IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 July 2024, accepted 2 August 2024, date of publication 5 August 2024, date of current version 16 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3439096

RESEARCH ARTICLE

Deciphering Knee Osteoarthritis Diagnostic Features With Explainable Artificial Intelligence: A Systematic Review

YUN XIN TEOH[1,2], ALICE OTHMANI[2], SIEW LI GOH[3,4], JULIANA USMAN[1], AND KHIN WEE LAI[1], (Senior Member, IEEE)

[1]Department of Biomedical Engineering, Faculty of Engineering, Universiti Malaya, Kuala Lumpur 50603, Malaysia
[2]Laboratoire Images, Signaux et Systèmes Intelligents (LISSI), Université Paris-Est Créteil, 94400 Vitry-sur-Seine, France
[3]Sports and Exercise Medicine Research and Education Group, Faculty of Medicine, Universiti Malaya, Kuala Lumpur 50603, Malaysia
[4]Centre for Epidemiology and Evidence-Based Practice, Faculty of Medicine, Universiti Malaya, Kuala Lumpur 50603, Malaysia

Corresponding authors: Khin Wee Lai (lai.khinwee@um.edu.my) and Alice Othmani (alice.othmani@u-pec.fr)

This work was supported in part by the Ministry of Higher Education, Malaysia; and in part by Universiti Malaya through the Fundamental Research Grant Scheme (FRGS) under Grant FRGS/1/2023/SKK05/UM/02/2.

ABSTRACT Existing artificial intelligence (AI) models for diagnosing knee osteoarthritis (OA) have faced criticism for their lack of transparency and interpretability, despite achieving medical-expert-like performance. This opacity makes them challenging to trust in clinical practice. Recently, explainable artificial intelligence (XAI) has emerged as a specialized technique that can provide confidence in the model's prediction by revealing how the prediction is derived, thus promoting the use of AI systems in healthcare. This paper presents the first survey of XAI techniques used for knee OA diagnosis. This survey identified 78 AI-based primary knee OA diagnostic test accuracy studies, of which 70 (89.7%) employed XAI. In 34 out of 70 (48.6%) of studies, XAI was utilized for the goal of visualization of predictions. Gradient-weighted class activation mapping (GradCAM) is the most common technique, being used in 24 out of 70 studies (34.3%), followed by SHapley Additive exPlanations (SHAP), being used in 9 out of 70 (12.9%) studies. All included studies analyzed the outcomes generated by XAI methods through qualitative analysis. However, only three studies utilized quantitative measures to evaluate the reliability of the XAI outcomes. We also observed that 64.3% of the studies utilized widely-circulated dataset, namely Osteoarthritis Initiative (OAI) extensively.The XAI techniques are discussed from two perspectives: data interpretability and model interpretability. Our paper provides an overview of XAI's potential towards a more reliable knee OA diagnosis approach and helps to encourage its adoption in clinical practice.

INDEX TERMS Computer aided diagnosis, explainable artificial intelligence, explanation representation, knee osteoarthritis, radiology.

I. INTRODUCTION

Osteoarthritis (OA) is a prevalent degenerative joint disease that affects millions of people worldwide [1], [2], with the weight-bearing knee joint being particularly susceptible. Radiography is a commonly used diagnostic tool for knee OA [3]. However, its diagnostic precision is often compromised due to the subjective nature of image interpretation and the perceptual differences among radiologists, which are influenced by their individual knowledge and experience [4]. To maximize the accuracy of OA diagnosis, researchers have also explored modeling OA using multimodal and multidimensional data to encompass a comprehensive range of patient information [5], [6], [7]. These data could be demographic, societal, symptomatic, medical history, biomechanical, biochemical, genetic, and behavioral characteristics. Artificial intelligence (AI) models have demonstrated the ability to automate diagnosis and have shown promising

The associate editor coordinating the review of this manuscript and approving it for publication was Inês Domingues.

results, achieving diagnostic accuracy on par with medical experts using either individual or combined data [8], [9], [10]. However, the specific impact of each factor within the data and the correlation between these factors remain largely unexplored.

Furthermore, there is a growing concern about the lack of transparency and interpretability of AI models in healthcare settings [11], [12], [13], [14]. The use of AI models in medical data for OA diagnosis shows potential in reducing the subjectivity and variability linked to human interpretation. However, those AI approaches predominantly rely on black-box models, which lack transparency and interpretability [11], [12]. In contrast to the human reasoning process, which depends on complex cognitive abilities, intuition, and the assimilation of diverse knowledge and experiences to make decisions, AI models make predictions based on the learning outcomes from training datasets. The internal workings of these models remain hidden or unknown, even to their designers. This lack of transparency can engender uncertainty and erode trust among patients and healthcare providers. Additionally, the use of black-box models impedes the development of health mobile applications for disease management [15]. According to a survey conducted by Mrklas et al. [15], a significant number of patients and physicians have expressed a strong desire for a visual symptom graph to aid in monitoring their condition.

Explainable AI (XAI) [16] offers a potential solution to these concerns by providing a transparent and interpretable framework for automated analysis of radiographic images. XAI algorithms can identify specific regions of interest within the image and provide a clear explanation of the factors that contributed to the final diagnosis [17], [18]. This could help to overcome the limitations of traditional radiographic diagnosis and increase the accuracy and consistency of knee OA diagnosis.

By leveraging XAI, healthcare providers could have a more objective and transparent method for diagnosing knee OA, leading to earlier detection and more timely treatment. XAI could also potentially reduce the need for costly and invasive diagnostic procedures, such as arthroscopy, which are currently used for further evaluation to confirm cartilage lesion.

A. MOTIVATIONS AND PAPER CONTRIBUTIONS

In recent years, as XAI gains popularity, numerous survey papers have emerged discussing its application in healthcare settings [16], [17], [18], [19], [20], [21], [22]. Despite this growing interest, there is still a noticeable lack of comprehensive survey papers that delve into the specific application of XAI for diagnosing knee OA. Furthermore, many existing XAI strategies have been designed with a general-purpose approach and may not fully address the unique clinical concerns and domain-specific knowledge required for accurate diagnosis of knee OA. Therefore, there is a need for specialized XAI frameworks that take into account the specific clinical considerations and incorporate relevant domain knowledge to enhance the application of XAI in diagnosing knee OA effectively. To address this gap, it is crucial to explore different explanation methods and evaluate their effectiveness. By conducting a comprehensive review of the literature on interpretability and explainability of AI models for knee OA diagnosis, we can gain a deeper understanding of these concepts and their potential applications. Such a review will provide valuable insights into how interpretability and explainability can be leveraged to improve AI and machine learning models for knee OA diagnosis. To the best of our knowledge, this paper represents the first survey dedicated to exploring the application of XAI in knee OA diagnosis. The contributions of the paper include:

- Systematic review of the current state-of-the-art explainability and interpretability methods for neural networks used in diagnosing knee OA from medical data;
- Comparison of the existing knee OA datasets and the performance analysis of different explainability and interpretability methods in AI models;
- Identification of the potential clinical impact of the most promising explainability and interpretability methods by assessing their practicality, scalability, and effectiveness in real-world clinical settings for improving diagnostic accuracy and reducing misdiagnosis rates in knee OA.

B. ORGANIZATION OF PAPER

This review paper is partitioned into seven sections. Firstly, Section II introduces the preliminaries and fundamental concepts of XAI. Section III describes the study protocol, including the search strategy, as well as the inclusion and exclusion criteria for selecting relevant studies. In Section IV, the overview of the included studies is presented. Next, Section V introduces an XAI taxonomy and explores various techniques for achieving data and model interpretability. The implications and potential applications of XAI are discussed in Section VI, which also suggests promising avenues for future research. Finally, Section VII offers a comprehensive conclusion to this study, summarizing its findings and highlighting its contributions to the field of XAI in knee OA assessment.

II. PRELIMINARIES AND FUNDAMENTAL CONCEPTS

A. CURRENT STATE OF AI TECHNOLOGIES FOR OA DIAGNOSIS AND LIMITATIONS

AI algorithms can accurately identify definite OA cases (Kellgren Lawrence grade ≥ 2) from healthy cases with an accuracy above 77% [23], [24], [25], [26]. Moustakidis et al. [26] further revealed that the OA diagnosis by the dense neural network (DNN) exhibits a high level of fairness and equity across various demographic groups, with a demographic parity of 98.5% and balanced equalized odds around 92%. However, AI models for the detection of early OA, especially Kellgren Lawrence grade 1, suffer from low accuracy at about 64.3%, due to less noticeable visual clues [23], [24]. Multi-class classification is often employed to grade the severity of OA structural damage. The reported multi-class accuracy for OA diagnosis ranges from 57.6% to 98.36% using pre-trained convolutional neural network (CNN) models, as indicated by previous studies [8], [9], [23], [27], [28], [29], [30], [31]

Despite the promising prediction outcomes, existing AI technologies for OA diagnosis heavily rely on black box supervised learning approaches, especially CNN and random forest models [32]. About 75% of the models remain unexplainable [32]. While these AI methods can make accurate predictions with rigorous training, they face a significant challenge known as overfitting. Overfitting occurs when the model becomes too focused on specific details of the training data, including noise, making its predictions biased and less reliable for new cases. Recent surveys and reviews of OA studies (Table 1) highlight a critical issue in OA research, which is the lack of external validation [32], [33], [34], [35]. Without thorough external validation, AI models may struggle to adapt to new datasets effectively. However, it is particularly challenging to acquire new datasets due to privacy issues.

