Information Sciences 662 (2024) 120212

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Fuzzy inference system with interpretable fuzzy rules: Advancing explainable artificial intelligence for disease diagnosis—A comprehensive review

Jin Cao [a], Ta Zhou [a,b], Shaohua Zhi [a], Saikit Lam [c,d], Ge Ren [a], Yuanpeng Zhang [a,e,f], Yongqiang Wang [a], Yanjing Dong [a], Jing Cai [a,d,f,*]

[a] Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Hong Kong, China
[b] School of Electrical and Information Engineering, Jiangsu University of Science and Technology, Zhenjiang 212100, China
[c] Department of Biomedical Engineering, The Hong Kong Polytechnic University, Hong Kong, China
[d] Research Institute of Smart Aging, The Hong Kong Polytechnic University, Hong Kong, China
[e] Department of Medical Informatics, Nantong University, Nantong 226007, China
[f] The Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen 518057, China

ARTICLE INFO

Keywords:
Explainable artificial intelligence
Interpretability
Fuzzy rule
Disease diagnosis
Fuzzy inference system

ABSTRACT

Interpretable artificial intelligence (AI), also known as explainable AI, is indispensable in establishing trustable AI for bench-to-bedside translation, with substantial implications for human well-being. However, the majority of existing research in this area has centered on designing complex and sophisticated methods, regardless of their interpretability. Consequently, the main prerequisite for implementing trustworthy AI in medical domains has not been met. Scientists have developed various explanation methods for interpretable AI. Among these methods, fuzzy rules embedded in a fuzzy inference system (FIS) have emerged as a novel and powerful tool to bridge the communication gap between humans and advanced AI machines. However, there have been few reviews of the use of FISs in medical diagnosis. In addition, the application of fuzzy rules to different kinds of multimodal medical data has received insufficient attention, despite the potential use of fuzzy rules in designing appropriate methodologies for available datasets. This review provides a fundamental understanding of interpretability and fuzzy rules, conducts comparative analyses of the use of fuzzy rules and other explanation methods in handling three major types of multimodal data (i.e., sequence signals, medical images, and tabular data), and offers insights into appropriate fuzzy rule application scenarios and recommendations for future research.

1. Introduction

Applications of artificial intelligence (AI), from smartphones to natural image recognition, navigation and autonomous vehicles, have rapidly proliferated in people's daily lives. In the medical domain, however, AI applications in real-world clinical settings have lagged behind applications in other settings, despite extensive AI research being conducted in recent decades to obtain insights to support clinical decision- making in a wide range of areas, such as cancer prognostication [1], tumor responses [2], medical image

* Corresponding author.
E-mail address: jing.cai@polyu.edu.hk (J. Cai).

https://doi.org/10.1016/j.ins.2024.120212
Received 2 August 2023; Received in revised form 4 January 2024; Accepted 22 January 2024
Available online 26 January 2024

synthesis [3], and the detection of Coronavirus disease-2019 (COVID-19) infection [4]. This protraction of act in the field of medicine is largely attributed to the increasing, yet unsatisfied, demand for interpretable AI.

Interpretable AI, also referred to as eXplainable AI (XAI), is important for establishing trustable AI models for bench-to-bedside translation [5]. XAI is particularly important given AI's potential far-reaching implications for human well-being. Conventional prediction models, developed using simple methods (e.g., linear regression) and comprising a limited number of predictors, are easy to comprehend and interpret using plain language. However, since the paradigm shift in the AI era, studies have mostly focused on designing complex and sophisticated methods, such as deep neural networks, to harvest highly accurate prediction models, regardless of whether the models are interpretable. The complexity of these models greatly restricts the possibility of key stakeholders (i.e., physicians, patients, policymakers, and even researchers) comprehending the processes underlying how and why the AI reaches the ultimate prediction results. There is also a lack of interaction between humans and machines, which is a prerequisite for implementing trustworthy AI, in the medical domain.

In light of the increasing awareness of the aforementioned issues, the European Union enforced a legal right in 2019 for individuals to access "meaningful information about the logic behind automated decisions using their data" [6]. In addition, research is growing on interpretable AI in terms of its precise definition, categorization, assessment criteria, and the interpretability of explanation methods.

Various explanation methods have been adopted by researchers. These methods can be broadly classified into two categories: model-agnostic and model-specific interpretability techniques. Model-agnostic techniques are independent of the model structure and work by unraveling the relationship between the input parameters and outcomes of the trained models. Examples of these techniques are SHapley Additive exPlanations (SHAP) and local interpretable model-agnostic explanations (LIME). These techniques have remained prevalent in the medical field due to their wide applicability. In contrast, model-specific techniques are based on specific model structures or architectures and use reverse engineering methods that are specific to particular model structures to explain how a given model generates prediction results. For instance, rule-based approaches have been adopted in binary tree-based modeling to elucidate the reasoning process behind final prediction outcomes. These approaches have received more attention than model-agnostic approaches due to their interpretability.

Among the model-specific explanation techniques, the use of fuzzy rules in a fuzzy inference system (FIS) has emerged as a powerful method. Its inherent capability in bridging the communication gap between humans and machines (including both machine learning (ML) and deep learning (DL) models) has been demonstrated for various medical tasks. This capability is built on two central pillars. First, fuzzy rules in an FIS enables human-like reasoning for inferring prediction results. These rules are formulated as IF–THEN statements based on expert knowledge or data and leverage fuzzy logic to provide a high level of semantic interpretability and understandability in reasoning processes that resemble human reasoning behaviors. Fuzzy rules are thus highly intuitive and comprehensible even for stakeholders outside the field of AI, such as clinicians, patients, and policymakers. The applicability of FISs has been extensively demonstrated for various medical tasks, such as the handling of sequential medical signals, (e.g., electroencephalogram (EEG) for epilepsy classification and recognition, or seizure identification [7,8,9]), and the use of tabular medical data (e.g., medical records, heart rate data, and blood test results) for the risk assessment of cardiovascular diseases [10] and the diagnosis of breast cancer [11]. Second, FISs have untapped potential in terms of integration with complex artificial neural networks (ANNs). Multiple research groups have successfully incorporated fuzzy logic theory into ANNs for medical image classification [12], COVID-19 detection [4], thyroid disease classification, and diabetes detection. Given their strengths, fuzzy rules in FISs are anticipated to gain prominence in the contemporary era of AI. This integration is expected to enhance the accuracy of complex and sophisticated prediction models while providing semantic human-like reasoning and interpretability of the prediction results.

Although reviews have been published on the interpretability of fuzzy rules in FISs, they are either irrelevant to medical disease diagnosis [13] or restricted to specific diseases [14], and do not provide a comprehensive overview of fuzzy rules for disease diagnosis. That is to say, there is a scarcity of reviews covering a broader spectrum of medical diagnosis. Moreover, the literature offers little discussion of the role of fuzzy rules in handling medical data with diverse modalities, yet such use of fuzzy rules is expected to provide valuable insights that will advance interpretable AI through the design of appropriate methodologies for the available datasets.

The overarching goals of this review are to enhance the understanding of fuzzy rules in FISs and to promote the adoption of fuzzy rules for advancing XAI in the field of medical diagnosis. In **Section 2**, we introduce the fundamental knowledge, definitions, categorizations, properties, and historical development of FISs and fuzzy rules. In **Section 3**, we present comparative analyses of fuzzy rules and other explanation methods used in handling multimodal medical data (including sequence signals, medical images, and tabular data) for disease diagnosis. In **Section 4**, we discuss appropriate scenarios for the application of fuzzy rules and provide recommendations for future research. In **Section 5**, we present the conclusions of the review.

2. Explainable artificial intelligence (XAI)

Over the past decade, the concepts, taxonomies, assessment criteria, and underlying opportunities and challenges of XAI have been discussed. In this section, we introduce the fundamental knowledge and terminologies of XAI and interpretable FISs for readers to better comprehend the subsequent discussions presented in this review. We begin by introducing the general context of interpretability, including the categorization of interpretability and the properties of explanation methods (**Section 2.1**). We then give an overview of the interpretable FIS, focusing on its historical development and fuzzy rules (**Section 2.2**).

2.1. Context of interpretability

Since 2017, there has been a notable increase in research exploring the interpretability of AI, ML, and DL models, as evinced by data from Google Trends [15] (Fig. 1). This trend suggests a growing global awareness of the necessity for transparent and interpretable AI models in various domains, including disease diagnosis.

2.1.1. Interpretability and explainability

In the field of disease diagnosis, the terms "interpretability" and "explainability" are often used interchangeably. Since the paradigm shift in the AI era, primary research has focused on designing complex and sophisticated methods, such as DL models, to obtain accurate and optimized model performance, regardless of the interpretability of the models. The complexity of the resulting models has greatly limited the possibility for key stakeholders (i.e., physicians, patients, policymakers, and even researchers) to understand how and why the AI reaches its ultimate prediction results. There is thus a lack of interaction and "trust" between humans and machines. To address this problem, the terms "interpretability" and "explainability" have emerged and been extensively discussed within the research community, although precise mathematical definitions have yet to be agreed upon. These terms highlight the importance of the comprehension and transparency of AI models and have been established as critical criteria for evaluating AI-generated models, alongside accuracy.

Although some researchers have attempted to distinguish the terms interpretability and explainability, no clear boundaries have been established, leading to these terms often being used interchangeably in the literature. We also use the two terms interchangeably throughout this review.

2.1.2. Category of interpretability

Numerous interpretable AI models and tools have been developed. Broadly, interpretability can be categorized in terms of: (1) the application scope, (2) the origin, and (3) the application phase.

Application scope. Interpretability can be classified as model-specific or model-agnostic. Model-agnostic techniques are applicable irrespective of the model structure, and work by unraveling the relationship between the input parameters and outcomes of trained models rather than delving into the inner structures of the model. SHAP and LIME are widely used model-agnostic techniques that are applied irrespective of the model type. For instance, SHAP has been used in ML with decision trees [16] and deep neural networks, such as the convolutional neural network (CNN) [17]. These techniques are commonly used in the medical field because of their wide applicability. In contrast, model-specific techniques are only suitable for specific model structures in terms of yielding explanatory prediction results. For instance, simple linear classifiers can be readily interpreted on the basis of their specific transparent model structure but are rarely applied in more advanced AI models for interpretability analyses. Model-specific techniques have attracted more attention than model-agnostic approaches owing to their peculiar interpretability.

Origin. Interpretability can be categorized as intrinsic or post-hoc. Intrinsic interpretability refers to explanations derived from constraints imposed on the principles of ML models. Intrinsic interpretability is also known as transparency and sheds light on how a model operates. Post-hoc interpretability involves the use of additional methods to analyze the trained model and it answers the question of what extra insights can be gained from the model.

