to find this set and estimate the true state of the system with this set.

Under the premise of protecting some IoMT sensors in the system, if $r \geq m$, then at this time, the protected IoMT sensor set can ensure the system's observability, so the corresponding estimator can be designed by using the protected set, and the specific format is as follows.

$$\sum O = \begin{cases} z(t+1) = Mz(t) + Nu(t) + L\left[y\{1,\ldots,r\}(t) - y(t)\right] \\ y(t) = P_{\{1,\ldots,r\}} z(t) \\ r(k) = y_a(t) - y(t) \end{cases} \tag{9}$$

If $r < m$, in this case, since the protected IoMT sensors cannot directly estimate the true state of the system, it is necessary to perform a traversal to find the set such that the above estimator converges to zero. Since there are already r IoMT sensors

protected, the set of IoMT sensors to be traversed at this point is C^s_{p-r} sets. Denote the set of all possible attacks as S, then

$$S = \left\{S_1, \ldots, S_{C^s_{p-r}}\right\}, |S_j| = s, j = 1, \ldots, C^s_{p-r} \quad (10)$$

V. Experiments and Comparison Analysis

The experiments focus on the need for disease-specific models and explanations. Additionally, the integration of MCPS into healthcare has shown significant potential for improving patient care and overall management. Cervical cancer is a common malignant tumor that seriously threatens women's health. It is estimated that as many as 100000 women have cervical cancer yearly in China alone [37]. Fortunately, cervical cancer is highly preventable and treatable with early detection and proper medical intervention. Cervical cancer screening plays a significant role in early prevention, and abnormal squamous cells of the cervix have great significance in diagnosing cervical cancer.

The framework incorporates advanced attack detection mechanisms tailored for MCPS. The framework can identify anomalies or potential threats by continuously monitoring the system's state and analyzing data patterns. The state estimator plays a crucial role in this process, assessing the system's current state against expected patterns and flagging any deviations. Furthermore, the framework emphasizes the role of protected IoMT sensors, which are safeguarded against external threats and play a pivotal role in enhancing attack detection efficiency. By integrating these mechanisms, the framework ensures robust security and timely detection of potential threats.

IoMT sensors under protection are equipped with enhanced security protocols and encryption mechanisms to safeguard against external threats and unauthorized access. These sensors provide reliable and secure data inputs in the framework. Their protection ensures that the data they transmit is authentic and untampered, enhancing the system's overall security. Furthermore, by ensuring data integrity from these sensors, the framework can achieve higher efficiency in attack detection and state estimation, as it can place higher trust in the data from protected sensors.

A. Setup

1) Sub-Feature Extraction: According to the abnormal squamous cells of cervix image and the expert knowledge of atypical squamous cells: cannot exclude high-grade squamous intraepithelial lesion (ASC-H) cell morphology, this study extracts four sub-features f_1, f_2, f_3, f_4,which are cell size, karyoplasmic ratio, size of nucleus, and degree of hyperchromatism of nucleus [38]. In this study, we choose to recognize ASC-H cells to verify the feasibility of the explainable framework. The ASC-H cell recognition framework has improved the recognition accuracy and achieved the explainability of the recognition results. When using this recognition framework, doctors can choose whether to trust the results according to the explanations given by the framework. It is worth mentioning that false negatives should be avoided in cervical cancer screening, that is, to avoid the situation that cells with original lesions are considered to have no lesions. Therefore, the ASC-H cell identification framework also takes suspected ASC-H as a recognition category to avoid missing diseased cells.

2) Ontology and Rule Libraries Construction: In this study, classes and relationships between classes to recognize ASC-H cells were extracted from medical knowledge about the morphology of ASC-H cells, OWL was used to construct the ASC-H cell recognition ontology library, and the construction platform was Protégé. Rule Library consists of four rules, which are transformed from ASC-H's expert knowledge of cell morphology medicine. (i) The properties of the constituent parts of the cell are also the properties of the cell. (ii) In cell morphology, if small cells, high karyoplasmic ratio, enlarged nucleus, and slightly hyperchromatic nucleus all match, the cells are considered ASC-H. (iii) In the cell morphology, if any of the high karyoplasmic ratios and the enlarged nucleus is met, the cell is suspected of ASC-H (Sus-ASC-H). (iv) In cell morphology, if small cells, high karyoplasmic ratio, enlarged nucleus, and slightly hyperchromatic nucleus all match, the cells are considered not to be ASC-H (Non-ASC-H).