To address these concerns, explainable models should be fully utilized to bridge the gap between AI and the human intelligence of medical experts and to enhance the reliability of the AI models. For instance, Tiulpin et al. [8] found that while a baseline model (fine-tuned ResNet-34) achieved higher performance metrics, it overlooked relevant OA radiological findings compared to their proposed model (deep Siamese CNN). Through further investigation, the authors discovered that the baseline model was prone to overfitting. It demonstrates that the adoption of explainable AI (XAI) ensures that general AI models make their decisions based on meaningful patterns, similar to how medical experts analyze radiology images, looking for visual indications of OA signs, such as joint space narrowing, bone spurs, etc. Moreover, Wang et al. [36] showed that the CNN model could outperform specialists in identifying surgical candidates who have Kellgren-Lawrence grades 3 and 4 with an F1 score of 0.923 when using attention maps for external validation.

B. XAI CONCEPTS AND FRAMEWORKS

Due to rapid development of artificial intelligence and machine learning technologies, it becomes increasingly important to understand how these models make predictions [37]. In this field, the terms "interpretability" and "explainability" are closely related and often used interchangeably [38], but they do have subtle differences in the context of deep learning. Here's a breakdown of each concept:

1) INTERPRETABILITY

Interpretability refers to the ability to understand and make sense of the internal workings of a deep learning model [39]. It involves gaining insights into how the model processes inputs, makes decisions, and generates outputs. An interpretable model allows humans to examine and comprehend the underlying mechanisms and logic employed by the model to arrive at its predictions or decisions [40].

2) EXPLAINABILITY

Explainability, on the other hand, focuses on providing human-understandable explanations for the model's outputs or predictions [39], [41]. It goes beyond mere interpretation and aims to make the decision-making process of the model transparent and understandable to non-experts. Explainable models not only produce accurate predictions but also provide intuitive explanations that can be easily comprehended by end-users or stakeholders.

In summary, while interpretability is primarily concerned with understanding the internal workings of a deep learning model, explainability goes a step further by providing human-understandable explanations for the model's outputs or decisions. Both concepts aim to enhance the transparency and trustworthiness of deep learning models, especially in high-stakes applications such as healthcare, finance, or autonomous systems.

Recently published XAI taxonomies [38], [42], [43] propose a conceptual framework for XAI, utilizing four evaluation dimensions to effectively describe the scope and characteristics of the XAI domain. These dimensions include:

- **Explanation scopes**, which can be divided into local (explaining individual prediction) or global (explaining the whole model) interpretability.
- **Model specificity**, which can be divided into model-specific and model-agnostic interpretability.
- **Interpretation types**, which can be divided into pre-model, intrinsic, post-hoc, and extrinsic interpretability.
- **Explanation forms**, which encompass various ways in which explanations can be presented or communicated.

The proposed XAI framework effectively tackles the technical concerns of general AI models. However, it lacks emphasis on the essential aspects of data and problem characteristics required for instilling domain knowledge into AI models. Moreover, it does not adequately consider the specific needs of lay users, such as medical experts [44]. These factors are crucial in ensuring that AI models are not only transparent and interpretable but also capable of effectively utilizing domain-specific information to enhance their performance and relevance in real-world applications. Therefore, [45] extend the general XAI framework by incorporating considerations for the type of input data, problem, and task. This extension aims to provide a more comprehensive and practical approach to XAI, catering to the specific needs of various domains and ensuring the successful integration of domain knowledge into AI models.

The realm of interpretability in XAI can be categorized into two distinct groups: perceptive interpretability and interpretability by mathematical structures, as proposed by Tjoa et al. [19]. Perceptive interpretability methods typically provide immediate interpretations, while methods that offer

TABLE 1. Summary of existing reviews and surveys on the topic of predicting OA disease using automated approaches.

Paper	Year	No. of included articles	Scope				
			Data		Model		
			Tabular	Image	ML	DL	XAI
[33]	2021	26	✗	✓	✓	✓	✗
[32]	2022	46	✓	✓	✓	✓	✗
[34]	2023	N/A	✗	✓	✗	✓	✗
[35]	2023	20	✓	✓	✓	✓	✗
Our paper	2024	70	✓	✓	✓	✓	✓

interpretation via mathematical structures produce outputs that require an additional layer of cognitive processing to reach a human-readable presentation. These taxonomies primarily focus on the transition from black-box models to white-box models, where the inner logic is fully explored and understood. Reference [39] introduce a novel approach by incorporating gray-box models. These models lie between black-box and white-box models, offering a partial understanding of the underlying mechanisms. By considering this intermediate category, the proposed taxonomy accounts for a broader range of interpretability levels and provides a more nuanced perspective on XAI. Compared to previous studies, their XAI taxonomy incorporates data explainability as an essential aspect to comprehend the datasets used in the AI models. This addition reflects their effort on providing insights into the transparency and interpretability of the data itself, in addition to understanding the model's decision-making process. By considering data explainability, the proposed taxonomy offers a more comprehensive approach in gaining a deeper understanding of AI systems and the role of data in shaping their predictions.

All previously proposed XAI taxonomies offer a structured framework for comprehending and classifying various aspects of XAI approaches and their applications. As highlighted by Nauta et al. [45], it is important to recognize that certain explanation methods have the ability to incorporate multiple types of explanations, thereby making the categories of explanation methods non-mutually exclusive.

In order to enhance the connection between users and XAI, [46] introduced a theoretical conceptual framework that establishes links between different XAI explanation facilities and user reasoning goals. Their work generated a concept called user-centric XAI, where the AI systems are designed by placing the end-users, such as healthcare professionals or patients, at the forefront of the explanation process, as illustrated in Figure 1. Their framework was meticulously designed to mitigate reasoning failures caused by cognitive biases. Additionally, [47] proposed a flowchart to guide the design of human-centered XAI systems. This flowchart incorporates three essential components: domain analysis, requirements analysis, and interaction design. By following this flowchart, XAI designers can ensure that their systems are aligned with user needs and provide effective explanations for improved user understanding and decision-making.

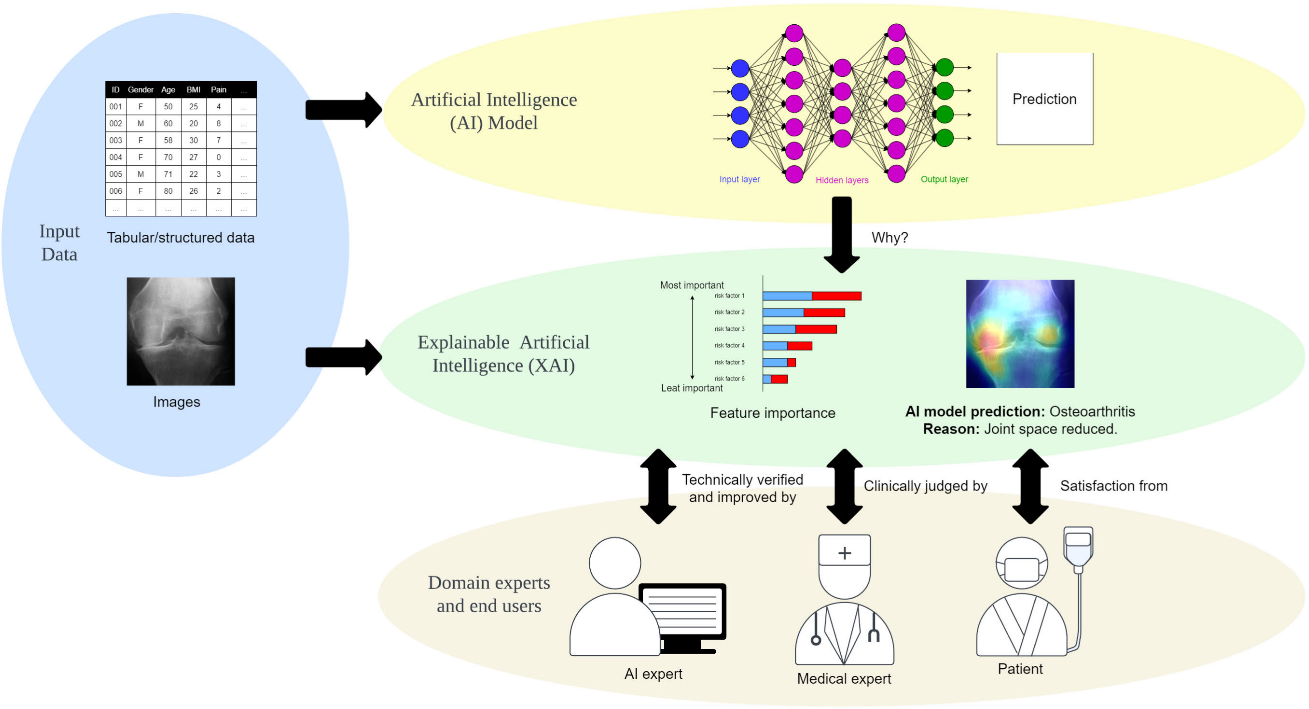
C. ETHICAL CONSIDERATIONS IN XAI

Global policy discussions are placing increasing emphasis on the integration of ethical standards into the design and implementation of AI-enabled technologies, highlighting the growing importance of Trustable AI. In 2018, the High-Level Expert Group on AI, established by the European Commission, published ethical guidelines focused on fostering trust in human-centric AI [48]. The guidelines highlighted seven key requirements for Trustable AI [49], as follows:

- **Human agency and oversight** that emphasize human autonomy and the importance of fundamental rights in decision-making.
- **Technical robustness and safety** that ensure AI systems are designed to prevent harm and promote resilience and security.
- **Privacy and data governance** that respect privacy and data protection while implementing sound data governance mechanisms.
- **Transparency** that advocates for transparency in data, system, and AI business models, complemented by traceability and explainability.
- **Diversity, non-discrimination, and fairness** that promote fairness and accessibility for all human while involve relevant stakeholders throughout the AI system's lifecycle.
- **Societal and environmental well-being** that focus on AI systems' positive impact on society and the environment, including sustainability considerations.
- **Accountability** that establishes mechanisms for responsibility and accountability, including auditability and accessible redress for AI system outcomes.