Application phase. Interpretability can be categorized in terms of the timing of the application of the explanation methods. Pre-model interpretability refers to applying explanation methods before model development, in-model interpretability refers to applying explanation methods during model development, and post-model interpretability refers to applying explanation methods after the model has been developed. Pre-model interpretability relates to the graphical representation of the descriptive statistics of data, such as principal component analysis and t-distributed stochastic neighbor embedding. In-model interpretability relates to the

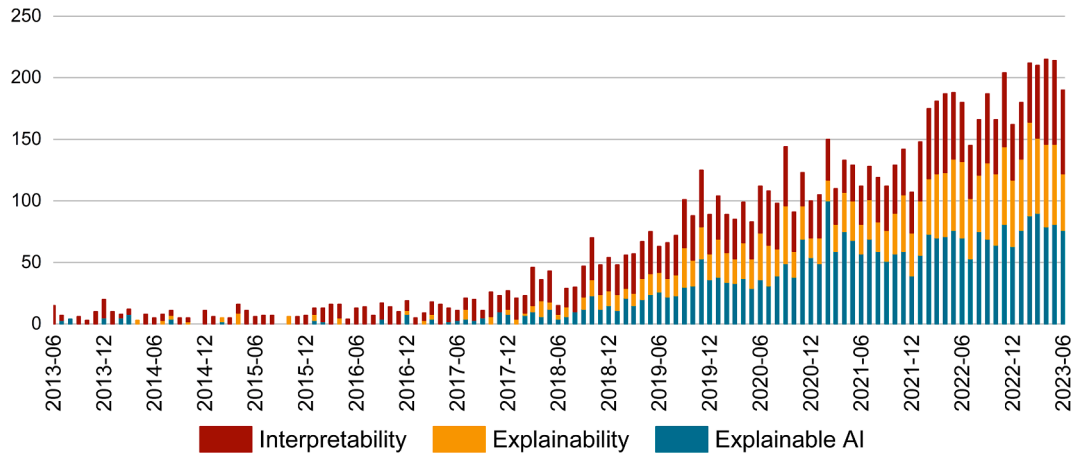

Fig. 1. Google Trends [15] data for the terms "interpretability," "explainability," and "explainable AI" worldwide over the past 10 years, revealing an increasing in search interest since 2017.

interpretability of the model itself, as well as the intrinsic interpretability described earlier. Post-model interpretability, similar to the concept of post-hoc interpretability, relates to improving interpretability after model development.

Although there are various methods of categorizing interpretability, many of them relate to each other to some degree [18]. For instance, the in-model, intrinsic model, and model-specific categories largely encompass the same XAI models that do not depend on external explanation or calculation tools. In contrast, the model structure is rarely considered in the application of the post-model, post-hoc, and model-agnostic categories.

2.1.3. *Properties of explanation methods*

Various explanation methods have been applied to enhance the interpretability of AI models. These methods have varying degrees of interpretability power based on their properties. Robnik-Sikonja et al. [19] defined four main properties of explanation methods, which were also discussed by Carvalho et al. [18]: *expressive power*, *translucency*, *portability*, and *algorithmic complexity*.

Expressive power refers to the structure or form of the output of the explanation method, including heat maps, rules, decision trees, and even natural language. *Translucency* is a measure of the extent to which the explanation method uses the parameters within the model. *Portability* is a measure of the application range of the explanation method. *Algorithmic complexity* is a measure of the computational cost of the explanation method.

Despite the introduction of the above assessment criteria, the direct quantification of explanation methods remains challenging. Generally, model-specific explanation methods have high *translucency* but low *portability*, whereas model-agnostic explanation methods have low *translucency* but high *portability* [18]. In any event, a comparative analysis of the two methods can provide deep insights into their capabilities. In **Section 4.1**, we discuss the capabilities of fuzzy rules in comparison with the capabilities two other popular explanation methods (SHAP and a heat map) applied to multimodal medical data.

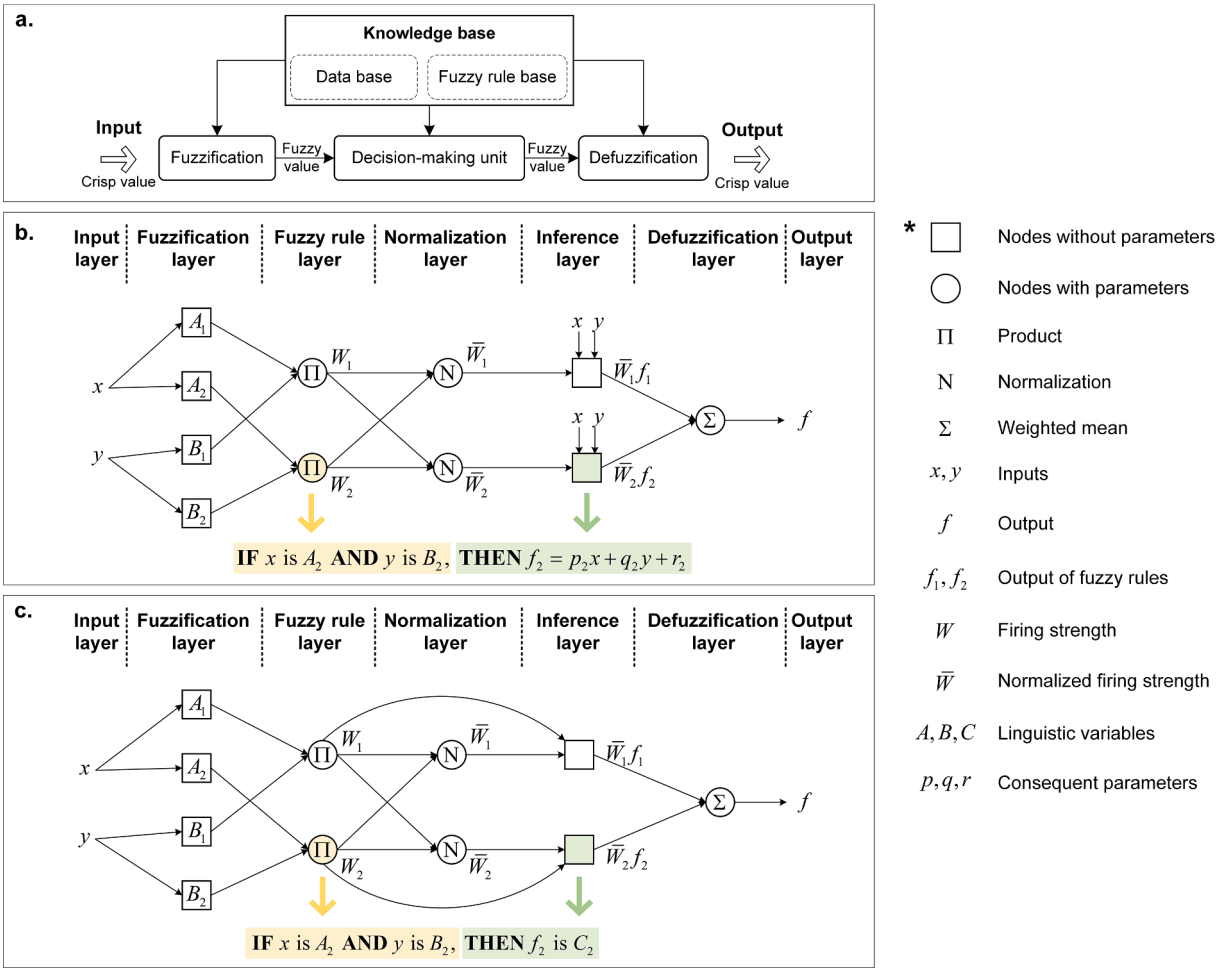

Fig. 2. Illustration of the basic structures of an FIS and an ANFIS. (a) Basic structure of an FIS. (b) Network structure of a TSK-based ANFIS. (c) Network structure of a Mamdani-based ANFIS.

2.2. Interpretable fuzzy inference system

The term FIS refers to a series of classifiers based on fuzzy set theory. A key feature of FIS is the use of crucial fuzzy rules, which provide semantic interpretability and enhance the understandability of the reasoning process. Therefore, this section focuses on the development and central principles of fuzzy rules in an FIS.

2.2.1. Development of FISs

Concerted efforts have been made to develop interpretable FISs. An FIS is governed by fuzzy rules, which were originally derived from Zadeh's theory of fuzzy sets proposed in 1965 [20]. Fuzzy sets are used to represent the degree of membership in the form of a real-value rather than a true and false statement of an object's characteristics, and they serve as a bridge between computer-friendly crisp numbers and human-friendly semantic expressions.

Fuzzy sets have developed into three types according to how they handle uncertainty. The type-1 (T1) fuzzy set, initially introduced by Zadeh to model linguistic uncertainty, has two popular modelling methods: the Mamdani type proposed in 1977 [21], and the Takagi-Sugeno-Kang (TSK) type proposed in 1985 [22]. These methods were later introduced to neural networks in research on the adaptive neural network based fuzzy inference system (ANFIS) conducted by Jang et al. in 1993 [23]. This integration allowed FISs to be driven by data instead of relying solely on experts' inputs. Although the membership grades of T1 fuzzy set are crisp value, there are scenarios with uncertain deviations in grades of membership. For instance, different physicians may have different interpretations of the linguistic term "*serious*" in the case of disease assessment.

The type-2 fuzzy set (T2 FS) was introduced to extend the modeling capacity of uncertainties by fuzzifying membership grades. Interval type-2 fuzzy logic controllers have attracted much research interests, and it has been shown that they are better than T1 fuzzy logic controllers at handling uncertainties. For instance, Wu et al. conducted extensive research in this field and recently compared T1 and interval type-2 fuzzy systems [24].

The type-3 fuzzy set (T3 FS) further extends the concept of the T2 FS by incorporating the notion of the footprint of uncertainty, allowing varying degrees of uncertainty within the membership functions. Castillo's team proposed multiple T3 FS based models such as the hybrid hierarchical neural network classification and prediction model with interval type-3 fuzzy aggregation and an ensemble model of the T3 FS and neural networks for COVID-19 time series prediction [25,26]. In addition, Zadeh investigated the potential of the T2 FS and further generalized the above concepts to type-*n* fuzzy sets [27]. Although there are challenges to revealing the following *n*-3 types of fuzzy set due to the rapidly growth of complexity, relevant research is still active. The FIS considered in the following sections of this paper is a T1 FS based inference system, as it has the most extensive applications.