3) Dataset and Classifier Group Construction: Dataset D, $D_1, D_2, \ldots, D_n$ are all composed of several images of squamous cells of the cervix with a size of 128 $\times$ 128. The annotation category of D is cell type, and the annotation category of D_1, $D_2, \ldots, D_n$ is based on sub-features f_1, f_2, f_3, f_4. The structure of the target feature classification neural network N is any neural network model used for classification. In this study, three models, convolutional neural networks (CNN), variational auto-encoder (VAE) [39], and VGG19 [40], are selected to implement three target feature classifiers, respectively.

4) Knowledge Reasoning Supporting Machine Learning Results: Assuming that a cell image t to be recognized is input into the ASC-H cell recognition framework after it is classified, the target feature classifier C gets the target feature result R_c. Sub-feature classifiers C_1, C_2, C_3, C_4 obtain four sub-feature results R_1, R_2, R_3, R_4, which are mapped to entity data of the corresponding sub-feature class in ontology library O of ASC-H cells. Then, knowledge reasoning is performed based on the ontology library O and rule library K, and the reasoning result R_r is obtained. In the following, the two results, R_c andR_r, will be evolved to achieve the explainability of the results.

B. Validation of the ASC-H Cell Recognition Framework

The validation set consists of 500 images of cervical squamous epithelial cells with a size of 128 $\times$ 128. To calculate each classifier's evaluation value, the positive class is set to Non-ASC-H, and the negative class is set as the sum of Sus-ASC-H and ASC-H. After training with different sizes of datasets, the accuracy and F1 value of each classifier on the verification set are shown in Table I. The reasoning results of each sample in the verification set are obtained through the knowledge reasoning in supporting machine learning results in the framework. The proposed method combines each sample's target feature classifier results with its reasoning for evolution. After the evolution of each target feature classifier combined with the knowledge reasoning in supporting machine learning results, the accuracy and F1 scores are improved, as shown in Table II.

TABLE I
EVALUATION VALUES FOR EACH CLASSIFIER

Classifier	Dataset size	Accuracy	F1
CNN	1000	0.8525	0.8526
	2000	0.8600	0.8539
	3000	0.8750	0.8671
VAE	1000	0.8025	0.7844
	2000	0.8175	0.8150
	3000	0.8425	0.8503
VGG19	1000	0.8875	0.8748
	2000	0.8950	0.8812
	3000	0.8900	0.8753

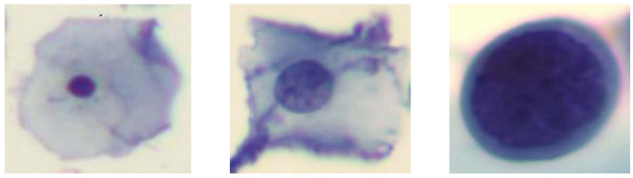
TABLE II
EVALUATION VALUES AFTER EVOLVED CLASSIFIER

Classifier	Dataset size	Accuracy	F1
CNN	1000	0.8675	0.8669
	2000	0.8800	0.8735
	3000	0.8900	0.8818
VAE	1000	0.8225	0.8154
	2000	0.8425	0.8441
	3000	0.8675	0.8732
VGG19	1000	0.8975	0.8851
	2000	0.9050	0.8915
	3000	0.9050	0.8907

Experiments show that the evolution method of machine learning results combined with the reasoning results proposed in this paper improve the performance of the target feature classifier, and the improvement effect is more pronounced when the accuracy of the classifier is low. When the data amount used by the classifier and its accuracy in the training process has reached a relatively saturated degree, the evolution method will play a small role in improving the classifier's performance. By combining the reasoning results, the evolutionary method can permanently eliminate part of the error results of the target feature classifier. When the framework is constantly used to classify cells, the evolutionary process is also iterative, and the target feature classifier will be continuously optimized.

The framework incorporates multiple layers of validation and verification to minimize potential errors or inaccuracies. In cases where discrepancies are identified between machine learning predictions and knowledge reasoning insights, the framework may flag such instances for manual review by medical professionals, ensuring that critical medical decisions are not solely based on automated processes but are subject to expert review and judgment.

The model places a strong emphasis on ethical considerations in AI-driven medical decisions. By providing clear and transparent explanations for its decisions, the model ensures that medical professionals remain informed and in control of the decision-making process. This fosters accountability, as doctors can evaluate and validate AI-driven decisions before implementing them. Furthermore, the model's design incorporates feedback loops, allowing continuous evaluation and improvement based on real-world outcomes. The model addresses the ethical concerns associated with AI-driven medical decisions by prioritizing transparency, accountability, and continuous improvement.