These requirements lead to the principles of Valid AI, Responsible AI, Privacy-preserving AI, and Explainable AI (XAI):

- **Valid AI** ensures that AI systems produce accurate and reliable results by using high-quality data, appropriate algorithms, and robust evaluation methods. It aims to



ID	Gender	Age	BMI	Pain	...
001	F	50	25	4	...
002	M	60	20	8	...
003	F	58	30	7	...
004	F	70	27	0	...
005	M	71	22	3	...
006	F	80	26	2	...
...	...	...	...	...	...

FIGURE 1. Illustration of XAI implementation for knee OA diagnosis. Through XAI, the decision-making process of AI models becomes interpretable and explainable, leading to the visualization of essential insights for AI expert, medical expert, and patient.

minimize errors and biases, making the AI outputs valid and trustworthy.

- **Responsible AI** involves designing and deploying AI systems in an ethical and socially conscious manner. It entails considering potential societal impacts, adhering to human values, and complying with legal and regulatory standards to minimize harm and promote positive outcomes.
- **Privacy-preserving AI** safeguards individuals' sensitive data during data processing and model training. These AI techniques ensure that personal information remains protected and confidential, preventing unauthorized access and preserving user privacy.
- **Explainable AI (XAI)** addresses the question of understanding the reasoning behind AI decisions. It provides transparency and interpretability to AI outputs, allowing users, including AI experts, medical professionals, and patients, to comprehend and trust the AI model's decisions.

In this framework, XAI plays a crucial role in addressing the fundamental question surrounding the rationale behind the decision-making process of AI systems, encompassing both human-level XAI (for human users) and machine-level XAI (for other AI models or systems). XAI techniques contribute to the transparency and interpretability required for achieving Trustable AI.

The European Union's High-Level Group on AI has made significant efforts to promote XAI by taking initiatives such to implement the General Data Protection Regulation (GDPR) [50], [51]. In addition, the proposal of the Artificial Intelligence Act by the European Commission represents their recent endeavors to foster a robust internal market for Artificial Intelligence (AI) systems [52], [53]. In the United States, the Defense Advanced Research Projects Agency (DARPA) has launched an XAI program aimed at tackling three key challenges: (1) developing more explainable models, (2) designing effective explanation interfaces, and (3) understanding the psychological requirements for effective explanations [54]. Despite considerable efforts, existing explainability methods still fall short in providing reassurance about the correctness of individual decisions, building trust among users, and justifying the acceptance of AI recommendations in clinical practice. Consequently, there is an immediate need to prioritize rigorous internal and external validation of AI models as a more direct approach to achieving the goals commonly associated with explainability [55].

III. SEARCH STRATEGY AND ELIGIBILITY CRITERIA

This systematic review was conducted based on the procedure proposed by Kitchenham et al. [56]. We conducted a comprehensive literature search using Boolean search strategy in five databases, namely Web of Science, Scopus, ScienceDirect, PubMed, and Google Scholar (Table 2). Our search included all publications up to May 20th, 2024. Eligibility screening was independently conducted by at least two of the authors.

TABLE 2. Boolean search strings employed for the corresponding bibliographic databases and search engines.

Database	Boolean search strings
Scopus	TITLE-ABS-KEY ("knee" OR "tibiofemoral joint") AND TITLE-ABS-KEY ("osteoarthritis" OR "degenerative arthritis" OR "degenerative joint disease" OR "wear-and-tear arthritis") AND TITLE-ABS-KEY ("XAI" OR (("explainable" OR "interpretable") AND ("AI" OR "artificial intelligen*" OR "deep learning" OR "machine learning")) OR "SHAP" OR "LIME" OR "gradcam" OR "grad cam" OR "heatmap" OR "saliency map" OR "attention map")
Web of Science	(((TS=(knee OR tibiofemoral joint)) AND TS=(osteoarthritis OR degenerative arthritis OR degenerative joint disease OR wear-and-tear arthritis)) AND TS=(XAI OR ((explainable OR interpretable) AND (AI OR artificial intelligen* OR deep learning OR machine learning)) OR SHAP OR LIME OR gradcam OR grad cam OR heatmap OR saliency map OR attention map))
PubMed	(knee OR tibiofemoral joint) AND (osteoarthritis OR degenerative arthritis OR degenerative joint disease OR wear-and-tear arthritis) AND (XAI OR ((explainable OR interpretable) AND (AI OR artificial intelligen* OR deep learning OR machine learning)) OR SHAP OR LIME OR gradcam OR grad cam OR heatmap OR saliency map OR attention map)
ScienceDirect	(knee) AND (osteoarthritis) AND (XAI OR ((explainable OR interpretable) AND (AI OR artificial intelligen OR deep learning OR machine learning)))
Google Scholar	intitle:((knee) AND (osteoarthritis OR joint disease) AND (XAI OR ((explainable OR interpretable) AND (AI OR artificial intelligen OR deep learning OR machine learning)) OR SHAP OR LIME OR gradcam OR grad cam OR heatmap OR saliency map OR attention map))

A. INCLUSION AND EXCLUSION CRITERIA

Papers will be included if they meet the following criteria:

- Focus on diagnostic tasks related to knee osteoarthritis (OA)
- Propose an end-to-end artificial intelligence (AI) model
- Utilize explainable AI (XAI) methods to provide explanations for the proposed model
- Not a review paper
- Published in English

B. DATA EXTRACTION

Our review identified a total of 78 studies that presented at least one knee OA computer-assisted diagnostic system utilizing an end-to-end AI approach. Among these studies, 70 out of 78 (89.7%) incorporated explainable AI (XAI) techniques and were included for our analysis (Figure 2). The earliest publication in this domain was found in 2017, which coincides with the introduction of popular XAI approaches such as gradient-weighted class activation map (GradCAM) [57] and self-attention mechanism [58]. The introduction of these techniques sparked increased interest and discussion surrounding XAI in the field of knee OA diagnosis, therefore the publication trend experienced exponential growth from 2017 to 2021, as depicted in Figure 3. Although there was a slight decrease in publications in 2022, the number of publications in high-quality (Q1) journals has been increasing steadily. This trend highlights the growing recognition and validation of XAI research in prestigious academic circles. Since 2022, the overall growth in publications has slowed, reflecting a shift towards more focused and impactful research in the field.

To provide an overview of our included papers, a bibliographic analysis and systematic review will be performed and presented in Section IV, based on the following breakdowns.

1) General query on knee OA assessment
2) Background of the study
3) Datasets for knee OA assessment
4) Classification systems for knee OA conditions

In Section V, detailed analysis of XAI methods for OA diagnosis will be presented. The XAI methods from all included articles will be analyzed based on their XAI explainability characteristics and categorized into either data, model, or post-hoc interpretability domains.

IV. OVERVIEW OF OA STUDY

A. GENERAL QUERY ON KNEE OA ASSESSMENT

We conducted an analysis of the general query to acquire an up-to-date comprehension of the topic on XAI application for knee OA diagnosis. This analysis aims to complement the qualitative literature review and provide valuable insights into the current state of research in this area. Co-occurrence analysis was performed using VOSviewer [59] to discover the relationships among terms extracted from the titles and