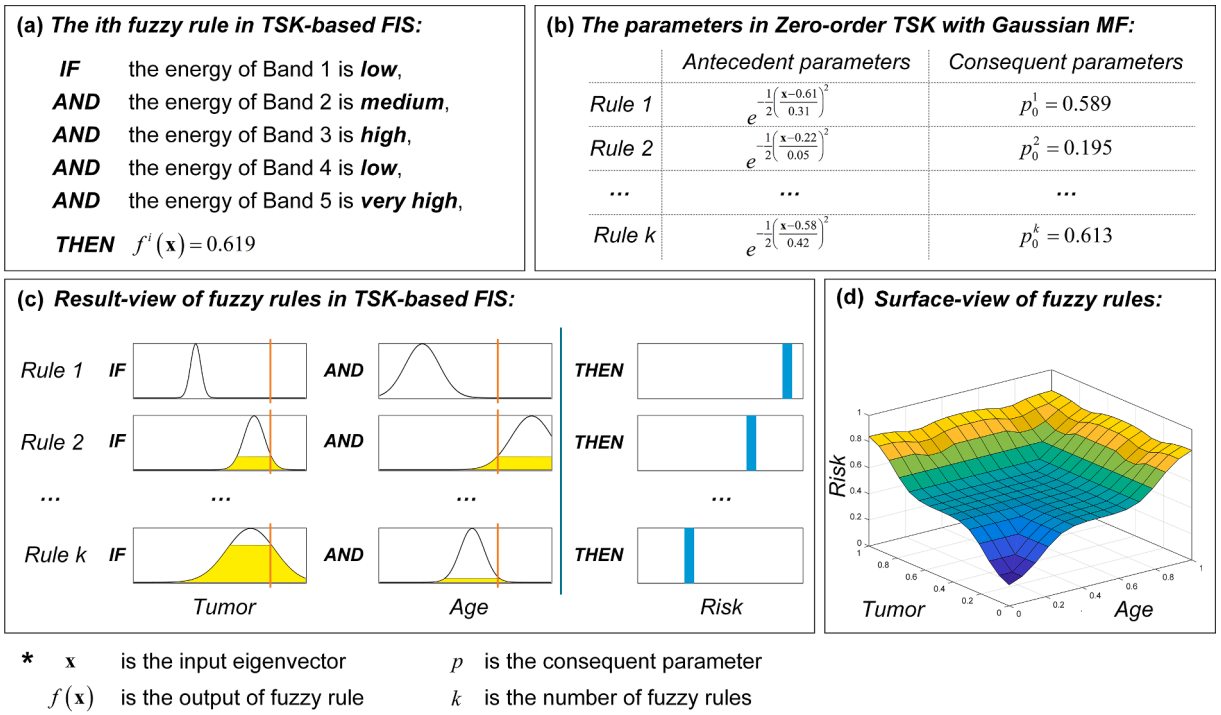

Fig. 3. Presentations of fuzzy rules. (a) An instance of fuzzy rules in the form of IF–THEN generated by the TSK-based FIS. (b) An instance of parameters in the zero-order TSK with a Gaussian membership function. (c) An instance of a result-view of fuzzy rules. (d) An instance of a surface-view of fuzzy rules. The IF–THEN form and the visualized presentation of fuzzy rules are respectively the most understandable presentation for end-users and researchers in the fuzzy logic field.

Besides the theoretical investigation of fuzzy sets, numerous studies on FISs have focused on optimizing the solution and calculation of parameter sets in the antecedent and consequent networks. This has been done using various strategies such as combining ANNs and fuzzy logic, adopting genetic-based algorithms, adopting hierarchical models [4,8] around the 2020 s, and adopting other fuzzy logic based techniques. Notably, the central pillar of the semantic interpretability of fuzzy rules in FISs has remained consistent throughout the development of FISs. FISs are thus known as fuzzy-rule-based systems [23].

2.2.2. Fuzzy rules in FISs

Each FIS (including its evolved variants) has five primary components, namely, a database, fuzzy rule base, fuzzification component, fuzzy inference component, and defuzzification component (Fig. 2**a**). The database plays a crucial role in determining the membership functions of fuzzy sets, which are used in fuzzification and defuzzification. It is noted that the membership function requires an adequate volume of sample data, or a direct definition given by experts at the initial stage. During fuzzification, crisp inputs are transformed into membership values, representing the degree of matching with linguistic terms in fuzzy sets. Conversely, the defuzzification step converts the inferred fuzzy results back into crisp outputs. The fuzzy rule base is elaborated in the following.

The fuzzy rule base, a critical component of an FIS, comprises logical fuzzy rules. a fuzzy rule is an effective representation of knowledge provided by an expert or data [27]. It has been shown that a system designed using such represented knowledge performs comparably to a system designed by experts and performs slightly better in terms of completeness of knowledge representation than a system designed using non-monotonic logic. For plain understanding, the *if* part in a fuzzy rule is a *condition*, and the *then* part is a *consequent*, and the *consequent* can thus be inferred from the *conditions*. Imagine a scenario, in which a person feels dizzy at home and does not have a thermometer, she/he habitually touches her/his forehead and senses her/his underlying body temperature to estimate whether she/he is having a fever. The inference logic behind this action is that *if* a person feels a "little dizzy" *and* the body temperature is "very high", *then* she/he "most likely" has a fever, *or if* a person feels a "little dizzy" *and* the body temperature is "low", *then* she/he may not have a fever. This logic of linguistic representation using the terms (*if*, *and*, *then* and *or*) has been seamlessly incorporated into the fuzzy rules of FISs (Fig. 3).

For instance, Casalino et al. [10] used a neuro-fuzzy model to evaluate the risk of cardiovascular disease based on four clinical features, namely, heart rate, breathing rate, blood oxygen saturation, and lip color. The neuro-fuzzy model enables the transformation of numerical crisp values into linguistic terms and uses fuzzy IF–THEN rules. In their study, the heart rate domain ranged from 10 to 180 bpm, and was manually fuzzified into three fuzzy sets representing the linguistic terms *Bradycardia*, *Normal*, and *Tachycardia*. The breathing rate domain ranged from 0 to 80 and was fuzzified into three fuzzy sets corresponding to *Bradypnea*, *Normal*, and *Tachypnea*. Blood oxygen saturation had a domain range of 75 to 100 and was fuzzified into the fuzzy sets *Critical*, *Low*, and *Normal*. Lip color had a domain range of 0 to 14 and was fuzzified into the fuzzy sets *Regular*, *Altered*, and *Purplish*. The output variable, the risk level, was associated with the linguistic terms *Low*, *Medium*, *High*, and *VeryHigh*. The relationship between the input variables and the potential cardiovascular risk was established through fuzzy rules according to.

IF (heart rate is *Bradycardia*) and (breathing rate is *Bradypnea*) and (blood oxygen saturation is *Low*) and (lips color is *Altered*), THEN (risk is *High*).

We see that both the condition part (i.e., the IF part) and the consequent part (i.e., the THEN part) comprise input or output variables and their corresponding linguistic terms. The definition and number of linguistic terms for each variable can be determined by experts or optimized using data-driven approaches (as explained in the following). In theory, the total number of fuzzy rules is the product of the linguistic terms of all variables, including both input and output variables. However, not all fuzzy rules are reasonable or efficient. It is thus crucial to effectively generate or select useful fuzzy rules that minimize assessment bias.

With such a linguistic representation, the fuzzy knowledge or experience of experts, which may not be easily quantifiable, can now be represented in the form of a human-like reasoning process and transferred or stored conveniently. In addition, this reasoning process is logical, non-linear, and fuzzy (owing to the use of fuzzy linguistic expressions such as "little", "low", "very high" and "most likely" in the descriptions), which implies that the process can handle a complex, abstract, non-linear, and fuzzy task. Furthermore, the fuzzy rules can be understood directly by various stakeholders in the medical domain, including physicians and patients, which increases the degree of transparency and approbation of the FIS-aided decision-making process. As a result, ML algorithms with fuzzy rules have great semantic interpretability in both the inference process and inferenced results. As the complexity of data and FISs grow rapidly with time, the adaptive and effective selection of fuzzy rules is expected to become a focus of FIS research.

Substantial efforts have been made to improve the methods of fuzzy rule selection and parameter resolution. ANFISs play a fundamental role in the construction of a set of fuzzy IF–THEN rules with appropriate membership functions to generate stipulated input–output pairs [23]. In contrast with the modularized flow depicted in Fig. 2**a**, an ANFIS combines fuzzy set theory with neural networks to form a hybrid neuro-fuzzy network. According to the model structure depicted in Fig. 2**b** and Fig. 2**c**, the fuzzy rules in an ANFIS are based on antecedent and consequent parameters. The optimization of fuzzy rules is thus equivalent to the tuning of antecedent and consequent parameters.

A key advantage of ANFISs (or neuro-fuzzy networks) is that the parameters can be optimized by adopting a neural-learning strategy (e.g., back propagation, or gradient descent strategy) and neuro-fuzzy networks can thus extract effective and human-like knowledge representations adaptively from data, resulting in outstanding interpretability. Sanz et al. [28] proposed a fuzzy association rule-based classifier that adopts a global fuzzy rule selection of all classes using the Apriori algorithm. Experimental results indicated that the minority class takes a larger number of fuzzy rules, which suggests high classification performance for imbalance data. The fuzzy rules used in such research offer semantic interpretability for diagnosis results and have been optimized to enhance training efficiency and understandability for end users.

In addition to the fuzzy rule logic, the methods of fuzzy rule presentation play a crucial role in enhancing the interpretability of

models. There are three ways to present fuzzy rules: adopting IF–THEN statements (Fig. 3**a**), making a list of antecedent and consequent parameters (Fig. 3**b**) and visualization (Fig. 3**c** and Fig. 3**d**). For most parameters in black-box models, such as the weights in a deep neural network, it is difficult to understand the logic of the model by directly listing the parameters in a table. However, this becomes understandable when an FIS is used. As mentioned earlier, the fuzzy rules are determined by the antecedent and consequent parameters in an ANFIS. Hence, the presentation of antecedent and consequent parameters is equivalent to the presentation of fuzzy rules. As a result, the presentation of antecedent and consequent parameters can reveal the reasoning process of an ANFIS. The IF–THEN form is the most logistic linguistic and human-like presentation of the reasoning process, and is intuitive for users who are unfamiliar with fuzzy set theory. In contrast, the visualized presentation of fuzzy rules is mainly associated with a result-view (Fig. 3**c**) or surface-view (Fig. 3**d**) of the corresponding rule sets. Specifically, the result view is useful for obtaining crisp output values and evaluating the performance of an FIS whereas the surface view provides a visual understanding of the fuzzy rules and their effects on the output according to different input combinations. Overall, the IF THEN form, and visualized presentation of fuzzy rules are respectively the most understandable presentations for end users and researchers in the field of fuzzy logic. The adopting of these presentations in previous studies is further discussed in **Section 3**.

3. XAI applications in disease diagnosis

Medical disease diagnosis has been recognized as a challenging and intricate task for healthcare professionals as the process of diagnosing patients requires physicians to carefully consider multiple factors and circumstances alongside medical evidence. However, disease diagnosis is prone to errors owing to its vague and complex nature when considering all factors, leading to great uncertainty in the process of disease diagnosis since different patients may have varying levels of confirmation for different diseases. In addition, it is critical in computer-assisted diagnosis modeling to clarify the process of reasoning out how to deal with all the factors. As shown in Fig. 4, the decision made by an understandable model using the same data as used by black box is more transparent and creditable for doctors, reducing the additional uncertainty introduced by the computer-aided diagnosis model and the difficulty in recognizing hidden relationship across factors.

Extensive studies have been carried out on XAI to reveal the predictive power of diverse data, including sequence signals from medical sensors (e.g., EEGs and electrocardiograms (ECGs)), medical images (e.g., chest X-ray (CXR), computed tomography (CT), magnetic resonance imaging (MRI), ultrasonography, and elastography images), and tabular data (e.g., medical health records, heart rate data, and blood test results). Fig. 5 provides an overview of XAI techniques commonly used in computer-aided disease diagnosis obtained from 50 relevant publications. The left panel gives different types of medical data. The right panel gives a series of explainable methods in XAI, including the use of fuzzy rules, a rule base, SHAP, LIME, a heat map, and other specialized but representative methods such as mathematical method, as described in Table 1. The middle panel of Fig. 5 presents a parallel sets plot generated from a summary of the 50 relevant publications, showcasing the proportions of research utilizing each of the explanation methods on different modalities of medical data. The plot shows that each explanation method has its advantages and is applicable to different proportions of diverse disease diagnosis scenarios with various types of data.