(a) (b) (c)

Fig. 2. Images of cells to be classified. (a) Non-ASC-H, (b) Sus-ASC-H, (c) ASC-H.

An image of a cell to be recognized is input into the ASC-H cell recognition framework, and the cell image is shown in Fig. 2(a). After it passes through the classifier group, classifier C gets the target feature result R_c as Non-ASC-H. The four sub-feature resultsR_1, R_2, R_3, and R_4 obtained by the classifier are respectively medium cell, low karyoplasmic ratio, normal nucleus and severe nuclear stained. Knowledge reasoning is carried out by combining R_1, R_2, R_3, R_4, ASC-H cell recognition ontology library O, and rule library$K$. The reasoning result is that R_r is Non-ASC-H.

Input a cell image to be recognized into the ASC-H cell recognition framework, and the cell image is shown in Fig. 2(b). R_c is obtained as Sus-ASC-H through the framework, and R_r is obtained as Non-ASC-H. Therefore, the framework decides that the cell is Sus-ASC-H and optimizes the classifier and rule library rules of R_r. According to the parameters recorded by the result evidence chain G, it can be found that the reliability evaluation value K_r based on the rule library is not high. That is, there may be errors in the rule library. The specificity of classifier C_1 is low; that is, C_1 has a low probability of correctly judging medium cells. According to the reason for the failure of R_r, check the rules in rule library K and correct any errors. C_1 is optimized to improve the classification accuracy of the framework.

Input a cell image to be recognized into the ASC-H cell recognition framework, and the cell image is shown in Fig. 2(c). R_c is obtained as ASC-H through the framework, and R_r is obtained as ASC-H. R_c and $R_r$are the same, and the reliability of both is higher than 0.8, so the framework considers R_c and R_r to be relatively reliable. Therefore, the framework decides that the cell is ASC-H and uses the sub-features resulting in small cells, high karyoplasmic ratio, enlarged nucleus, and slightly stained, as well as rule2 of rule library *K*, to explain the decision that the cell image t is ASC-H. Nevertheless, the doctor can understand the framework's logic in recognizing the cell as ASC-H and decide whether to believe the recognition based on the explanation.

It can be seen that the cell image in Fig. 2(b) found the reasons for Rr failure through the evidence chain G, and the classification

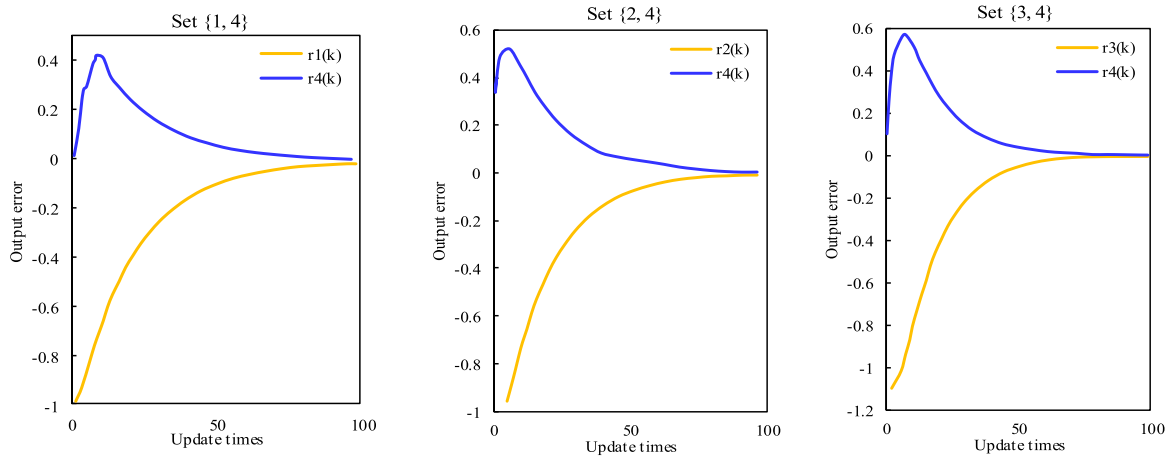


Fig. 3. Detector residuals without IoMT sensor attack.

accuracy of the frame in Fig. 2(c) is improved after optimizing the corresponding part based on these reasons. When two results are the same and relatively reliable, the framework endows the results with explainability, which solves the issue that the rules cannot reflect the actual decision of the model to a large extent.