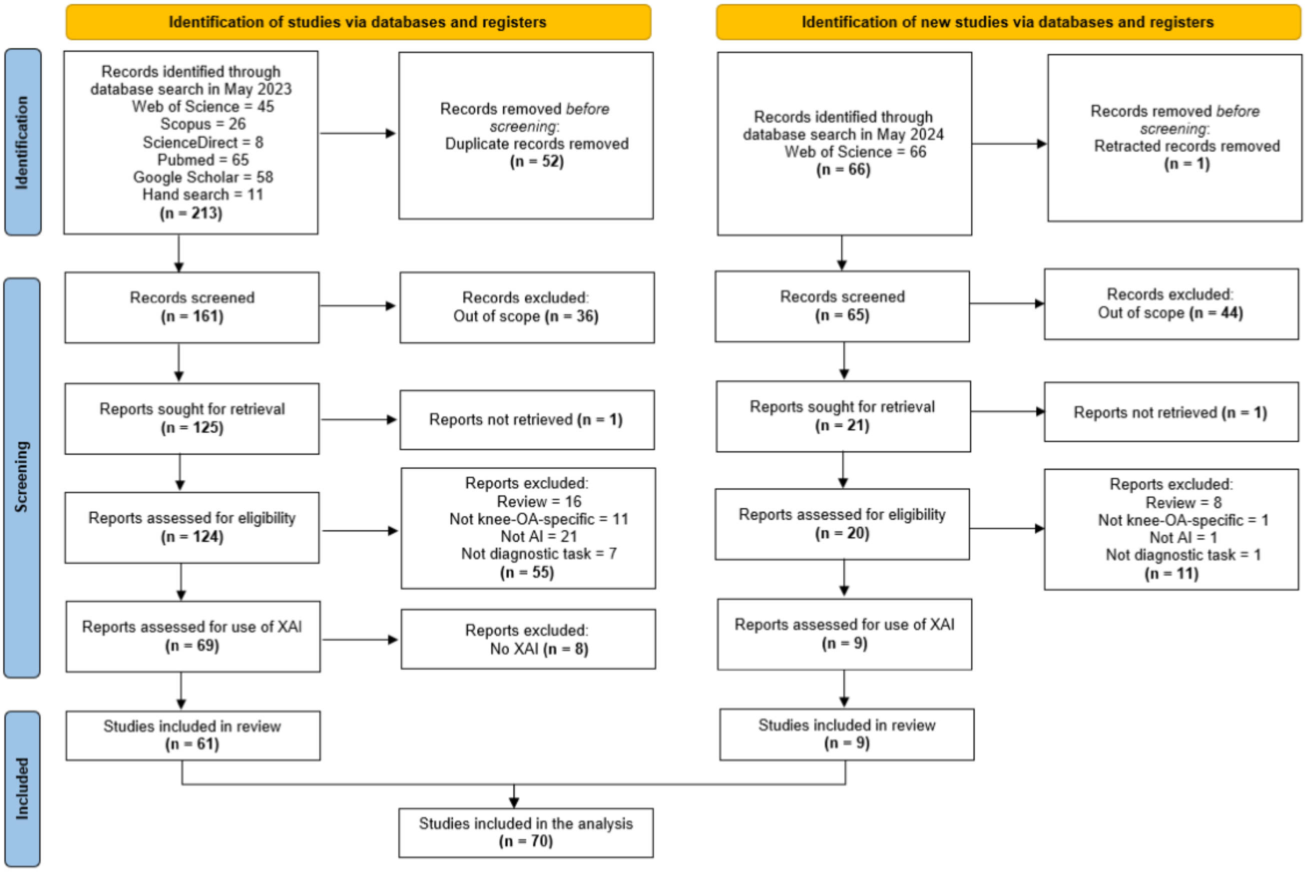

Identification of studies via databases and registers
Identification of new studies via databases and registers
Identification
Screening
Included
Records identified through database search in May 2023
Web of Science = 45
Scopus = 26
ScienceDirect = 8
Pubmed = 65
Google Scholar = 58
Hand search = 11
(n = 213)
Records removed before screening:
Duplicate records removed
(n = 52)
Records screened
(n = 161)
Records excluded:
Out of scope (n = 36)
Reports sought for retrieval
(n = 125)
Reports not retrieved (n = 1)
Reports assessed for eligibility
(n = 124)
Reports excluded:
Review = 16
Not knee-OA-specific = 11
Not AI = 21
Not diagnostic task = 7
(n = 55)
Reports assessed for use of XAI
(n = 69)
Reports excluded:
No XAI (n = 8)
Studies included in review
(n = 61)
Records identified through database search in May 2024
Web of Science = 66
(n = 66)
Records removed before screening:
Retracted records removed
(n = 1)
Records screened
(n = 65)
Records excluded:
Out of scope (n = 44)
Reports sought for retrieval
(n = 21)
Reports not retrieved (n = 1)
Reports assessed for eligibility
(n = 20)
Reports excluded:
Review = 8
Not knee-OA-specific = 1
Not AI = 1
Not diagnostic task = 1
(n = 11)
Reports assessed for use of XAI
(n = 9)
Studies included in review
(n = 9)
Studies included in the analysis
(n = 70)

FIGURE 2. PRISMA flowchart depicting the study selection process for this systematic review.

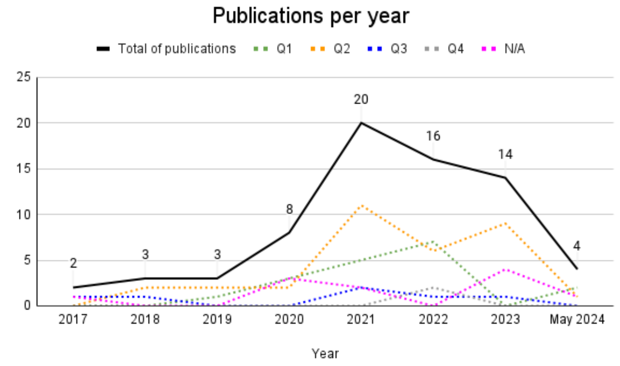


FIGURE 3. Trend of publications between 2017 to 2024.

abstracts of the selected studies. Out of the 1,692 terms identified, a subset of 222 terms with an occurrence frequency of at least three were chosen for analysis. Out of these 222 terms, we focused on the top 133 terms based on their relevance score, which fell within the top 60% range. These 114 terms were then included in our analysis to gain insights into their co-occurrence patterns and relationships (Figure 4).

As a result, the analysis revealed nine distinct clusters. Cluster 1 (19 items), Cluster 2 (19 items), Cluster 6 (15 items), and Cluster 7 (14 items) primarily focused on various aspects of knee OA symptoms and underlying risks, including bone and cartilage conditions, anterior cruciate ligament injury, demographics, and risk of OA deterioration, respectively. Cluster 3 (17 items), Cluster 4 (16 items), and Cluster 5 (15 items) emphasized model interpretability and clinical practitioners. Cluster 8 (13 items) mainly encompassed studies related to automatic early diagnosis of knee OA. Cluster 9 (5 items) was specifically associated with patient data.

The ten most cited terms included "severity" (26 occurrences), "radiograph" (22 occurrences), "detection" (22 occurrences), "pain" (15 occurrences), "cluster" (13 occurrences), "parameter" (13 occurrences), methodology (13 occurences), "risk factor" (12 occurrences), "mri" (12 occurrences), and "task" (12 occurrences).

B. BACKGROUND OF THE STUDY

1) GEOGRAPHICAL DISTRIBUTION

Extensive data collection for OA research was conducted in both Western (n = 16) and Eastern (n = 10) countries (Fig. 5), with a particular focus on the United States and China regions. Existing research heavily relied on data from United States, where European American is the largest population in

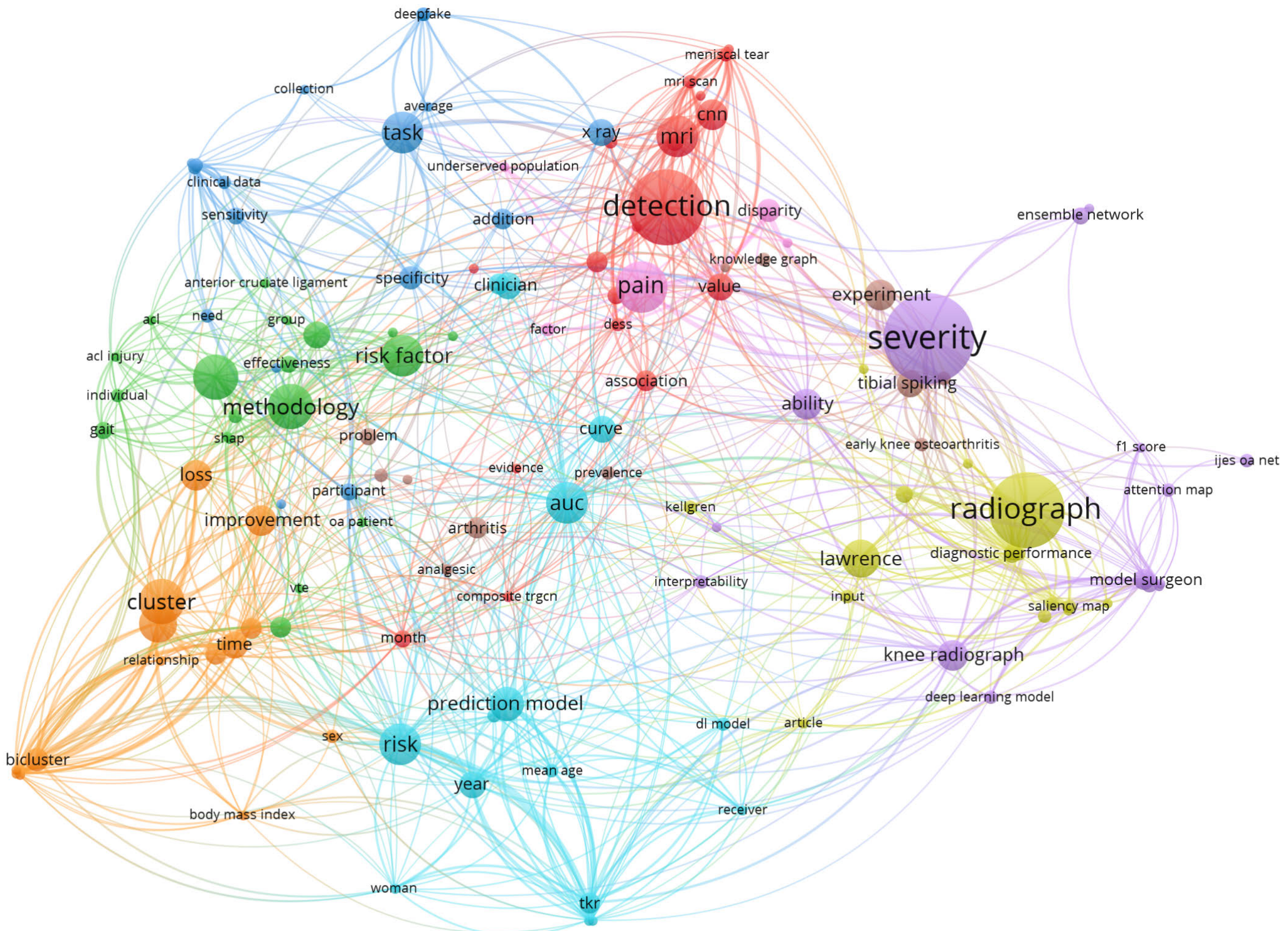


FIGURE 4. Visual representation of the scientific landscape of the selected studies using VOSviewer's mapping function.

the datasets. It is worth noting that limited research has been carried out in South American countries, and no research has been conducted in African countries.

2) SUBJECT AND INPUT DATA CHARACTERISTICS

Overall, the identified datasets included a wide range of sample size, varying from 40 to 4,796 individuals. Notably, 64.3% of the studies (45 out of 70) utilized the Osteoarthritis Initiative (OAI) dataset from United States for training or testing purposes. In 49 out of 70 studies, imaging data, primarily X-ray images (38 out of 49 studies), were utilized for clinical confirmation of OA disease. Approximately 38.6% of the studies (27 out of 70) employed tabular or structured data, such as demographics, clinical characteristics, and laboratory examinations, to predict the risk of OA incidence.

The use of single-channel data, whether images or tables, poses significant challenges in knee OA research, as it may limit the comprehensive understanding of the condition. To address this limitation, two studies [5], [9] adopted a data fusion approach, leading to the development of multi-modal data models that maximize the utilization of patient information. By integrating diverse data types (Figure 6), these innovative approaches achieved more comprehensive and accurate predictions for knee OA assessment.