This section is organized into three subsections according to the type of multimodal medical data, namely, sequence data, medical image data, and tabular data, and presents the actual application modes of fuzzy rules and other explanation methods to facilitate a comprehensive comparison.

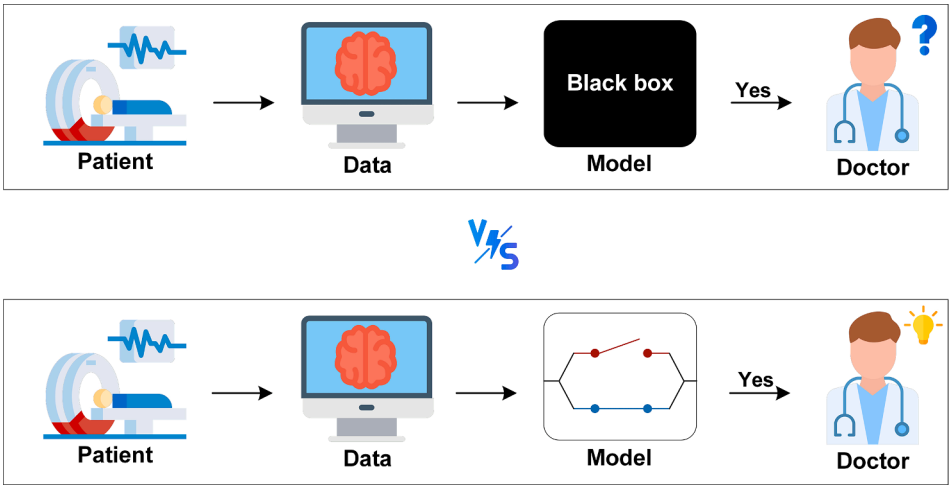

Fig. 4. Illustration of a computer-aided diagnosis system. Compared with a decision made using a black box model, the decision made using an interpretable model is more understandable for doctors.

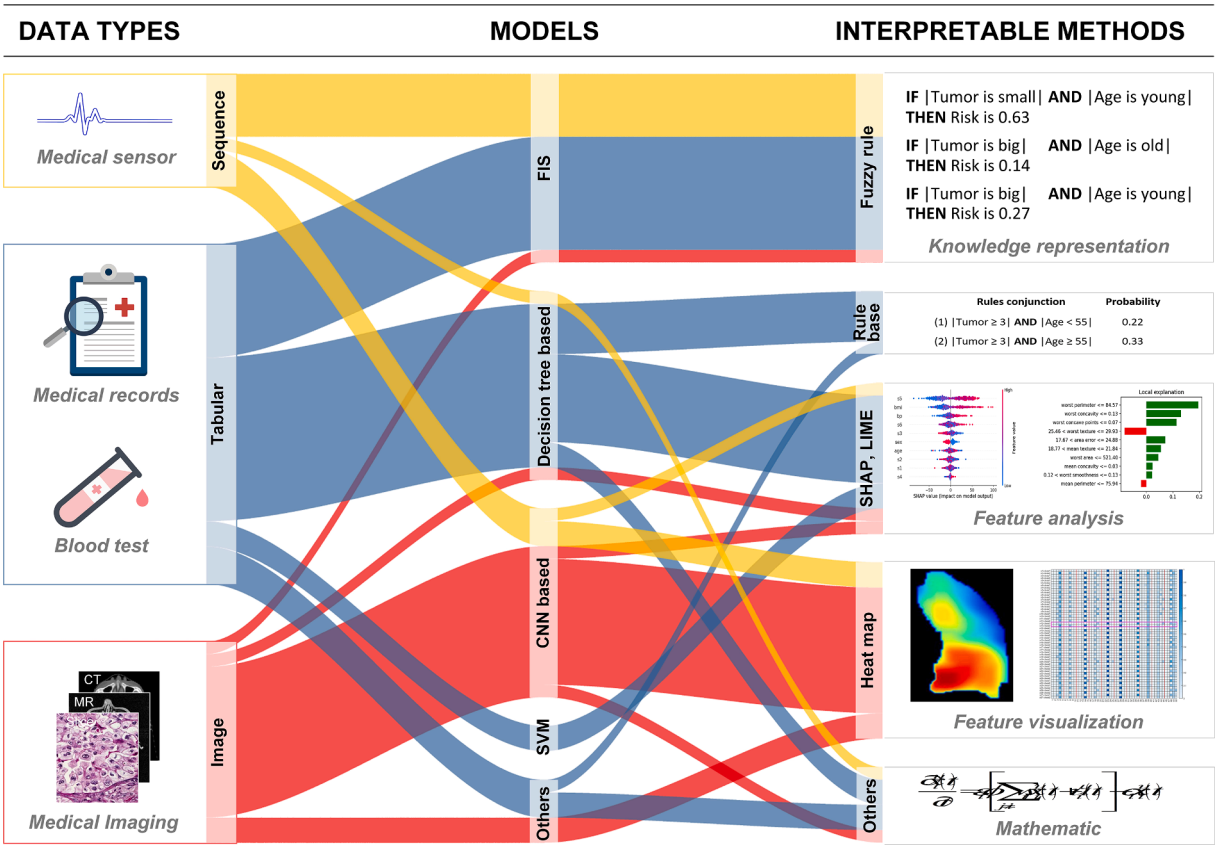

Fig. 5. Overview of XAI techniques commonly used in computer-aided disease diagnosis scenarios, based on the analysis of 50 relevant literature sources. The line width in the parallel set figure is determined by the number of relevant sources. The examples of data types and interpretability methods mentioned are not limited to those presented on the sides of the parallel set plot.

Table 1
Brief description of other interpretability methods.

Interpretability methods	Description
Rule base	Rule base can be generated by decision tree-based model, including Decision Tree, Random Forest, XGBoost etc. These rules describe the conditions that lead to specific decisions, making the model easily interpretable. These models provide insights into the importance of each feature in the decision-making process, and it can also be well integrated with the SHAP principle. In the view of this, they are often used together. However, its biggest difference from fuzzy rules is that it does not include fuzzy linguistic variables, instead it relies entirely on crisp values, as shown in the panel of Fuzzy rule and Rule base in Fig. 5.
SHAP	SHAP is a game-theoretic approach that provides a unified framework for explaining the output of any machine learning model. It is based on concepts from cooperative game theory, specifically Shapley values, which allocate the contribution of each feature toward the prediction outcome. SHAP values represent the impact of each feature on the predicted outcome for a specific instance. These values enable us to understand the importance and influence of features in the model's output. An example is shown in the feature analysis panel of Fig. 5.
LIME	LIME is a technique for explaining the predictions of any black-box machine learning model. It aims to provide local and interpretable explanations by approximating the behavior of the model around specific instances. By examining the coefficients of the approximated model, LIME identifies which features were the most influential in influencing the prediction for that particular instance. These explanations help users understand the model's decision-making process at an individual instance level, thus increasing transparency and trustworthiness. An example is shown in the feature analysis panel of Fig. 5.
Heat map	A heatmap is a visualization technique used to represent the importance or relevance of features in a model. The color gradient in the heatmap helps identify patterns and correlations between features and instances. A higher intensity or a distinct color in a cell or pixel signifies a stronger influence of that feature on the model's decision, while lower intensity or a different color suggests a relatively lesser impact, as shown in the heat map panel in Fig. 5.

3.1. Sequence data

Sequence data, such as EEG and ECG data, are harvested from medical sensors and widely utilized for diagnosing diseases related to the heart, brain, and mental health owing to their non-invasive recording and suitability for longitudinal disease monitoring. This subsection outlines the application of sequence data in disease diagnoses, focusing on the explanation methods of fuzzy rule, SHAP and LIME based methods developed to handle sequence data.

3.1.1. Fuzzy rule approaches

Directly handling temporal features in sequence data can be challenging for fuzzy rule-based algorithms. Therefore, feature extraction methods are commonly used to pre-process raw sequence data. In the case of multi-view EEG data, Zhang et al. [9] used various feature extraction methods, such as wavelet packet decomposition (WPD), short time Fourier transform (STFT), and kernel principal component analysis (KPCA), to construct training and testing datasets. They used a specific model called the deep view-reduction TSK fuzzy system to determine the weight of each view and automatically reduce the effect of weak views, achieving a testing accuracy of 84.89 % in the classification of epilepsy. Moreover, they used fuzzy rules in the specific TSK fuzzy neural network to enhance the interpretability of prediction results, enabling researchers to understand the meaning of parameters during the reasoning process and providing linguistic interpretability support for the results. Similarly, Xue et al. [8] proposed a novel deep ladder-type TSK fuzzy classifier for epilepsy recognition using EEG signals. They also used WPD, STFT, and KPCA for feature extraction in dataset construction and summarized the antecedent and consequent parameters, as well as fuzzy rules, to enhance the understandability of the results. Other studies have developed specific TSK-based FIS models and used fuzzy rules to provide transparency in the reasoning process [7,29].

Different representations of fuzzy rules, each having its advantages and uses, have been used in medical diagnosis based on sequence data. Xue et al. [8] used all three representations discussed in **Section 2.2.2**: the parameters, IF–THEN, and visualization representations. Each of these three representations of fuzzy rules has its advantages, but the simultaneous adopting of all three maximizes the ability to comprehend the reasoning process according to fuzzy rules and ensures a higher level of interpretability for the widest audience. Most previous studies [9,30] have used both the parameter and IF–THEN representations of fuzzy rules as the represented knowledge. Fuzzy rules have been used to make medical diagnosis from sequence data to realize high interpretability and to generate knowledge representations. Indeed, there have been researchers who have taken the interpretability of fuzzy rules as a default advantage and only briefly mentioned it in their works [29].

3.1.2. Conventional approaches

Besides the use of fuzzy rules, studies have used model-agnostic explanation methods to enhance the interpretability of results. For instance, Smith et al. [31] explored a combination of variational mode decomposition and the Hilbert transform to extract hidden information from EEG signals. They used five traditional classifiers and four model-agnostic explanation methods (i.e., the use of LIME, SHAP, a partial dependence plot, and Morris sensitivity) to detect attention deficit hyperactivity disorder, achieving an accuracy of 99.81 %. However, the extracted features were mainly statistical measures, such as standard deviations and mean values, and the

Table 2
Literatures related to the interpretability of a disease diagnosis made using sequence data.