C. Numerical Simulation for MCPS

In this subsection, the proposed attack detector and state estimator algorithms are used for simulation verification to prove the rationality of the theory. The adopted MCPS model is as follows.

$$\sum\nolimits_a = \begin{cases} x(t+1) = Mx(t) \\ y(t) = Px(t) + a(t) \end{cases} \tag{11}$$

where $M = \begin{bmatrix} 1 & 0.4 \\ 0 & 0.6 \end{bmatrix}$, $C = \begin{bmatrix} 1 & 0 \\ 1 & -1 \\ 1 & -2 \\ 0 & 5 \end{bmatrix}$, and $a(t)$ is the attack vector. However, the parameter L of the attack detector is $L = 10^{-2} \begin{bmatrix} 1 & 2 \\ 5 & 1 \end{bmatrix}$.

First, it is verified that there is no attack in the system, and the attack detector residual eventually converges to zero. Assuming that sensor 1 is protected in the system, different sensor sets are selected to form maximal linearly independent groups under the design of the attack detection algorithm, and the residuals of the attack detectors corresponding to different sets are shown in Fig. 3. Obviously, if there is no attack, any set satisfying the conditions can be used to ensure that the residual of the detector eventually converges to zero. It is easy to verify that the linear representation relationship between the relevant measurement data holds. Therefore, when the attack does not exist, the attack detector can ensure that no false alarm occurs; the system is correctly judged not to be attacked.

The sparse observability index of the above system is $s = 2$. Therefore, a necessary and sufficient condition for an attack to be detectable for the system is that at most two IoMT sensors are under attack. The rationality of attack detectors will be analyzed in the following for detectable attacks. Suppose the system protects the fourth IoMT sensor while the attack occurs at IoMT sensors 2 and 3. In this case, the $\{P_i, P_4\}$ corresponding to any IoMT sensor and IoMT sensor 4 can be chosen as the maximal linearly independent group of matrix P, and $i = 1, 2, 3$. The corresponding attack vector is $a_i(t) = P_i M^{t-1} e(0)$, $i = 2, 3$, where $e(0) = [1 \quad 0.5]^T$. Then, under different sets, the residuals of the attack detector of the attack detector are shown in Fig. 4.

According to Fig. 4, when the sets $\{2, 4\}$ and $\{3, 4\}$ are selected, an attack can be detected immediately. When the set $\{1, 4\}$ is selected, it is not immediately possible to determine whether the IoMT sensor is attacked. However, it is easy to verify that $y_2(t) \neq y_1(t) - 1/5 y_4(t)$ by using the linear representation relationship between IoMT sensor data, $P_2 = P_1 - 1/5 P_4$. Therefore, the presence of an attack can also be judged.

Since the sparse observability index is 2, it is possible to perform secure state estimation when the number of IoMT sensors under attack is 1. Without considering the number of protected IoMT sensors, the set of IoMT sensors to be searched is $C_4^2 = 6$. If IoMT sensor 4 is protected, the set to be searched is $C_{4-1}^{2-1} = 3$. At this point, the corresponding attack detector residuals are shown in Fig. 5.

Fig. 5 shows that the residual corresponding to the set $\{1, 4\}$ converges to zero, so it can be judged that IoMT sensor 1 is under attack, and the measured data of IoMT sensors 2 and 3 can be used for state estimation. It is not difficult to analyze that when only one IoMT sensor is protected, the number of elements in the exhaustive search set for state estimation can be reduced from C_p^s to C_{p-1}^{s-1}. Therefore, when the system becomes more complex, protecting some IoMT sensors can significantly reduce the set of state estimators to search. As the number of protected IoMT sensors increases, the set to search becomes smaller and smaller. Therefore, protecting some IoMT sensors is of great significance in improving the estimation efficiency of the system.

The verified rationality of attack detection and state estimation in MCPS can significantly impact medical image analysis, enhancing transparency and trustworthiness in several ways. Attack detection and state estimation mechanisms help identify and mitigate cybersecurity threats within MCPS. Ensuring the integrity and security of the system can improve the accuracy