C. DATASETS FOR KNEE OA ASSESSMENT

Knee OA is a complex and multifactorial disorder, and as such, a wide variety of data can be utilized to gain insights and explanations related to this health condition. In this review, we specifically focus on tabular data and image pixels. To track the evolving landscape of knee OA research, we performed an analysis of available datasets for knee OA assessment. This analysis provides a comprehensive understanding of the Western and Eastern data sources in knee OA diagnosis. We also highlight the role of datasets for generalizability and applicability of AI-based approaches.

1) TABULAR DATA

Tabular data in knee OA research is a collection of structured information encompassing both objective and subjective measurements of the condition. Within this tabular data, we have identified six distinct domains: demographic, clinical, imaging, patient-reported outcomes, biomechanics, and biomarkers. Demographic data represents information about participants' characteristics, such as medical history, symptoms, demographics, nutrition, physical activity, comorbidity, and behavioral aspects. Clinical data involves physical exam and blood measures, outlining patients' essential health information. Imaging data consists of medial imaging outcomes and anthropometrics for quantifying anatomical structures. Patient-reported outcomes focus on data collected through questionnaires to assess patient-reported symptoms and health-related quality of life. Biomechanical data involves the mechanics and movement of the knee during various activities. Biomarkers data includes measurable indicators found in bodily fluids, offering insights into disease status and treatment response. A comprehensive comparison of the accessibility, cost, and level of knowledge required for each domain is presented in Table 3. This evaluation aids researchers in understanding the strengths and limitations of each data domain.

2) IMAGE PIXELS

Image pixels in knee OA research consist of 2D or 3D data that allow for the visualization of human bone and tissue structures. This visual representation aids in gaining a deeper understanding of the anatomical aspects of the knee, enabling researchers to analyze and assess the condition more effectively. When paired with the tabular data of medical imaging outcomes, the combination of image pixels and structured data provides a more comprehensive approach to both qualitative and quantitative assessments of knee OA. This integration enhances the overall analysis and contributes to better knowledge discovery opportunities. However, handling image pixels data comes with challenges such as noise and resolution issues. High-resolution images offer improved visualization outcomes but also create a heavier computational load. Thus, a trade-off between image quality and computational efficiency needs to be carefully considered for practical implementation.

3) PUBLIC DATASETS FOR KNEE OA ASSESSMENT

Due to ethical concerns and strict institutional regulations, there are limited public datasets available for knee OA assessment. Despite the availability of a greater number of private datasets, public datasets play a dominant role in establishing benchmark results and facilitating continuous improvement in the field of knee OA research. In this section, we present the publicly accessible datasets for knee OA diagnosis as outlined in Table 4.

Osteoarthritis Initiative (OAI) [60] is an open access dataset provided by the National Institutes of Health (NIH). It focuses on identifying the most promising biomarkers of development and progression of symptomatic knee OA. This dataset includes 4,796 subjects between the ages of 45 and 79 years who either have knee OA or are at an increased risk of developing the condition. The data was collected from four clinical centers (Ohio State University, University of Maryland School of Medicine/Johns Hopkins University School of Medicine, University of Pittsburgh School of Medicine, and Brown University School of Medicine and Memorial Hospital of Rhode Island). Over a period of ten years, all participants underwent annual radiography and MRI scan of the knee, along with clinical assessments of disease activity. Furthermore, genetic and biochemical specimens were collected annually from all participants, providing rich data for researchers to explore novel knee OA diagnosis and treatment approaches.

Multicenter Osteoarthritis Study (MOST) [61] is a public dataset funded by the National Institutes of Health (NIH) and National Institute on Aging (NIA). The primary objective of this dataset is to study symptomatic knee OA in a community-based sample of adults with or at high risk of developing knee OA. About 3,026 subjects between the ages of 50 and 79 years from two clinical sites (Iowa City, Iowa and Birmingham, Alabama) participated the study. The dataset contains essential information related to biomechanical factors (such as physical activity-related factors), bone and joint structural factors (such as knee MRI assessment), and nutritional factors.

MRNet [62] is a collection of MRI data created by the Stanford University Medical Center. This dataset aims to investigate two common types of knee injuries: anterior cruciate ligament tears and meniscal tears which are contributing factors to knee OA disorder. The study involved 1,312 subjects and generated a total of 1,370 MRI scans. The MRI examinations were conducted using GE scanners (GE Discovery, GE Healthcare, Waukesha, WI) with a standard knee MRI coil and a routine non-contrast knee MRI protocol, comprising several key sequences: coronal T1 weighted, coronal T2 with fat saturation, sagittal proton density (PD) weighted, sagittal T2 with fat saturation, and axial PD weighted with fat saturation. Among the knee examinations, about 56.6% were performed using a 3.0 Tesla magnetic field, while the remaining used a 1.5 Tesla magnetic field. Furthermore, the authors provided a benchmark MRNet single model, intended to support further research endeavors in the field.

FastMRI+ [63] is a publicly available MRI dataset that extended the work of the FastMRI dataset [64]. This extended dataset includes 1,172 MRI scans acquired at 1.5 or 3.0 Tesla and provides 22 different pathology labels in knee anatomical areas such as bone, cartilage, ligament, meniscus, and joint. Notably, many of the pathologies, such as cartilage loss and joint effusion, are closely related to knee OA. Each knee MRI scan comprises a single series of coronal images in PD or T2-weighted sequence. The primary focus of the FastMRI+ dataset is to facilitate the study of MRI image reconstruction, particularly in regions that could potentially contain clinical pathology. This dataset provides detailed pathology labels, researchers can explore and develop advanced image reconstruction techniques that cater to specific clinical conditions.

Cohort Hip and Cohort Knee (CHECK) [65], [66] is a research initiative sponsored by the Dutch Arthritis Foundation, in collaboration with ten general and university hospitals in The Netherlands, situated in semi-urbanized regions. The study recruited a total of 1,002 subjects aged between 45 and 65 years. The primary goal of this dataset is to explore and analyze the clinical, biochemical, and radiographic signs and symptoms associated with early OA. Moreover, the dataset aims to identify prognostic factors that may contribute to the diagnosis and progression of OA. The study spans a duration of seven years, during which 846 subjects actively participated in annual clinic visits, providing valuable longitudinal data for comprehensive OA research.

Private research at Danderyd University Hospital is used in [67] to develop a predictive model for classification of OA stage. The dataset consists of 6,103 X-ray images acquired from Danderyd University Hospital. Unlike other datasets that undergo extensive preprocessing for artifact removal, this dataset used the entire image series, including X-ray images with visual disturbances like implants, casts, and non-degenerative pathologies. This unique approach provides a more realistic representation of clinical scenarios and enhances the dataset's value for studying OA progression and prediction in real-world conditions.

Mendeley VI is a unique public dataset that focuses on the Eastern population. It contains 1,650 X-ray images collected from Indian institutions. The X-ray images were captured using the PROTEC PRS 500E X-ray machine. All images are 8-bit grayscale and have been cropped to focus on the cartilage region. They have been manually annotated by two experienced medical experts with their respective Kellgren and Lawrence grades. The intention of this dataset is to facilitate in the development of AI models for classifying osteoarthritis severity.

Private research at Chang Gung Memorial Hospital comprises a small dataset of 400 X-ray images collected from the Taiwanese population. These 8-bit grayscale images have been manually annotated by a doctor with their respective Kellgren and Lawrence grades. The dataset is intended to aid in the development of AI models for classifying the severity of radiographic OA.

D. CLASSIFICATION SYSTEMS FOR KNEE OA CONDITIONS

In this section, we conducted a comparison of the employment of classification systems from the medical domain that establish the ground truth data for predictive models. Approximately half of the studies (40 out of 70) utilized medical experts' knowledge for classification or clustering tasks. Within this subset of studies, 34 employed Kellgren Lawrence (KL) grading system to rate the OA severity. Original Kellgren Lawrence (KL) grading system comprises five ordinal classes based on composite score of radiographic OA symptoms. However, the number of classes used in the top layer of the KL prediction models varied from two to five across the reviewed studies, depending on their respective research purposes. A commonly used standard threshold for radiological OA is a KL≥2. Most of the studies (23 out of 34) were dedicated to developing AI models specifically for the five-grade KL classification. Binary classification was designed to identify the presence of OA (KL1 to KL4) (4 out of 34) or early OA (KL2) (6 out of 34). In addition, one study classified the change in KL grade after 60 months.

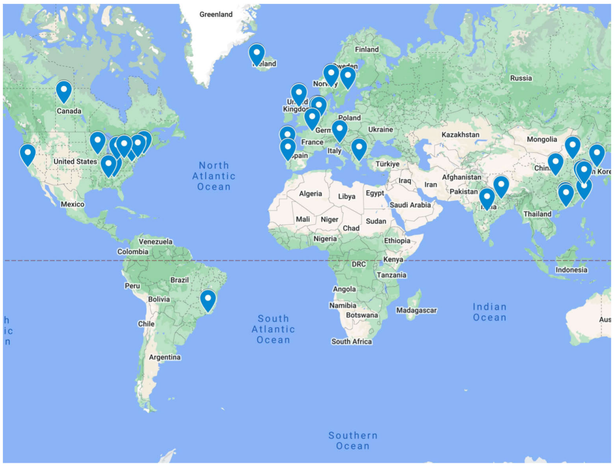


FIGURE 5. Geographical distribution of OA data sources.

Besides KL grading, there was one study employed Osteoarthritis Research Society International (OARSI) atlas joint-space narrowing for medial tibiofemoral OA. Another research developed radiographic spiking criteria to guide the generation of ground truth data [100]. Whole-Organ Magnetic Resonance Imaging Score (WORMS) (n = 1) and MRI Osteoarthritis Knee Score (MOAKS) (n = 1) were employed for knee OA detection on MRI data. Both classification systems emphasized cartilage damage related to OA.