No.	Author	Disease	Data	Preprocessing	Model	ACC	Explanation method
1	Tao et al. [30]	Epilepsy	EEG	WPD, STFT, KPCA	Domain adaptation learning, semi-supervised learning, and a fuzzy system	96.8 %	Fuzzy rules
2	Li et al. [7]	Epilepsy	EEG	WPD, STFT, KPCA	Multi-view TSK fuzzy system	98.87 %	Fuzzy rules
3	Zhang et al. [9]	Epilepsy	EEG	WPD, STFT, KPCA	Deep View-reduction TSK fuzzy system	84.89 %	Fuzzy rules
4	Gu et al. [29]	Seizure	EEG	WPD, STFT, KPCA	Multiple-source transfer learning-based TSK	97.1 %	Fuzzy rules
5	Xue et al. [8]	Seizures	EEG	STFT, KPCA	Deep ladder-type TSK fuzzy classifier	88 %	Fuzzy rules
6	Rahman et al. [33]	Cardiac arrhythmia	ECG	Butterworth bandpass filter	FIS	~100 %	Fuzzy rules
7	Smith et al. [31]	Attention deficit hyperactivity disorder	EEG	Variational mode decomposition, Hilbert transform	DT, Nearest Neighbor, Medium NN, Random Forest and Explainable Boosting Machine	99.81 %	Heat map (SHAP, LIME, PDP based)
8	Zhang et al. [34]	Cardiac arrhythmia	ECG	N/A	Deep neural network	96.6 %	Heat map
9	Rashed et al. [32]	Cardiovascular	ECG	N/A	VGG16-based CNN	99.1 %	Heat map (SHAP based)
10	Agrawal et al. [35]	Changes in the ECG of the post-COVID	ECG	HRV-analysis module	Convolutional Neural Network without HRV	100 %	Heat map (SHAP based)
11	Rashed et al. [36]	Epileptic seizures	EEG	Signal-to-image conversion methods	FT-VGG16 classifier	99.21 %	Heat map (SHAP based)

explanation methods enhanced the understandability of the statistical features instead of the raw sequence data. In contrast, CNN related models handle sequence data without additional feature extraction algorithms. For instance, Rashed et al. [32] proposed a VGG16-based CNN adopting the time–frequency representation of temporal ECG signals to diagnose cardiovascular conditions. In addition, they applied SHAP values in the time domain of frequency mapped ECG to highlight key features and its range in the original domain, confirming that the model learnt useful information from effective regions. Numerous studies have performed diagnostic tasks with or without the preprocessing of sequence data and have explored the interpretability of the models (Table 2).

Compared with heat map explanation methods, the current use of fuzzy rules may be suboptimal for end-users to understand, even though the rules elucidate the reasoning process. As depicted in Fig. 6, a heat map directly illustrates the importance or contribution of each point in the raw data, whereas fuzzy rules and SHAP plots mainly explain abstract statistical features that are extracted from sequence data. Therefore, in scenarios of disease diagnosis from sequence data, deep neural networks with heat map explanation methods are recommended to improve interpretability.

3.2. Medical image data

Medical imaging techniques, such as CXR analysis, CT, MRI, ultrasonography, and elastography, are commonly used in disease diagnosis as they provide valuable tissue-related information for clinicians in a non-invasive manner. These types of data can be directly used by CNN based models, and in FISs after feature extraction. The two applications are further discussed in this subsection.

3.2.1. Feature extraction

Two main approaches for feature extraction from medical images are used in disease diagnosis: the use of specific filters and the use of deep neural networks.

Specific filters are used to extract descriptors that capture specific information from medical images, such as texture features and contour features. Radiomics feature extraction methods are commonly used to extract quantitative features from medical images, including but not limited to CXR, CT, and MRI images. Zhang et al. [4], for example, extracted radiomic features, including the intensity, shape, texture, and wavelet features, from segmented CXR images using U-net. Subsequently, these features were used for COVID-19 detection using a novel TSK fuzzy classifier with a soft label-driven mechanism. They used fuzzy rules in their model to give transparency to the reasoning process by adopting both IF–THEN and parameter representations (as elaborated in **Section 2.2.2**). The

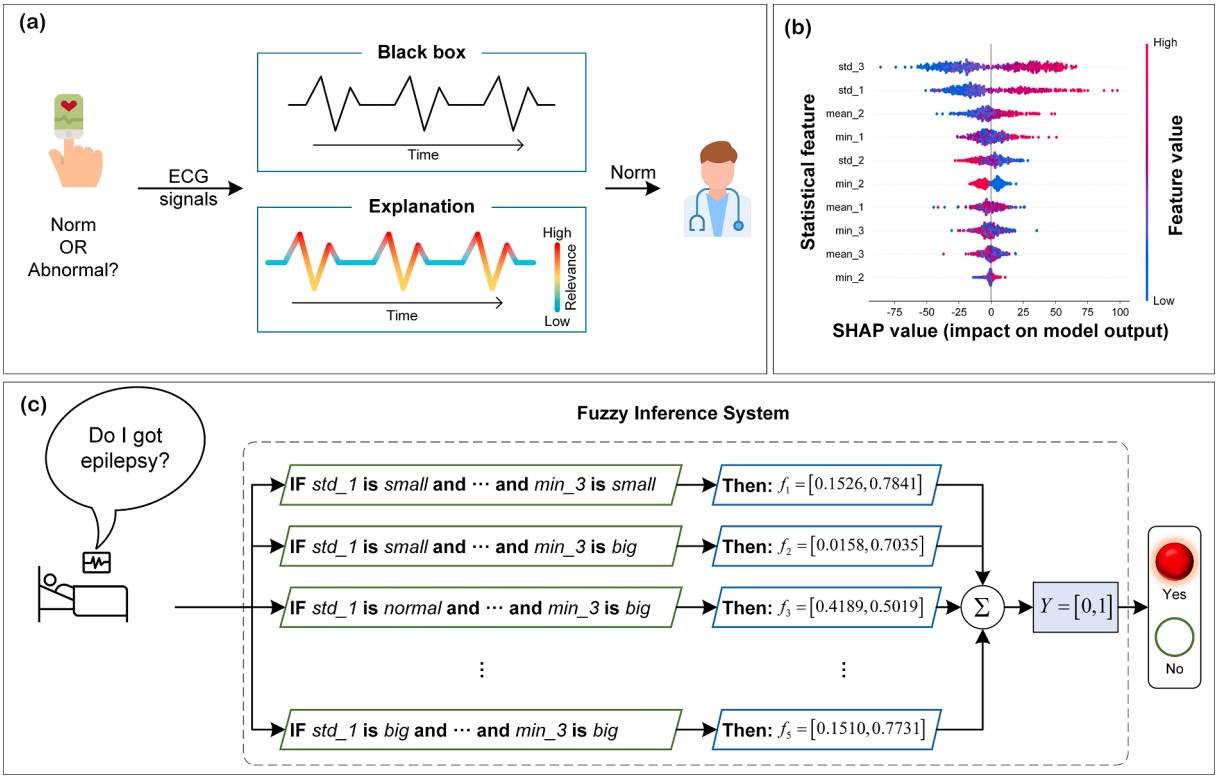

Fig. 6. Representative examples of explanation methods used in disease diagnosis with sequence data. (a) A heat map used as an explanation to highlight fragments with diverse relevance in ECG data. (b) A SHAP plot of statistical features calculated from ECG sequence data for the analysis of the impact of features on the model output. (c) A method of applying fuzzy rules to improve the interpretability of the reasoning process and results for epilepsy recognition based on statistical features calculated from EEG sequence data.

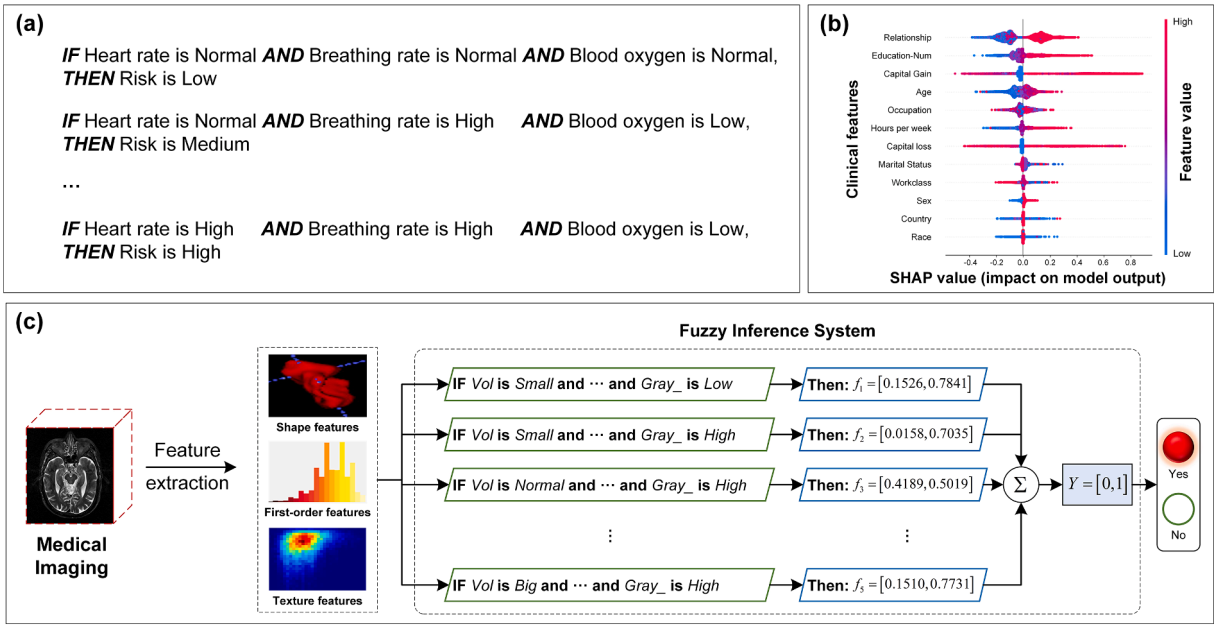

Fig. 7. Representative examples of explanation methods used in disease diagnosis with medical image and tabular data. (a) Fuzzy rules used in an FIS model with clinical features to assess the risk of cardiovascular disease. (b) SHAP plot that showing importance distribution of clinical features. (c) Use of an FIS in disease detection based on radiomic features (including Shape features, First-order features, and texture features) extracted from medical images, with fuzzy rules in the form of IF–THEN providing interpretability of the reasoning process.

variables in fuzzy rules are features extracted from medical images and are as abstract as those in scenarios involving extracted sequence data (Fig. 7).

Deep neural networks have been used to improve the interpretability of the feature extraction and reasoning process. Ma et al. [37] extracted mammography and ultrasound features based on the BI-RADS manual from ultrasound images and used SHAP to determine the contributions of the features, thereby offering a degree of model interpretability. Explanation methods used in scenarios involving eigenvectors extracted from image data are slightly superior to those used for sequence data because the extracted features retain spatial information from the raw image, which can be understood by users. However, interpretability in disease diagnosis using eigenvectors extracted from image data remains limited.

3.2.2. *Convolutional neural networks*

Currently used visualization maps include feature maps, attention maps, gradient-based saliency maps, and other specific heat maps. These maps offer varying levels of clarity regarding the calculation process and pixel contributions in CNN-based models. These visualization maps enhance the trust of clinicians in the results of computer-aided diagnosis systems. Previous reviews have summarized the explanation methods for models using image-based data, and these techniques can also be applied in disease diagnoses using medical image data [38]. Gradient-weighted class activation mapping (Grad-CAM), a specific type of attention mapping, is a technique widely used to provide visual explanations for decisions made using CNN-based models. Panwar et al. [39] used Grad-CAM to visualize regions contributing to the detection of COVID-19 in X-ray and CT images, thereby enhancing model interpretability. Grad-CAM remains a representative and useful model-agnostic explanation tool for improving the interpretability of CNN-based models.