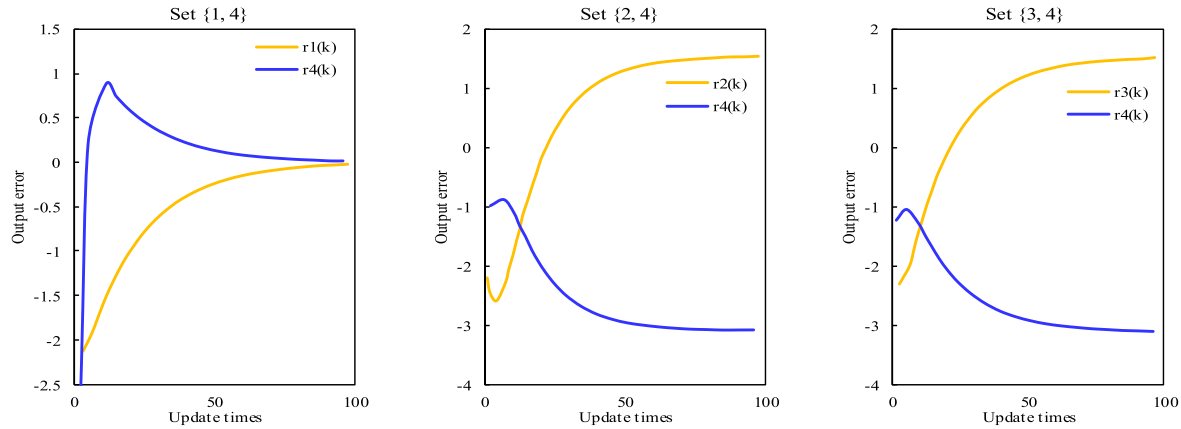


Fig. 4. Detector residuals under different detection sets when protecting IoMT sensor 4, IoMT sensors 2 and 3 under attack.

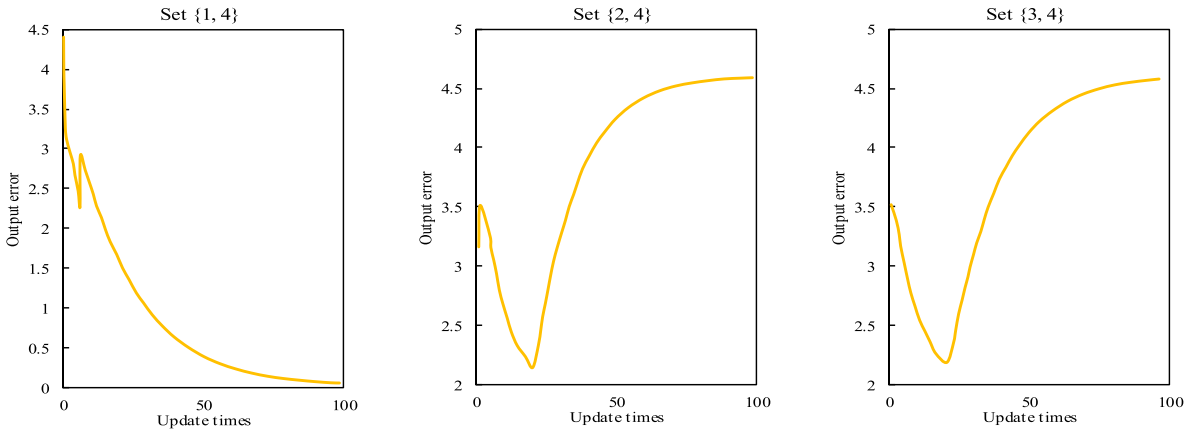


Fig. 5. Estimator residuals for different IoMT sensor sets.

and reliability of medical image analysis. This is crucial in maintaining the trustworthiness of the analysis results and preventing potential malicious attacks that could compromise patient data or influence the analysis outcomes. Adversarial attacks in the context of medical image analysis refer to deliberate manipulations of images or algorithms to mislead the analysis process. The verified rationality of attack detection mechanisms can detect and prevent such attacks, ensuring that the medical images used for analysis are authentic and unaltered. This protects the integrity of the analysis process and enhances transparency by assuring that the results are based on legitimate and unmanipulated data [41], [42]. By ensuring the verified rationality of attack detection and state estimation in MCPS, medical image analysis can benefit from improved accuracy, protection against adversarial attacks, preservation of data privacy, and enhanced transparency and trustworthiness.

Transparency and trustworthiness are paramount in MCPS, especially when integrated with the IoMT. The proposed framework ensures transparency by providing clear explanations for its decisions, which is achieved through the explainability feature, where the framework's evolution target feature results and reasoning results align and are deemed reliable. Medical professionals can understand the logic behind the system's decision-making by offering clear explanations. The rigorous testing and validation of the framework ensures trustworthiness. The experimental results demonstrate the framework's capability to improve lesion classification accuracy and effectively filter out low-quality medical images. The framework establishes its trustworthiness in the medical community by showcasing its reliability and effectiveness in real-world scenarios.