In contrast, patient-reported outcome measure were only used in four studies. Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) (n = 2) and the Knee Injury and Osteoarthritis Outcome Score (KOOS) (n = 1) were frequently employed as patient-reported outcome measures, particularly for assessing pain. However, due to their subjective nature, the analysis process for these measures was complex and required careful statistical analysis [76]. Moreover, transforming the data from these measures into a format suitable for modeling presented a significant challenge. Reference [101] established a direct binary classification system for chronic knee pain based on patient self-reporting. They defined chronic knee pain as pain that persists for more than half of the days in a month for at least six out of the past 12 months. Moreover, two studies utilized knowledge-based and patient-based outcome measures [71], [95].

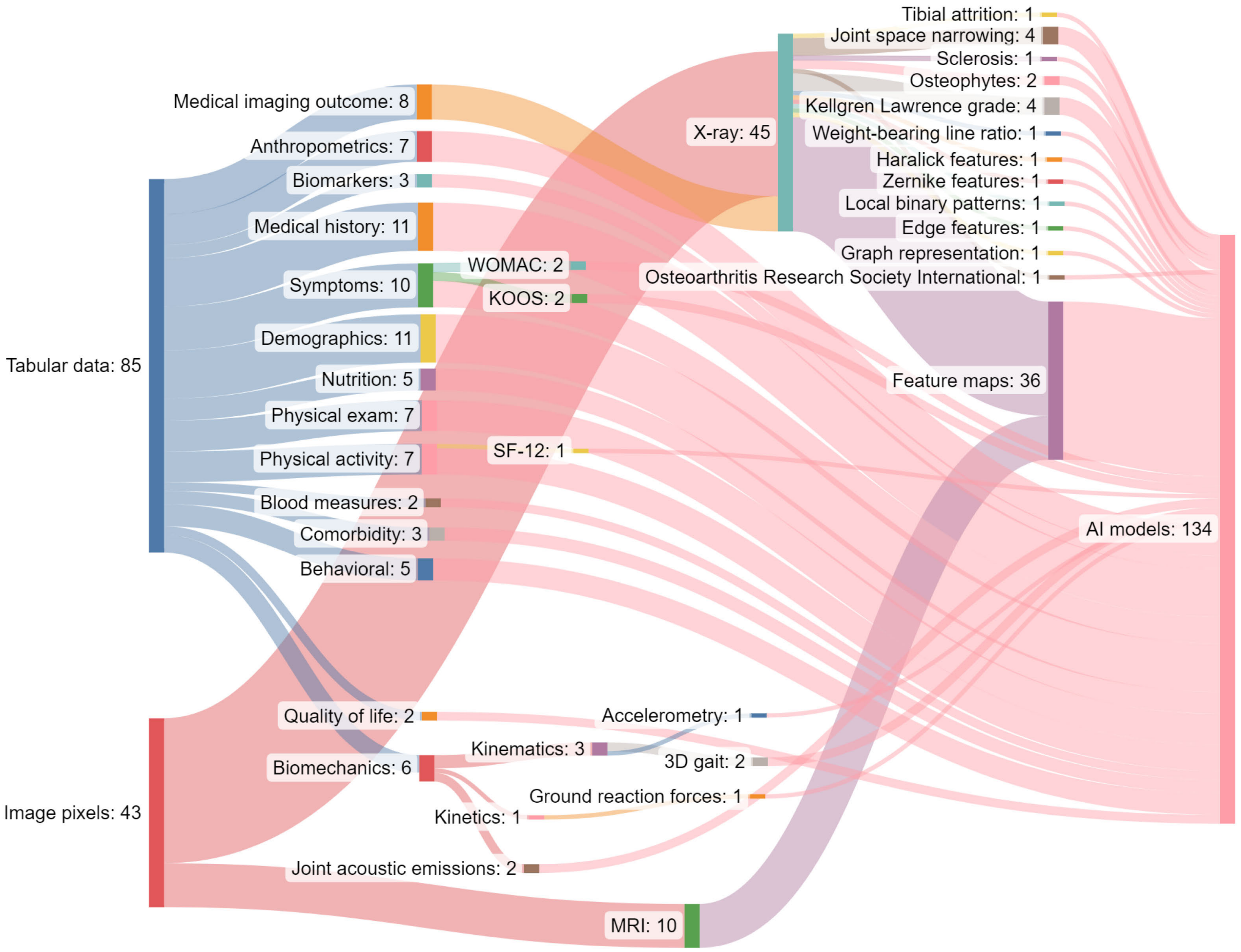


FIGURE 6. Categorization and distribution of input data for AI models.

TABLE 3. Comparison of input data types in knee OA diagnosis. Threshold for Accessibility - High: easily acquired; Moderate: requires equipment and technician; Low: necessitates equipment and technician, available in a limited number of labs.

Input data type	Accessibility	Cost	Knowledge required	References
Demographic	High	Low	No	[68, 5, 69, 70, 71, 72, 66, 73]
Clinical	High	Low	Yes	[68, 5, 69, 70, 74, 75, 76, 77, 78, 79, 71, 72, 80, 66, 73, 81]
Imaging	Moderate	High	Yes	[82, 5, 69, 77, 83, 71, 75, 78, 66, 72, 84, 73, 85, 86, 87, 88, 89, 90]
Patient-Reported Outcomes	Moderate	Low	No	[5, 75, 71, 78, 79, 72, 66, 91]
Biomechanics	Low	Low	Yes	[92, 70, 93, 94, 95, 96, 97, 91]
Biomarkers	Low	High	Yes	[68, 70, 71, 98, 73, 81]

V. XAI APPROACHES FOR KNEE OA ASSESSMENT

The role of XAI in knee OA assessment is to offer comprehensible explanations regarding the input data. These explanations are intended to be understood by humans. Thus, we take into account the interests of data scientists and domain experts in the development of XAI methods.

For data scientists, knowing the internal workings of the model and comprehending how the data is applied are crucial for improving the model's performance and preventing overfitting. This knowledge enables them to fine-tune the model, optimize its architecture, and make informed decisions during the development process. Post-hoc

TABLE 4. List of open access OA-related data sources and descriptions.

Data source (website)	Year of data	No. of subjects (age range)	Data types	Geographic representation
Osteoarthritis Initiative (OAI) [60] (https://nda.nih.gov/oai/)	2004-2015	4,796 (45-79)	Image data, tabular data at baseline, and 12, 24, 36, 48, 60, 72, 84, 96, and 108 months	USA
Multicenter Osteoarthritis Study (MOST) [61] (https://most.ucsf.edu/)	2003-2018	3,026 (50-79)	Image data, tabular data at baseline, 15, 30, 60, 72, and 84 months	USA
MRNet [62] (https://stanfordmlgroup.github.io/competitions/mrnet/)	2001-2012	1,312 (-)	Image data: 1,370 MRI images	USA
FastMRI+ [63] (https://github.com/microsoft/fastmri-plus and https://fastmri.med.nyu.edu/)	N/A	- (-)	Image data: 1,172 coronal MRI scans either proton density-weighted and T2-weighted	USA
Cohort Hip and Cohort Knee (CHECK) [65, 66]	2002-2012	1,002 (45-65)	Image data and tabular data at baseline, 2, 5, 8, and 10 years	The Netherlands
Private research at Danderyd University Hospital [67] (https://datahub.aida.scilifelab.se/10.23698/aida/koa2021)	2002-2016	- (-)	Image data: 6,403 X-ray images	Sweden
Mendeley V1 [99] (https://data.mendeley.com/datasets/t9ndx37v5h/1)	N/A	- (-)	Image data: 1,650 X-ray images	India
Private research at Chang Gung Memorial Hospital [88] (https://www.kaggle.com/datasets/tommyngx/cgmh-oa)	N/A	400 (-)	Image data: 400 X-ray images	Taiwan

explanations may be of lesser concern to them, as they prioritize optimizing the model itself.

On the other hand, domain experts especially medical experts who may not have the technical expertise of data scientists are more interested in understanding how and why a model generated a particular result. They seek clear and interpretable explanations to trust the model's decisions and insights. Knowing the key characteristics that led to a conclusion helps them validate the model's outputs and make informed decisions based on the AI system's recommendations.

By considering the specific needs and interests of both data scientists and domain experts, we propose the XAI taxonomy as shown in Figure 7 to provide valuable insights into the diverse requirements of different stakeholders. Understanding data interpretability, model interpretability, and post-hoc interpretability, along with XAI evaluation approaches, is crucial in building transparent, trustworthy, and effective AI models that cater to various real-world applications.

A. DATA INTERPRETABILITY

The importance of data interpretability arises from the substantial impact of the training dataset on an AI model's behavior. To facilitate a better understanding of the input data, numerous data analysis techniques and mathematical algorithms have been developed to quantify the intrinsic data characteristics. In the context of knee OA, data interpretability can uncover valuable clinical patterns that might not have been captured in traditional evidence-based research. This can empower the researchers to glean new insights and knowledge from the data, contributing to more informed and effective decision-making in knee OA assessment.

In the following sections, we will discuss a few approaches that provide interpretability for knee OA data. This includes feature extraction, explainable feature engineering, and knowledge graphs, which are widely recognized as pre-modelling approaches. These approaches help extract useful information from the data and represent different steps in achieving data interpretability. Feature extraction extracts relevant features, explainable feature engineering transforms data for better understanding, and knowledge graphs connect related points for a comprehensive disease overview.