Traditionally, the direct application of FISs to image data has been rare and challenging. Concerted efforts have enabled the incorporation of fuzzy logic theory into CNNs in various ways. Sharma et al. [12] explored the use of fuzzy-based pooling in CNNs for image classification. To tackle the uncertainty in the extraction of useful information due to the unclear intuition of conventional pooling methods, type-2 fuzzy logic was utilized to identify the dominant convolved features of the pixels within a window to be pooled, and type-1 fuzzy logic with the weighted average of these dominant features was adopted to reduce the spatial size. Their proposed method, achieving an accuracy of 94.4 % for fuzzy pooling, 94 % for average pooling, and 88.4 % for max pooling, outperformed conventional pooling methods on challenging image datasets, such as MNIST. Recently, Ruixuan et al. [40] proposed a Reference-guided Fuzzy Integral Generative Adversarial Network to nonlinearly fuse the textural and structural features of ultrasound images into the convolutional layer of a generative adversarial network. Wang et al. [41] converted an input medical image into a fuzzy domain and processed the uncertainty of the pixels utilizing proposed fuzzy rules, and fused the outputs of the fuzzy rule layer and convolution layer. They achieved a high reconstruction performance for high-resolution medical images. It is widely agreed that even with a powerful CNN, there is uncertainty in the network structure, which can be addressed using fuzzy logic. Owing to the

inherent strength of CNN, the improvements that result from the introduction of fuzzy logic may not be large. However, in the medical field (a highly interactive and high-risk domain), it is crucial to thoroughly address this uncertainty. Although this technique has commonly been tested on MNIST data and is still in its infancy, it has provided evidence that fuzzy logic can handle uncertainties in image classification, such as noise, perturbation, and multi-center variations and has application potential in the field of medical diagnosis.

3.3. Tabular data

Various tabular data, such as medical records, blood test reports, and radiomic features, can be used in disease diagnosis. These data are highly convenient for use in ML models and contribute to the wide use of explanation methods such as fuzzy rules and SHAP.

Clinical records (including but not limited to records of patient characteristics, clinical history, and heart rate) are directly obtained from electronic health records and can be used in ML algorithms after normalization. Fuzzy rules are commonly used in explanation methods for this scenario. For example, Casalino et al. [10] used tabular data on four clinical features (i.e., heart rate, breathing rate, blood oxygen saturation, and color of the lips) to assess the risk of cardiovascular diseases using an ANFIS. They achieved 91 % accuracy and used 80 fuzzy rules in IF–THEN form (as elaborated in **Section 2.2.2**), which are largely understandable. Algehyne et al. [11] incorporated tumor features into a five-layered FIS for breast cancer diagnosis and stored IF–THEN fuzzy rules in a fuzzy rule base for inference of the output. Several other studies have used fuzzy rules in the context of tabular data from clinical records to enhance the understandability of the reasoning process and results (Table 3). When fuzzy rules incorporate clinical records, they represent logical relationships in IF–THEN form rather than quantitative rankings of clinical or tumor features. This characteristic of fuzzy rules makes the reasoning process more comprehensible to end users owing to its human-like nature (Fig. 7). Therefore, we strongly recommend the application of an FIS with interpretable fuzzy rules in disease diagnosis involving tabular-based data such as the data of clinical records.

4. Discussion

The aforementioned literature indicates that different explanation methods exhibit varying performance across diverse application scenarios. In this section, we first discuss the advantages of using FISs in disease diagnosis applications. We than compare the performance of fuzzy rules with that of two other explanation methods (i.e., the adoption of SHAP and a heat map) considering the four key properties described in **Section 2.1.3**. Finally, we suggest several future research directions, including the application of novel FISs, enhancement of interpretability in FISs, and reduction of the complexity of fuzzy rules, aimed at advancing the use of XAI in disease diagnosis.

4.1. FISs in disease diagnosis

In this subsection, we discuss the strengths of FISs in disease diagnosis scenarios and present a comparative analysis of fuzzy rules and two other explanation methods (i.e., the adoption of SHAP and a heat map).

4.1.1. Advantages of fuzzy rules in disease diagnosis

FISs use fuzzy logic to generate prediction results and thus offer a high level of semantic interpretability for both the reasoning process and results. Although fuzzy rules can only be used in conjunction with an FIS, the FIS has great adaptability and generalization

Table 3
Literatures on the interpretability of disease diagnosis with tabular data.

No.	Author	Disease	Data	Feature No.	Model	Best ACC	Explanation method
1	Casalino et al. [10]	Cardiovascular risk	Clinical features	4	ANFIS	91 %	Fuzzy rules
2	Bai et al. [42]	Breast cancer	Clinical features, tumor features	10	A broad learning-based dynamic FIS	96.74 %	Fuzzy rules
3	Thani et al. [43]	Breast cancer	Clinical features	10	FIS	90.3 %	Fuzzy rules
4	Algehyne et al. [11]	Breast cancer	Tumor features	30	FIS	99.33 %	Fuzzy rules
5	Murugesan et al. [44]	Chronic kidney disease	Clinical features	7	FIS	96 %	Fuzzy rules
6	Dong et al. [45]	3-year risk of diabetic kidney	Clinical features	46	LightGBM, XGBoost, Adaboost, ANN, DT, SVM, LR	0.815 (AUC)	SHAP
7	Liu et al. [46]	Cardiovascular	Clinical features	11	SVM, KNN, LR, RF, ET, GBDT, XGBoost, LightGBM, CatBoost, MLP	89.86 %	SHAP
8	Hakkoum et al. [14]	Breast cancer	Tumor features	30	Multilayer perceptron, and Radial Basis Function Network	96.63 %	LIME, PDP
9	Pal et al. [47]	Lung cancer	Tumor features	10	SVM, KNN, GBM, XGBoost, RFC and feed forward neural network	94.2 %	SHAP

capabilities. In particular, ANFISs [23] excels at adaptively tuning fuzzy rules by optimizing antecedent and consequent parameters based on large datasets, as detailed in **Section 2.2**. Consequently, in disease diagnosis applications, the explanation method of fuzzy rules retains the advantage of being highly comprehensible without requiring specific task. Even in the case of sequence data, such as EEG and ECG data, and medical images, ANFISs can handle the classification task when the data have been transformed to tabular or eigenvector data in advance, as presented in **Section 3.1** and **Section 3.2**. Fuzzy rules, especially when expressed in the form of IF–THEN statements, present a reasoning process that closely resembles human reasoning. They are thus highly intuitive and comprehensible even for individuals outside of the field of fuzzy logic, such as clinicians and patients. This is particularly evident in disease diagnosis scenarios involving clinical records, as discussed in **Section 3.3** and presented in Fig. 7.

This resemblance to human reasoning means that fuzzy rules are highly suitable for applications in the medical domain (Fig. 8). For instance, in a scenario of sequence data, Xue et al. [8] presented the fuzzy rules in parameter, IF–THEN, and visualization representations to show the interpretability of a TSK-based ANFIS in the recognition of epilepsy from EEG signals. Even when sequence signals had been transformed to abstract features, the role of each parameter in the human-like reasoning process and the generated prior knowledge representation remained understandable (Fig. 6). For medical imagery, the FIS can be utilized as an interpretable classification method after feature extraction. Zhang et al. [4] extracted radiomic features, including intensity, shape, texture, and wavelet features, from segmented CXR images using U-Net as the input of a novel TSK neural network for COVID-19 detection, and presented fuzzy rules in parameter and IF–THEN form for the interpretation of the reasoning process (Fig. 7**c**). Similarly, Algehyne et al. [11] adopted an FIS based on clinical features, such as the radius, texture, area, and smoothness, for breast cancer detection; their model achieved 99.33 % accuracy and presented an interpretable knowledge base of the predictors (Fig. 7**a**). However, it is worth noting that although fuzzy rules have strong interpretability, the current variants of fuzzy rules may still encounter limitations when compared with model-agnostic explanation methods, such as SHAP and LIME. These limitations are explored in the following subsection.

4.1.2. Comparative analysis

The use of fuzzy rules, as a model-specific explanation method, exhibit varying interpretability performance for different types of multimodal medical data in disease diagnosis scenarios. This discussion centers around the three representative modalities of medical data (i.e., sequence signals from medical sensors, medical images, and tabular data), as presented in **Section 3**. Fig. 9 offers a comparative analysis of the four properties of explanation methods detailed in **Section 2.1.3**: *expressive power*, *translucency*, *portability*, and *algorithmic complexity*. *Algorithmic complexity*, *portability*, and *translucency* pertain to the interaction with the model whereas *expressive power* relates to the interaction with the end-user. Fig. 9 reveals that the properties of *algorithmic complexity*, *portability*, and *translucency* remain consistent across various application scenarios, whereas *expressive power* differs.

From our analysis, we reach a conclusion similar to that of Carvalho [18] regarding the consistent properties of explanation methods. The use of fuzzy rules, as a model-specific explanation method, exhibit the highest *translucency* and *algorithmic complexity* and lowest *portability* owing to its dependence on the specific model structure and parameters. Conversely, the use of SHAP, as a model-agnostic explanation method, has the lowest *translucency* and highest *portability* owing to its independence of the model. Both SHAP and a heat map have high *algorithmic complexity* owing to their reliance on external computational resources. However, heat maps are more generic and moderate as there are many ways to draw a heat map. The advantages of each interpretability method are elaborated in the following for three scenarios.

In scenarios involving sequence data (Fig. 4), heat maps directly show the importance and contribution of each data point in the raw data. In contrast, fuzzy rules and a SHAP plot explain only abstract features extracted from the data, which may not provide adequate understanding from the perspective of the end-user (as elaborated in **Section 3.1**). In disease diagnosis scenarios involving tabular data, particularly clinical records, fuzzy rules offer substantial advantages in terms of the understandability of explanation methods owing to their human-like inference processes (as elaborated in **Section 3.3**). SHAP provides the importance or contributions of features to the model's output, which is result interpretable rather than process interpretable, and this is the reason why its *expressive power* is not as good as that of fuzzy rule. A heat map is commonly used in the correlation analysis of features, which is data interpretable. This is efficient but there are other methods that can also be utilized to analyze data. In disease diagnosis scenarios involving image data, heat maps (such as Grad-CAM) and attention maps effectively illustrate what has been learned by the CNN and the contributions made to the recognition results (as elaborated in **Section 3.2**). Therefore, in scenarios of disease diagnosis involving tabular data, such as data on clinical features, tumor features, and radiomic features, we highly recommend that researchers apply fuzzy rules as the explanation method to enhance model interpretability.