VI. Conclusion

This study demonstrates the potential of XAI in medical image analysis within MCPS to enhance transparency and trustworthiness in decision-making. The proposed framework effectively integrates machine learning and knowledge reasoning to explain the model's decision-making process. This approach improves the accuracy of lesion classification, removes low-quality medical images, and enables the explainability of recognition results. However, using XAI in medical image analysis also presents new challenges, including the need to ensure the security and privacy of patient data. Attack detection is a crucial aspect of MCPS security, and the study proposes necessary and sufficient conditions for detecting attacks based on sparse observability. Future work should address the security and privacy challenges

associated with XAI in medical image analysis and develop advanced techniques for secure data transmission, storage, and access control. Additionally, novel approaches to improving the interpretability of the XAI framework and enhancing its robustness against adversarial attacks are essential research directions. Integrating XAI in medical image analysis can revolutionize healthcare systems of IoMT by providing transparent and trustworthy decision-making processes, ultimately improving patient care.

APPENDIX

Proof of Theorem 2

Proof. Since the first m rows of P are the maximum linearly independent group of P, there is, $r+1, r+2, \ldots, p$, rows of P can be linearly represented by the first m rows. Taking the $r+1$th row as an example, there are coefficients $k_1^{r+1}, \ldots, k_m^{r+1}$ that are not all zero making the following equations true.

$$P_{r+1} = k_1^{r+1} P_1 + \ldots + k_m^{r+1} P_m$$

If the ith IoMT sensor is not attacked, $y_i(t) = P_i M^t x(0)$ holds. Therefore, if the $r+1$th IoMT sensor is not attacked, then we have

$$\begin{aligned} y_{r+1}(t) &= P_{r+1} M^t x(0) \\ &= \left(k_1^{r+1} P_1 + \ldots + k_m^{r+1} P_m\right) M^t x(0) \\ &= k_1^{r+1} P_1 M^t x(0) + \ldots + k_m^{r+1} P_m M^t x(0) \\ &= k_1^{r+1} y_1(t) + \ldots + k_m^{r+1} y_m(t) \end{aligned}$$

If the $r+1$th IoMT sensor is not attacked, then we have

$$y_{r+1}(t) = k_1^{r+1} y_1(t) + \ldots + k_m^{r+1} y_m(t)$$

If the $r+1$th IoMT sensor is attacked, then $y_{r+1}(t) = P_{r+1} M^t x(0) + a_{r+1}(t)$, since $a_{r+1}(t)$ is not zero, we have

$$y_{r+1}(t) \neq k_1^{r+1} y_1(t) + \ldots + k_m^{r+1} y_m(t)$$

Proof of Theorem 4

Proof: (i)→(ii), we use proof by contradiction to prove. Hypothesis (ii) is not satisfied; the system does not satisfy the $2s$-sparse observable. It is shown that there exists a IoMT sensor set $\Gamma \subseteq \{1, \ldots, p\}, |\Gamma| = p - 2s$, which makes the matrix (M, P_Γ) unobservable, that is, there exists a non-zero initial state $x(0)$, and the output $y_\Gamma(t)$ of the system is always zero. Then, there is $\overline{\Gamma}$ and it satisfies $|\overline{\Gamma}| = 2s$, S_1 and S_2 are subsets of $\overline{\Gamma}$, and $S_1 \cup S_2 = \overline{\Gamma}$, $S_1 \cap S_2 = \Phi$.

The attack vector is constructed as follows.

$$\begin{aligned} a_1(t) &= P_{S_1} M^t x(0) \\ a_2(t) &= -P_{S_2} M^t x(0) \\ Y^1(t) &= P M^t 0_{n\times 1} - P_{S_1} M^t x(0) \\ Y^2(t) &= P M^t 0_{n\times 1} - P_{S_2} M^t x(0) \end{aligned}$$

Then, $Y^1(t) = P M^t x(0) + a_1(t)$, $Y^2(t) = P M^t 0_{n\times 1} + a_2(t)$, and $Y^1(t) = Y^2(t)$. Since there is no undistinguishable attack, $x(0) = 0_{n\times 1}$. Therefore, there exists contradiction.

(ii)→(i), we still use proof by contradiction to prove. Suppose there are indistinguishable attacks $a_1(t)$ and $a_2(t)$, and the attack sets $|S_1| = s$ and $|S_2| = s$ are satisfied. Then there are different initial states $x(0)_1$ and $x(0)_2$, such that

$$\begin{aligned} P M^t x(0)_1 + a_1(t) &= P M^t x(0)_2 + a_2(t) \\ P M^t \left(x(0)_1 - x(0)_2\right) &= a_2(t) - a_1(t) \end{aligned}$$

Since $|S_1 \cup S_2| \leq |S_1| + |S_2| = 2s$, we have

$$P M^t \left(x(0)_1 - x(0)_2\right) + \left(a_1(t) - a_2(t)\right) = 0$$

Again, the system satisfies $2s$-sparse observability, so $x(0)_1 - x(0)_2 = 0$. Therefore, there exists a contradiction.