1) FEATURE EXTRACTION

Feature extraction plays a critical role in capturing a representative set of features. In our survey, we found three types of feature extraction: exploratory data analysis, image descriptors, and dimensionality reduction.

a: EXPLORATORY DATA ANALYSIS

Exploratory data analysis (EDA) is a data analysis approach that involves summarizing the main characteristics of the data and visualizing the data summary using appropriate representations [102]. EDA is an essential process for understanding the structure and distribution of tabular data, as well as identifying important features and patterns that

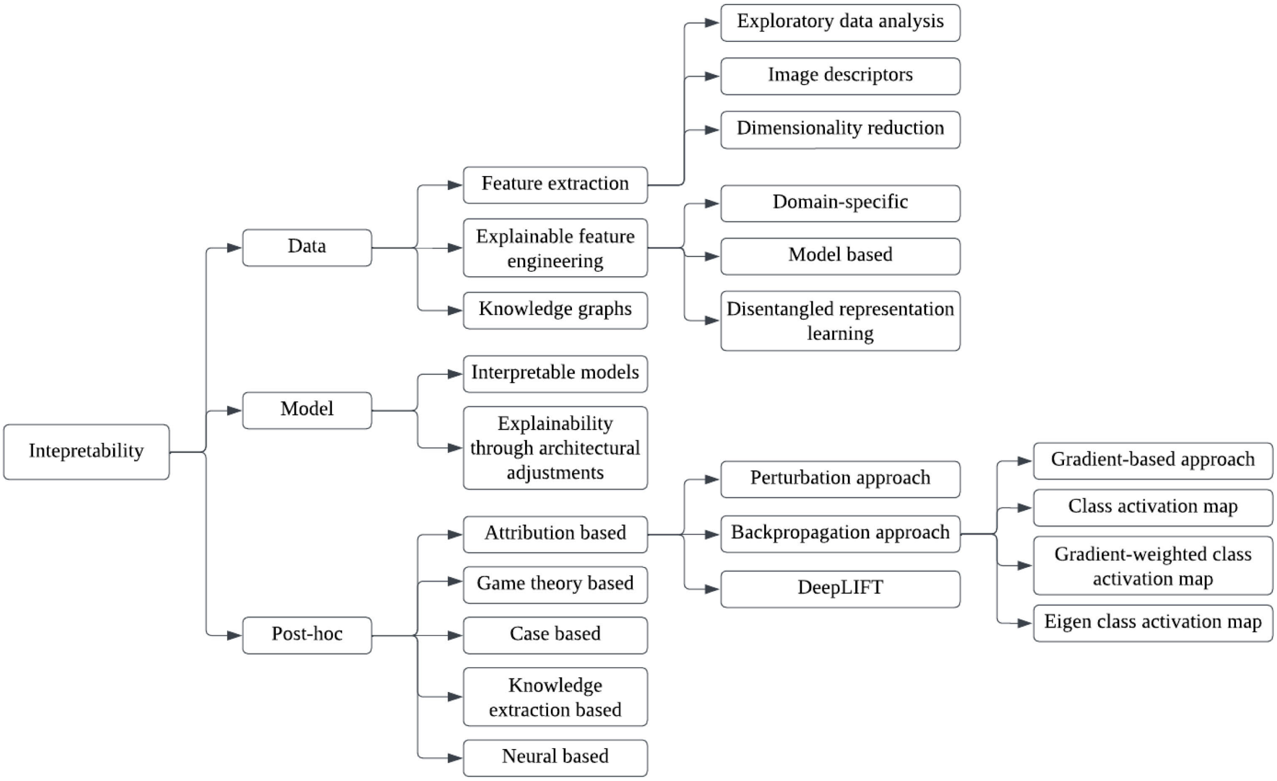


FIGURE 7. Proposed XAI taxonomy.

can guide subsequent analysis. The significant contribution of EDA in knee OA data is analysis of population-based samples to provide the disease overview and to detect biases in data [98].

General EDA outcomes included dimensions, mean [88], median [98], standard deviation [88], range, and missing samples. To deal with missing samples, [98] implemented imputation models using random forest (RF) and k-nearest neighbor (KNN) regression models, which were further refined through a bootstrapping-like procedure. Reference [103] performed an analysis on the brightness value distributions in the lateral and medial sides of the OA and control groups. The results showed that the mean brightness in the OA group was higher than in the control group in both sides. The observation suggested that the higher mean brightness may be indicative of increased bone density in patients from the OA case group.

b: IMAGE DESCRIPTORS

Image descriptors are typically used to capture and describe the shape of an object in an image. Jakaite et al. [103] utilized Zernike moments (Equation 1) to capture knee X-ray textural details at the bone microstructural level. By using this image descriptor and the Group Method of Data Handling (GMDH), they were able to effectively identify patients at risk of early knee OA, even with a relatively small dataset of 40 samples.

The Zernike moments A_{nm} are defined as:

$$A_{nm} = \frac{n+1}{\pi} \sum_{x} \sum_{y} f(x, y) V^{*}_{nm}(\rho, \theta) \qquad (1a)$$

where $f(x, y)$ represents the image, n denotes the number of order, m denotes the number of repetition, V_{nm} represents orthogonal complex polynomials, ρ represents the length of a vector to a (x, y) pixel, and θ represents the angle between x-axis and ρ.

The orthogonal complex polynomials V_{nm} are defined as:

$$V_{nm}(x, y) = V_{nm}(\rho, \theta) = R_{nm}(\rho)e^{(-jm\theta)} \qquad (1b)$$

where R_{nm} represents radial polynomials, ρ represents the length of a vector to a (x, y) pixel, and θ represents the angle between x-axis and ρ.

The radial polynomials R_{nm} are defined as:

$$R_{nm} = \sum_{k=0}^{(n-|m|)/2} \frac{(-1)^k (n-k)!}{k!((n+|m|)/2-k)!((n-|m|)/2-k)!} \rho^{n-2k} \qquad (1c)$$

where n denotes the number of order and m denotes the number of repetition.

A more detailed bone analysis work was performed by Bayramoglu et al. [104]. The authors conducted a comparison of five image descriptors: local binary pattern (LBP), fractal dimension (FD), Haralick features, Shannon entropy, and histogram of gradient (HOG). Based on their findings, they recommended the use of LBP as it preserved the most discriminative features among the descriptors. They also pointed out that LBP and HOG descriptors are less sensitive to changes in radiographic acquisition protocols and could be applied in clinical decision support tools in the future.

Fatema et al. [88] utilized six sets of features to describe the geometry of knee X-rays. These included morphological features, gray level co-occurrence matrix (GLCM) features, statistical features, texture features, LBP features, and a novel set of features inspired by unique characteristics of knee bone structures. The proposed features computed several key metrics: the vertical distance between femur and tibial bones, joint space area, peak curvature angle at the femur center, and gradient features representing both magnitude and direction derived from function derivatives in horizontal and vertical directions. Computation was compartmentalized for medial and lateral sides. Heatmap analysis of feature correlations revealed strong correlations in bone distances across compartments, except between the middle of lateral knee segments and the medial side of medial knee segments. Moderate positive correlations were observed between the joint space areas of the lateral and medial compartments. Notably, gradient features at the lateral and medial knees exhibited negative correlations with several distance and peak features, suggesting an inverse relationship. The analysis also indicated weaker or negative correlations between the medial and lateral compartments, highlighting distinct behavioral characteristics between these compartments.

A range of texture features were extracted from knee MRI images [86], including first-order statistical features, GLCM, gray level size zone matrix (GLSZM), gray level run length matrix features (GLRLM), gray level dependence matrix (GLDM), and neighboring gray tone difference matrix (NGTDM) to characterize bone and cartilage textures. Feature extraction was conducted from 32 anatomical sub-regions identified as potentially influential in knee OA progression [87] at three time points: baseline, 12 months, and 24 months. These features served as input for deep learning models. Due to the extensive feature set and the risk of spatial heterogeneity, the Minimum Redundancy Maximum Relevance (mRMR) and Least Absolute Shrinkage and Selection Operator (LASSO) algorithms were employed to select the optimal feature sets. However, this approach was not applicable to severely damaged cartilage because the texture patterns in such cartilage were too degraded to provide consistent and reliable feature measurements. In these cases, the variability introduced by the damaged tissue could obscure meaningful patterns, leading to less accurate and less robust feature extraction.

Besides texture analysis, image descriptor could be used to extract object edge information. Adaptive Canny algorithm was employed to extract the edges of the knee joint from X-ray images by dynamically adjusting the threshold values based on the local image characteristics [105]. The low α and high adaptive thresholds β are defined as:

$$\alpha = \max\left(0, (1-\sigma) \times \text{median}(x_i)\right) \tag{2a}$$

$$\beta = \min\left(255, (1+\sigma) \times \text{median}(x_i)\right) \tag{2b}$$

where α denotes upper limit pixel value, β denotes bottom limit pixel value, and x_i represents median pixel value.

c: DIMENSIONALITY REDUCTION

OA datasets are typically complex and multidimensional, containing a vast amount of variables. Visualizing such high-dimensional data can be challenging since human perception is limited to three dimensions. Hence, researchers tend to find lower-dimensional representations of the original data [40]. Dimensionality reduction techniques are employed to reduce the number of parameters while preserving the underlying structure as much as possible. Two commonly used methods in this field are Principal Component Analysis (PCA) [98] and t-Distributed Stochastic Neighbor Embedding (t-SNE) [25], [27], [71], [106], [107].

Nielsen et al. [81] performed OA patient clustering and characterization using the Louvain clustering algorithm based on SHapley Additive exPlanations (SHAP) values. A total of 12 clusters were obtained after dimensionality reduction using PCA (10 PCs) and further visualized with Uniform Manifold Approximation and Projection (UMAP).

2) EXPLAINABLE FEATURE ENGINEERING

There are two main approaches developed for explainable feature engineering: domain-specific methods and model-based methods. Another emerging approach in explainable feature engineering is disentangled representation learning, which has gained traction with the introduction of various generative models.

a: DOMAIN-SPECIFIC

Domain-specific approaches for knee OA diagnostic task utilize the knowledge and expertise of medical experts, along with insights derived from EDA to extract features. Many studies in this field have focused on developing knee-specific approaches that capture and characterize key aspects on bone and cartilage, as well as the limb alignment.