When assisting doctors in diagnosing diseases from tabular data, the FIS has a distinct advantage of providing highly

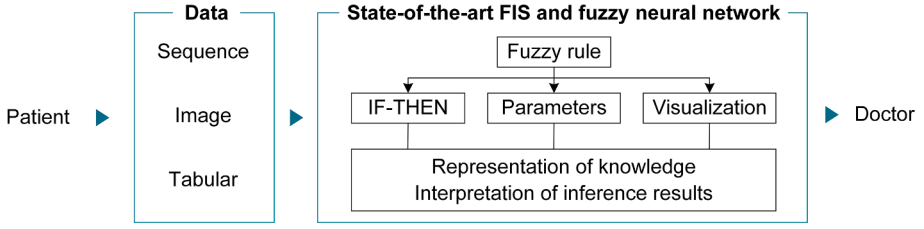

Fig. 8. Interpretability of an FIS in the scenario of disease diagnosis.

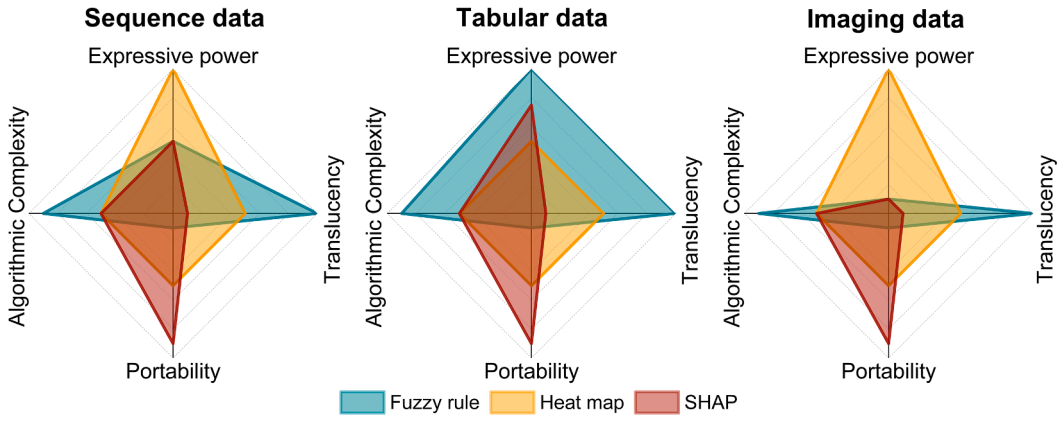

Fig. 9. Subjective comparisons of the properties of explanation methods in disease diagnosis scenarios revealing that the properties of *algorithmic complexity*, *portability*, and *translucency* remain consistent regardless of the application scenario, whereas the *expressive power* of various explanation methods varies across different scenarios.

understandable interpretability in the form of IF–THEN rules for the reasoning process, inferred results, and represented knowledge. Moreover, the FIS has the ability to extract and handle fuzzy and uncertain features in data, and its inference process is inherently interpretable. However, it is indisputable that some explanation methods offer better interpretability than fuzzy rules in other applications due to various limitations, indicating that there is room for further improvement of FISs. The primary difficulty in adopting an FIS lies in the high requirement for the understandability of the input features. The inclusion of easily comprehensible data features, such as clinical characteristics, enhances the *expressive power* of fuzzy rules. However, when dealing with medical imaging data or sequence data, which are also commonly used in the medical field, both the pixel-level information in the original images and the abstract features extracted from the images pose challenges in terms of understandability for physicians. Another limitation is the absence of an intuitive visual representation of fuzzy rules. The commonly used result view or surface view of fuzzy rules (as shown in Fig. 3) primarily serve as means for researchers to validate their theories rather than to facilitate the interpretation of the reasoning process and output results. As a result, the representation remains abstract and less comprehensible for physicians.

Although this review centers on FISs, we strongly recommend that researchers adopt different complementary interpretation methods, individually or in combination, whenever appropriate according to the data type to obtain simple and transparent medical decision-making insights into disease diagnosis.

4.2. Future Trends of FIS

4.2.1. Medical applications of novel FISs

Currently, researchers are focusing on the development of novel and powerful FISs, which hold great potential in the medical field. Notably, in the field of medical disease diagnosis, classical FISs continue to be extensively applied [10]. Therefore, the development of novel FISs that cater to the substantial demand for interpretability and recognition accuracy in the contemporary AI paradigm is crucial. Numerous emerging methods have been introduced to optimize fuzzy neural networks and have been evaluated with public datasets, as discussed in **Section 2.2.1**. Zhang et al. [48] evaluated a sensitivity-ensemble-level-based TSK fuzzy system with epilepsy EEG data and reported great accuracy and high interpretability. In medical applications, it is common for a single patient to undergo multiple medical tests to confirm a disease, leading to the adoption of multi-view or multi-modality approaches. The research of Zhang et al. concentrated on state-of-the-art FIS methods and is representative of tentative and potential studies of novel FISs in a spectrum of medical applications.

4.2.2. Interpretability enhancement of FISs

As discussed in previous sections, the interpretability of FISs in disease diagnosis applications is limited when the input features become increasingly abstract and complex. In such cases, ensemble learning can be implemented through the fusion of multiple TSK fuzzy systems and the adoption of appropriate ensemble learning strategies. This approach has been shown to be effective in eliminating the curse of dimensionality problem and reducing the number of fuzzy rules, thus enhancing the interpretability of TSK fuzzy systems. Zhang et al. [49] provided a comprehensive survey of TSK fuzzy system fusion strategies to elaborate the TSK fuzzy system embedding methods and their interpretability when dealing with high-dimension and complex data, such as multimodal data in the field of medicine. In contrast, model-agnostic explanation methods, such as SHAP and Grad-CAM, can be integrated with FISs to maximize model interpretability. For instance, with an increasing awareness of the importance of FISs, concerted efforts have been made to simultaneously apply multiple explanation methods, such as drawing heat maps based on SHAP theory [31]. This paves the way for the incorporation of model-agnostic explanation methods to further enhance the interpretability of FISs. Given this strength, the role of fuzzy rules in FISs is expected to become increasingly prominent in the AI era in terms of leveraging the enhanced accuracy of complex and sophisticated prediction models while providing semantic human-like reasoning and interpretability of the prediction

results. These improvements, in turn, are expected to bridge the communication gap between humans and machines, leading to advancements in XAI in the medical domain.

4.2.3. Reduction of the complexity of fuzzy rules

FISs with lightweight fuzzy rules provide high understandability and generalizability. Studies have attempted to define evaluation metrics for fuzzy rules, such as the length and the number of the rules. These studies have reported that a greater complexity of fuzzy rules is associated with a greater difficulty for individuals to understand and interpret the rules. Conversely, mitigating complexity facilitates the learning of more general fuzzy rules that are suitable for diverse application scenarios, thus improving the generalizability of FISs. In disease diagnosis, differences in data across different medical institutions, such as variations in equipment, operational procedures, and image algorithms, necessitate the development of generalizable fuzzy rules, which is expected to become an interesting and important topic of research on FIS. For instance, Zhou et al. [50] recently adopted a deep TSK fuzzy classifier with a random rule heritage (Drrh-TSK-FC) to recognize the sleep stage from EEG signals. In their research, several randomly generated short fuzzy rules were used to imitate the cognitive behavior of past experiences (i.e., the represented knowledge base-fuzzy rules) to solve new yet similar problems. Moreover, the simplicity of fuzzy rules in Drrh-TSK-FC enabled the expansion of fuzzy neural networks to deeper levels, providing fuzzy sub-classifiers with better uncertainty handling and generalization capabilities.

5. Conclusion

There is an increasing, yet unsatisfied, demand for interpretable AI (or XAI), particularly to achieve human machine interactions and trustworthy AI for bench-to-bedside translation in clinics. Among various explanation methods for interpretable AI, the use of fuzzy rules embedded in FISs is a novel and powerful technique. However, there are few reviews of the use of FISs in medical diagnosis. In addition, the role of fuzzy rules for different types of multimodal medical data has been little discussed. In this review, we not only provide the fundamental and historical knowledge of fuzzy rules to facilitate readers better appreciate fuzzy logic, working principles, and the underlying semantic interpretability but also discuss the strengths and weaknesses of fuzzy rules for three major types of multimodal medical data used in diagnosis compared with those of two other popular methods (i.e., the adoption of SHAP and a heat map). In the field of medical diagnosis, we strongly recommend the adoption of fuzzy rules when using tabular data (such as clinical parameters, diseases characteristics, and radiomic features) owing to their high *expression power* and *translucency*. Given the current limitations of fuzzy rules discussed in this review, we recommend that researchers combine fuzzy rules with other explanation methods (such as the use of SHAP and a heat map), whenever appropriate, to maximize the interpretability of AI models used in the bench-to-bedside translation of prediction models applied to clinical decision-making processes. In the contemporary AI paradigm, model interpretability has become indispensable in clinical implementation for various stakeholders, specifically medical practitioners and patients. The advantages of emerging FIS technique, as discussed in this review, mean that FISs outperform other explanation methods in terms of interpretability. As both FIS and AI techniques are evolving, it is anticipated that the role and impact of FISs will become even greater in the medical domain.

CRediT authorship contribution statement

Jin Cao: Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing. **Ta Zhou:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Shaohua Zhi:** Methodology, Writing – review & editing. **Saikit Lam:** Methodology, Writing – review & editing. **Ge Ren:** Methodology, Writing – review & editing. **Yuanpeng Zhang:** Methodology, Writing – review & editing. **Yongqiang Wang:** Methodology, Writing – review & editing. **Yanjing Dong:** Methodology, Writing – review & editing. **Jing Cai:** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

This research was partly supported by research grants of Shenzhen Basic Research Program (JCYJ20210324130209023), Mainland-Hong Kong Joint Funding Scheme (MHKJFS) (MHP/005/20), Health and Medical Research Fund (HMRF 09200576), The Health Bureau, The Government of the Hong Kong Special Administrative Region, Project of Strategic Importance Fund (P0035421), Project of RISA Fund (P0043001) and Centrally Funded Postdoctoral Fellowship Scheme (P0045698) from The Hong Kong Polytechnic University, and by the Project of Ministry of Education 'Chunhui plan' cooperative Scientific Research (HZKY20220133), the national natural science foundation of Jiangsu, China under Grant BK20191200, and by the natural science foundation of Jiangsu Universities, China under Grant 19JKD520003, and by the national defense basic research program of China under Grant JCKY2020206B037 and

by Jiangsu Graduate Scientific Research Innovation Project under Grant KYCX21_3506 and KYCX22_3825.