REFERENCES

[1] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electron. Markets*, vol. 31, no. 3, pp. 685–695, 2021.

[2] T. Hagendorff, "The ethics of AI ethics: An evaluation of guidelines," *Minds Mach.*, vol. 30, no. 1, pp. 99–120, 2020.

[3] G. Yang, Q. Ye, and J. Xia, "Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond," *Inf. Fusion*, vol. 77, pp. 29–52, 2021.

[4] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[5] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning explainability methods," *Entropy*, vol. 23, no. 1, 2021, Art. no. 18.

[6] C. Longoni, A. Bonezzi, and C. K. Morewedge, "Resistance to medical artificial intelligence," *J. Consum. Res.*, vol. 46, no. 4, pp. 629–650, 2020.

[7] J. Zhang and D. Tao, "Empowering things with intelligence: A survey of the progress, challenges, and opportunities in artificial intelligence of things," *IEEE Internet Things J.*, vol. 8, no. 10, pp. 7789–7817, May 2021.

[8] I. Lee et al., "Challenges and research directions in medical cyber-physical systems," *Proc. IEEE*, vol. 100, no. 1, pp. 75–90, Jan. 2012.

[9] M. R. Kanjee and H. Liu, "Authentication and key relay in medical cyber-physical systems," *Secur. Commun. Netw.*, vol. 9, no. 9, pp. 874–885, 2016.

[10] B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis," *Med. Image Anal.*, vol. 79, 2022, Art. no. 102470.

[11] G. Litjens et al., "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.

[12] A. Alzahrani, M. Alshehri, R. AlGhamdi, and S. K. Sharma, "Improved wireless medical cyber-physical system (IWMCPS) based on medical learning," *Healthcare*, vol. 11, no. 3, 2023, Art. no. 384.

[13] M. Shehab et al., "Machine learning in medical applications: A review of state-of-the-art methods," *Comput. Biol. Med.*, vol. 145, 2022, Art. no. 105458.

[14] N. Dey, A. S. Ashour, F. Q. Shi, S. J. Fong, and J. M. R. S. Tavares, "Medical cyber-physical systems: A survey," *J. Med. Syst.*, vol. 42, no. 4, pp. 1–3, 2018.

[15] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Mueller, "Explaining deep neural networks and beyond: A review of methods and applications," *Proc. IEEE*, vol. 109, no. 3, pp. 247–278, Mar. 2021.

[16] I. Ahmed, G. Jeon, and F. Piccialli, "From artificial intelligence to explainable artificial intelligence in industry 4.0: A survey on what, how, and where," *IEEE Trans. Ind. Inform.*, vol. 18, no. 8, pp. 5031–5042, Aug. 2022.

[17] D. Gunning and D. W. Aha, "DARPA's explainable artificial intelligence program," *AI Mag*, vol. 40, no. 2, pp. 44–58, 2019.

[18] R. K. Sheu and M. S. Pardeshi, "A survey on medical explainable AI (XAI): Recent progress, explainability approach, human interaction and scoring system," *Sensors*, vol. 22, no. 20, 2022, Art. no. 8068.

[19] A. M. Antoniadi et al., "Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: A systematic review," *Appl. Sci.*, vol. 11, no. 11, 2021, Art. no. 5088.
[20] R. Nyrup and D. Robinson, "Explanatory pragmatism: A context-sensitive framework for explainable medical AI," *Ethics Inf. Technol.*, vol. 24, no. 1, 2022, Art. no. 13.
[21] Y. Zhang, Y. Weng, and J. Lund, "Applications of explainable artificial intelligence in diagnosis and surgery," *Diagnostics*, vol. 12, no. 2, 2022, Art. no. 237.
[22] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G. Yang, "XAI-explainable artificial intelligence," *Sci. Robot.*, vol. 4, no. 37, 2019, Art. no. eaay7120.
[23] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021.
[24] R.-K. Sheu, M. S. Pardeshi, K.-C. Pai, L.-C. Chen, C.-L. Wu, and W.-C. Chen, "Interpretable classification of pneumonia infection using eXplainable AI (XAI-ICP)," *IEEE Access*, vol. 11, pp. 28896–28919, 2023.
[25] A. B. Tosun, F. Pullara, M. J. Becich, D. L. Taylor, J. L. Fine, and S. C. Chennubhotla, "Explainable AI (xAI) for anatomic pathology," *Adv. Anatomic Pathol.*, vol. 27, no. 4, pp. 241–250, 2020.
[26] K. Siddiqui and T. E. Doyle, "Trust metrics for medical deep learning using explainable-AI ensemble for times series classification," in *Proc. IEEE Can. Conf. Elect. Comput. Eng.*, 2022, pp. 370–377.
[27] F. Pasqualetti, F. Dorfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Trans. Autom. Control*, vol. 58, no. 11, pp. 2715–2729, Nov. 2013.
[28] Z. Wang, P. Ma, X. Zou, J. Zhang, and T. Yang, "Security of medical cyber-physical systems: An empirical study on imaging devices," in *Proc. IEEE INFOCOM Conf. Comput. Commun. Workshops*, 2020, pp. 997–1002.
[29] Z. T. Li, L. Cheng, Y. Zhang, and D. G. Feng, "Understanding and mitigating security risks of networks on medical cyber physical system," in *Proc. 16th Int. Conf. Wireless Algorithms, Syst., Appl.*, 2021, pp. 123–134.
[30] O. Kocabas, T. Soyata, and M. K. Aktas, "Emerging security mechanisms for medical cyber physical systems," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 13, no. 3, pp. 401–416, May/Jun. 2016.
[31] J. S. Priya, S. P. Rajagopalan, and M. Ramakrishnan, "Medical cyber physical system security-mitigating attacks using trust model," *IEEE/ACM Trans. J. Med. Imag. Health Inf.*, vol. 6, no. 7, pp. 1572–1575.
[32] H. Almohri, L. Cheng, D. Yao, and H. Alemzadeh, "On threat modeling and mitigation of medical cyber-physical systems," in *Proc. IEEE/ACM Int. Conf. Connected Health: Appl., Syst. Eng. Technol.*, 2017, pp. 114–119.
[33] J. Chen, P. Hu, E. Jimenez-Ruiz, O. M. Holter, D. Antonyrajah, and I. Horrocks, "OWL2Vec*: Embedding of OWL ontologies," *Mach. Learn.*, vol. 110, no. 7, pp. 1813–1845, 2021.
[34] B. Pittl and H. G. Fill, "A visual modeling approach for the semantic web rule language," *Semantic Web*, vol. 11, no. 2, pp. 361–389, 2020.
[35] N. Tarfulea, "Observability for initial value problems with sparse initial data," *Appl. Comput. Harmon. Anal.*, vol. 30, no. 3, pp. 413–427, 2011.
[36] Y. Shoukry, P. Nuzzo, A. Puggelli, A. L. Sangiovanni-Vincentelli, S. A. Seshia, and P. Tabuada, "Secure state estimation for cyber-physical systems under sensor attacks: A satisfiability modulo theory approach," *IEEE Trans. Autom. Control*, vol. 62, no. 10, pp. 4917–4932, Oct. 2017.
[37] H. Wen, Q. Guo, X. Zhou, X. Wu, and J. Li, "Genomic profiling of Chinese cervical cancer patients reveals prevalence of DNA damage repair gene alterations and related hypoxia feature," *Front. Oncol.*, vol. 11, 2022, Art. no. 792003.
[38] A. Goyal, A. P. Patel, T. L. Dilcher, and S. A. Alperstein, "Effects of implementing the dual papanicolaou test interpretation of ASC-H and LSIL following bethesda 2014: A retrospective comparison with LSIL-H and ASC-H," *Amer. J. Clin. Pathol.*, vol. 154, no. 4, pp. 553–558, 2020.
[39] X. Jin, W. Gong, J. Kong, Y. Bai, and T. Su, "PFVAE: A planar flow-based variational auto-encoder prediction model for time series data," *Mathematics*, vol. 10, no. 4, 2022, Art. no. 610.
[40] M. M. Hasan, N. Islam, and M. M. Rahman, "Gastrointestinal polyp detection through a fusion of contourlet transform and Neural features," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 3, pp. 526–533, 2022.
[41] S. H. Wang, Y. Zhang, X. C. Cheng, X. Zhang, and Y. D. Zhang, "PSSPNN: PatchShuffle stochastic pooling neural network for an explainable diagnosis of COVID-19 with multiple-way data augmentation," *Comput. Math. Methods Med.*, vol. 2021, 2022, Art. no. 6633755.
[42] Y. D. Zhang, X. Zhang, and W. G. Zhu, "ANC: Attention network for COVID-19 explainable diagnosis based on convolutional block attention module," *Comput. Model. Eng. Sci.*, vol. 127, no. 3, pp. 1037–1058, 2021.