References

[1] Y. Zhang, S. Lam, T. Yu, X. Teng, J. Zhang, F.-K.-H. Lee, K.-H. Au, C.-W.-Y. Yip, S. Wang, J. Cai, Integration of an imbalance framework with novel high-generalizable classifiers for radiomics-based distant metastases prediction of advanced nasopharyngeal carcinoma, Knowl.-Based Syst. 235 (2022) 107649.
[2] S.-K. Lam, J. Zhang, Y.-P. Zhang, B. Li, R.-Y. Ni, T. Zhou, T. Peng, A.-L.-Y. Cheung, T.-C. Chau, F.-K.-H. Lee, et al., A multi-center study of CT-based neck nodal radiomics for predicting an adaptive radiotherapy trigger of ill-fitted thermoplastic masks in patients with nasopharyngeal carcinoma, Life 12 (2) (2022) 241.
[3] G. Ren, B. Li, S.-K. Lam, H. Xiao, Y.-H. Huang, A.-L.-Y. Cheung, Y. Lu, R. Mao, H. Ge, F.-M.-S. Kong, et al., A transfer learning framework for deep learning-based CT-to-perfusion mapping on lung cancer patients, Front. Oncol. 12 (2022) 883516.
[4] Y. Zhang, D. Yang, S. Lam, B. Li, X. Teng, J. Zhang, T. Zhou, Z. Ma, T.C. Ying, J. Cai, Radiomics-Based Detection of COVID-19 from Chest X-ray Using Interpretable Soft Label-Driven TSK Fuzzy Classifier, Diagnostics 12 (11) (2022) pp.
[5] W. Ding, M. Abdel-Basset, H. Hawash, A.M. Ali, Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey, Inf. Sci. (2022).
[6] "Explainable ai: the basics." https://futurium.ec.europa.eu/system/files/ged/ai-and-interpretability-policy-briefing_creative_commons.pdf, Nov. 2019.
[7] Y. Li, P. Qian, S. Wang, S. Wang, "Novel multi-view Takagi-Sugeno-Kang fuzzy system for epilepsy EEG detection", Journal of Ambient Intelligence and Humanized, Computing 5 (0123456789) (2021) pp.
[8] W. Xue, T. Zhou, J. Cai, Horizontal progressive and longitudinal leapfrogging fuzzy classification with feature activity adjustment, Appl. Soft Comput. 119 (2022) 108511.
[9] Y. Zhang, X. Li, J. Zhu, C. Wu, Q. Wu, Epileptic EEG signals recognition using a deep view-reduction tsk fuzzy system with high interpretability, IEEE Access 7 (2019) 137344–137354.
[10] G. Casalino, G. Castellano, U. Kaymak, G. Zaza, Balancing accuracy and interpretability through neuro-fuzzy models for cardiovascular risk assessment, in: In 2021 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 2021, pp. 1–8.
[11] E.A. Algehyne, M.L. Jibril, N.A. Algehainy, O.A. Alamri, A.K. Alzahrani, Fuzzy neural network expert system with an improved Gini index random forest-based feature importance measure algorithm for early diagnosis of breast cancer in saudi arabia, Big Data and Cognitive Computing 6 (1) (2022) 13.
[12] T. Sharma, V. Singh, S. Sudhakaran, N.K. Verma, Fuzzy based pooling in convolutional neural network for image classification, in: In 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, 2019, pp. 1–6.
[13] J. M. Alonso Moral, C. Castiello, L. Magdalena, C. Mencar, J. M. Alonso Moral, C. Castiello, L. Magdalena, and C. Mencar, "Interpretability constraints and criteria for fuzzy systems," *Explainable fuzzy systems: paving the way from interpretable fuzzy systems to explainable AI systems*, pp. 49–89, 2021.
[14] H. Hakkoum, A. Idri, I. Abnane, Assessing and comparing interpretability techniques for artificial neural networks breast cancer classification, Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization 9 (6) (2021) 587–599.
[15] "Google trends." https://trends.google.com/trends/, 2023.
[16] Y. Deng, R. He, Y. Liu, Crime risk prediction incorporating geographical spatiotemporal dependency into machine learning models, Inf. Sci. 646 (2023) 119414.
[17] T. Szandaa, Unlocking the black box of CNNs: Visualising the decision-making process with prism, Inf. Sci. 642 (2023) 119162.
[18] D.V. Carvalho, E.M. Pereira, J.S. Cardoso, Machine learning interpretability: A survey on methods and metrics, Electronics 8 (8) (2019) 832.
[19] M. Robnik-Šikonja and M. Bohanec, "Perturbation-based explanations of prediction models," *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*, pp. 159–175, 2018.
[20] L. Zadeh, "Fuzzy sets," *Inform Control*, vol. 8, pp. 338–353, 1965.
[21] Mamdani,, Application of fuzzy logic to approximate reasoning using linguistic synthesis, IEEE Trans. Comput. 100 (12) (1977) 1182–1191.
[22] M. Sugeno, T. Takagi, Fuzzy identification of systems and its applications to modeling and control, IEEE Trans. Syst. Man Cybern. 15 (1) (1985) 116–132.
[23] J.-S. Jang, ANFIS: adaptive-network-based fuzzy inference system, IEEE Trans. Syst. Man Cybern. 23 (3) (1993) 665–685.
[24] D. Wu, R. Peng, J.M. Mendel, Type-1 and interval Type-2 fuzzy systems [ai-explained], IEEE Comput. Intell. Mag. 18 (1) (2023) 81–83.
[25] O. Castillo, J.R. Castro, P. Melin, Forecasting the COVID-19 with interval Type-3 fuzzy logic and the fractal dimension, Int. J. Fuzzy Syst. 25 (1) (2023) 182–197.
[26] P. Melin, D. Sánchez, J.R. Castro, O. Castillo, Design of Type-3 fuzzy systems and ensemble neural networks for COVID-19 time series prediction using a firefly algorithm, Axioms 11 (8) (2022) 410.
[27] L.A. Zadeh, The concept of a linguistic variable and its application to approximate reasoning – I, Information Sciences 8 (3) (1975) 199–249.
[28] J. Sanz, M. Sesma-Sara, H. Bustince, A fuzzy association rule-based classifier for imbalanced classification problems, Inf. Sci. 577 (2021) 265–279.
[29] Y. Gu, K. Xia, K.-W. Lai, Y. Jiang, P. Qian, X. Gu, Transferable takagi-sugeno-kang fuzzy classifier with multi-views for EEG-based driving fatigue recognition in intelligent transportation, IEEE Trans. Intell. Transp. Syst. (2022).
[30] Y. Tao, Y. Jiang, K. Xia, J. Xue, L. Zhou, P. Qian, Classification of EEG signals in epilepsy using a novel integrated tsk fuzzy system, J. Intell. Fuzzy Syst. 40 (3) (2021) 4851–4866.
[31] S.K. Khare, U.R. Acharya, An explainable and interpretable model for attention deficit hyperactivity disorder in children using EEG signals, Comput. Biol. Med. 155 (2023) 106676.
[32] M. Rashed-Al-Mahfuz, M.A. Moni, P. Lio, S.M.S. Islam, S. Berkovsky, M. Khushi, J.M. Quinn, Deep convolutional neural networks based ECG beats classification to diagnose cardiovascular conditions, Biomed. Eng. Lett. 11 (2021) 147–162.
[33] M.Z. Rahman, M.A. Akbar, V. Leiva, A. Tahir, M.T. Riaz, C. Martin-Barreiro, An intelligent health monitoring and diagnosis system based on the internet of things and fuzzy logic for cardiac arrhythmia COVID-19 patients, Comput. Biol. Med. 154 (2023) 106583.
[34] D. Zhang, S. Yang, X. Yuan, P. Zhang, Interpretable deep learning for automatic diagnosis of 12-lead electrocardiogram, Iscience 24 (4) (2021) 102373.
[35] A. Agrawal, A. Chauhan, M.K. Shetty, M.D. Gupta, A. Gupta, et al., ECG-ICOVIDnet: Interpretable ai model to identify changes in the ECG signals of post-covid subjects, Comput. Biol. Med. 146 (2022) 105540.
[36] M. Rashed-Al-Mahfuz, M.A. Moni, S. Uddin, S.A. Alyami, M.A. Summers, V. Eapen, A deep convolutional neural network method to detect seizures and characteristic frequencies using epileptic electroencephalogram (EEG) data, IEEE J. Translat. Eng. Health Med. 9 (2021) 1–12.
[37] M. Ma, R. Liu, C. Wen, W. Xu, Z. Xu, S. Wang, J. Wu, D. Pan, B. Zheng, G. Qin, et al., Predicting the molecular subtype of breast cancer and identifying interpretable imaging features using machine learning algorithms, Eur. Radiol. (2022) 1–11.
[38] Q. Teng, Z. Liu, Y. Song, K. Han, Y. Lu, A survey on the interpretability of deep learning in medical diagnosis, Multimedia Syst. (2022) 1–21.
[39] H. Panwar, P. Gupta, M.K. Siddiqui, R. Morales-Menendez, P. Bhardwaj, V. Singh, A deep learning and Grad-CAM based color visualization approach for fast detection of covid-19 cases using chest X-ray and CT-scan images, Chaos Solitons Fractals 140 (2020) 110190.
[40] R. Zhang, W. Lu, J. Gao, Y. Tian, X. Wei, C. Wang, X. Li, M. Yu, RFI-GAN: A reference-guided fuzzy integral network for ultrasound image augmentation, Inf. Sci. 623 (2023) 709–728.
[41] C. Wang, X. Lv, M. Shao, Y. Qian, Y. Zhang, A novel fuzzy hierarchical fusion attention convolution neural network for medical image super-resolution reconstruction, Inf. Sci. 622 (2023) 424–436.
[42] K. Bai, X. Zhu, S. Wen, R. Zhang, W. Zhang, Broad learning based dynamic fuzzy inference system with adaptive structure and interpretable fuzzy rules, IEEE Trans. Fuzzy Syst. 30 (8) (2021) 3270–3283.
[43] I. Thani, T. Kasbe, Expert system based on fuzzy rules for diagnosing breast cancer, Heal. Technol. 12 (2) (2022) 473–489.
[44] T.I. Ahmed, J. Bhola, M. Shabaz, J. Singla, M. Rakhra, S. More, I.A. Samori, et al., Fuzzy logic-based systems for the diagnosis of chronic kidney disease, Biomed Res. Int. 2022 (2022).
[45] Z. Dong, Q. Wang, Y. Ke, W. Zhang, Q. Hong, C. Liu, X. Liu, J. Yang, Y. Xi, J. Shi, et al., Prediction of 3-year risk of diabetic kidney disease using machine learning based on electronic medical records, J. Transl. Med. 20 (1) (2022) 1–10.

[46] J. Liu, X. Dong, H. Zhao, Y. Tian, Predictive classifier for cardiovascular disease based on stacking model fusion, Processes 10 (4) (2022) 749.
[47] M. Pal, S. Mistry, and D. De, "Interpretability approaches of explainable ai in analyzing features for lung cancer detection," in *Frontiers of ICT in Healthcare: Proceedings of EAIT 2022*, pp. 277–287, Springer, 2023.
[48] Y. Zhang, G. Wang, X. Huang, W. Ding, TSK fuzzy system fusion at sensitivity-ensemble-level for imbalanced data classification, Information Fusion 92 (2023) 350–362.
[49] Y. Zhang, G. Wang, T. Zhou, X. Huang, S. Lam, J. Sheng, K.S. Choi, J. Cai, W. Ding, Takagi-Sugeno-Kang fuzzy system fusion: A survey at hierarchical, wide and stacked levels, Information Fusion 101 (2024) 101977.
[50] T. Zhou, G. Wang, K.-S. Choi, S. Wang, "Recognition of sleep-wake stages by deep Takagi-Sugeno-Kang fuzzy classifier with random rule heritage", IEEE Transactions on Emerging Topics, Comput. Intell. (2023).