Reference [108] developed a measure called cartilage damage index (CDI) to quantify cartilage thickness by measuring specific informative locations on the reconstructed cartilage layer instead of evaluating the entire cartilage. In a cartilage assessment conducted by Ciliberti et al. [84], two volumetric analyses were employed. The first analysis focused on wall thickness, where the cartilage mesh was examined, and the thickness of each element was calculated from surface to surface. The hypothesis underlying this analysis was that patients with degenerative and traumatic cartilages would exhibit thinner cartilage in specific regions compared to the control group. The second analysis focused on cartilage curvature by measuring the Gaussian curvature of cartilage element based on its neighboring elements. This analysis hypothesized that areas with higher cartilage degradation would exhibit increased curvature due to the formation of holes and depressions surrounding those regions. In [101], cartilage thickness was determined for femoral, tibial, and patellar cartilage masks per sagittal slice by performing an Euclidean distance transform along the morphological skeleton of each mask. Furthermore, the shape of the bone was characterized by measuring the distance from the bone surface of each bone mask to its volumetric centroid.

A recent study by Zhuang et al. [109] proposed a unified graph representation approach to construct personalized knee cartilages that are attached to the femur, tibia, and patella, respectively. They used the patient-specific cartilage graph representation to guide their DL model. Additionally, to assess the coronal limb alignment through radiographic means, weight-bearing line (WBL) ratio was derived by calculating the ratio between the crossing point of the mechanical axis, measured from the medial edge of the tibial plateau, and the total width of the tibial plateau [110]. Recognizing the multi-compartment nature of the knee joint, Hu et al. [86] employed an undirected regional adjacency graph to represent the interrelations among various knee sub-regions. In this graphical representation, nodes corresponded to distinct knee MRI-extracted structures such as bone and cartilage, while edges denoted their anatomical connections. For instance, given the connection between the tibia and femur through the meniscus, edges were assigned between the meniscus and nearby cartilage. Similarly, connections between the medial and lateral tibiofemoral joints and the patellofemoral joints were determined by their direct structural contacts. This approach resulted in the creation of four anatomic compartment-based regional graphs, each capturing the intricate anatomical relationships within the knee joint.

b: MODEL BASED

Model-based feature engineering leverages an automatic approach to unveil the inherent structure of a dataset, leading to the extraction of relevant and informative features [40]. One such example is unsupervised clustering, a technique that groups similar data points together based on their intrinsic characteristics, without the need for labelled target variables [40]. For instance, [101] developed a fully automatic landmark-matching algorithm based on Coherent Point Drift to map the bone surfaces into reference space. Reference [104] used simple linear iterative clustering based superpixel segmentation to extract the region of interest as a pre-processing strategy. Reference [98] applied k-means clustering to analyze biochemical marker data and figure out prominent subgroups among patients with OA. This approach enabled them to identify three dominant OA phenotypes. Reference [72] conducted similar work using biclustering, but their work was extended to more inclusive clinical data, including demographics, medical history, symptoms, physical activity, physical exam, and medical imaging outcome. Through their analysis, they identified two significant clusters. One cluster represented individuals who exhibited structural progression over time but experienced improvements in pain. The other cluster represented individuals who had stable pain scores and were less affected by OA. Additionally, model-based feature engineering techniques were employed to analyze gait data, which is known for its complexity with multidimensional and time-series properties. Leporace et al. [93] utilized self-organizing maps (SOM) on principal components to detect gait similarity patterns in individuals with high grade OA. The resulting patterns were visualized using a unified distances matrix (U-matrix). Subsequently, the U-matrix was subjected to the k-means clustering algorithm, leading to the formation of four distinct gait kinematic clusters. Liu et al. [111] employed a feature optimization learning method, namely the integrated multi-modal learning method (MMLM) to enhance the classification of early-stage knee OA. Their study addressed the challenges posed by the complexity and high dimensionality of multi-modal data, which includes clinical, imaging, and demographic features. By utilizing L1-norm-based optimization, they were able to regularize and select the most relevant features from each modality, effectively reducing the overall dimensionality of the data, thereby improving the performance of machine learning classifiers in detecting early knee OA. Consequently, MMLM not only tackles the curse of dimensionality but also enhances interpretability and efficiency in the diagnostic process.

c: DISENTANGLED REPRESENTATION LEARNING

Disentangled representation learning is a significant and closely related area of research that focuses on acquiring a dataset representation where the generative latent variables are disentangled or separated. Latent variables in this context can be regarded as interpretable or explainable features of the dataset. Reference [112] were pioneers in applying the DeepFake concept in this medical domain, specifically by utilizing Wasserstein generative adversarial neural networks with gradient penalty (WGAN-GP). Their model managed to preserve important OA anatomical information during the generation process. The authors utilized DeepFake generated data to substitute real data during the training of a pre-trained VGG model for classification task. Remarkably, they achieved a mere 3.79% decrease in accuracy compared to the baseline when classifying real OA X-rays. Reference [113] advanced the field of disentangled representation learning by introducing a novel approach called key-exchange convolutional autoencoder (KE-CAE). This method was designed to extract specific radiograph features related to early knee osteoarthritis (OA) from latent space through cross image reconstruction. Their proposed approach successfully captured crucial information from radiographs that represents early knee OA, enabling effective analysis. Notably, their model not only achieved high-quality reconstruction of the original images but also generated synthetic images that accurately represented different stages of knee OA. This noteworthy contribution holds promise for the early detection and diagnosis of knee OA.

3) KNOWLEDGE GRAPHS

Knowledge graph (Figure 8) is a structured representation of knowledge that captures relationships between entities in a particular domain. Reference [74] established a medical knowledge graph using unstructured data from an electronic medical record (EMR) database. The EMR data was in Mandarin language, which posed a computational challenge for processing Mandarin words. To address this, the authors adopted five feature extraction methods, including bi-directional long short-term memory (Bi-LSTM), bag of characters, natural language processing with the Chinese Academy of Sciences Word Segmentation Tool, dictionary features, and the k-means algorithm for word clustering. These diverse feature sets were utilized in the conditional random field (CRF++) algorithm for entity recognition. Following entity recognition, the authors combined the extracted features and implemented a CNN model, comprising a convolutional layer, pooling layer, fully connected layer, and softmax classifier, to extract entity relations from the identified entities. This step allowed for a deeper understanding of the interconnections within the medical data. In the final stages of building the medical knowledge graph, the authors employed Neo4j graph database. They achieved this by batch importing the previously identified medical entities and their corresponding relationships into the Neo4j database, forming a comprehensive and interconnected representation of medical knowledge in knee OA domain. The resulting knowledge graph encompassed 2,518 distinct entities and an impressive 29,972 different relationships related to knee OA condition. The knowledge graph spans a diverse range of entity types, comprising 368 diseases, 706 symptoms, 421 treatments, 859 examination descriptions, 72 examinations, 43 aggravating factors, 35 mitigating factors, and 14 inducing factors. This comprehensive repository of information serves as a valuable resource, empowering researchers and medical practitioners to gain deeper insights into knee OA.

B. MODEL INTERPRETABILITY

While clean and carefully prepared data, aided by data interpretability techniques, is crucial for training models, it is equally important for the model itself to possess a clear understanding. Without this understanding, developers may face challenges when incorporating their domain knowledge into the learning process to achieve improved results. Therefore, alongside data interpretability, model interpretability plays a vital role.

In many instances, analyzing outputs or examining individual inputs is insufficient for comprehending why a training procedure failed to yield the desired outcomes. In such cases, it becomes necessary to investigate the training procedure in the model. The objective of model explainability is to develop models that are inherently more interpretable and understandable. This approach is also called intrinsic XAI.

1) INTERPRETABLE MODELS

Interpretable models, also known as white-box models, are models that provide self-explanatory insights [44]. Examples of such models include rule-based model, linear regression, logistic regression, and decision trees.

In the realm of rule-based models, [76] devised an objective algorithm for pain prediction and compared it to a general KL grade-based algorithm. Their proposed algorithm incorporated racial disparities (Black versus non-Black) and two socioeconomic measures, namely annual income below $50,000 and educational attainment (college graduation). The authors examined the differences in pain scores between groups and quantified the pain disparities using non-parametric means. By employing a regression model, the proposed algorithm successfully addressed the inequalities faced by under-served patients.

Linear regression was utilized when researchers assumed a linear relationship between the severity of OA disease and the KL grade. Nichols et al. [91] employed Least Absolute Shrinkage and Selection Operator (LASSO) regression on knee acoustic emission features to predict the KL grade. The predicted KL grade was then combined with Knee Injury and Osteoarthritis Outcome Score (KOOS) scores and used to determine the stage of OA, categorizing it as either early or late, using a linear discriminant analysis model. This two-stage classification approach achieved improved balanced accuracy and area under a receiver operating characteristic curve (ROC-AUC) over the models using single-stage or single-input methods. The proposed approach demonstrated the enhanced model's intrinsic interpretability for knee evaluation, despite the use of subjective patient-recorded data (KOOS) as input.

Zeng et al. [95] utilized binary logistic regression to detect knee OA and recommend appropriate treatment options, including conservative or surgical approaches. Although the authors claimed the interpretability of their model, however they did not provide detailed analysis or explanations to support their claim.

In terms of tree-based models, Kotti et al. [92] utilized a regression tree to analyze and interpret the rule induction process for detecting OA cases from a biomechanical perspective. A random subset of parameters extracted from ground reaction forces in the z-axis was employed to construct the regression tree, as illustrated in Figure 9. This approach provided insights into the biomechanical factors that may contribute to the presence of OA and offered a means of interpreting the rule induction process in the context of OA detection.

[68] employed splitting nodes algorithm to assess the importance of each feature in a tree generated by eXtreme Gradient Boosting (XGBoost). They found that

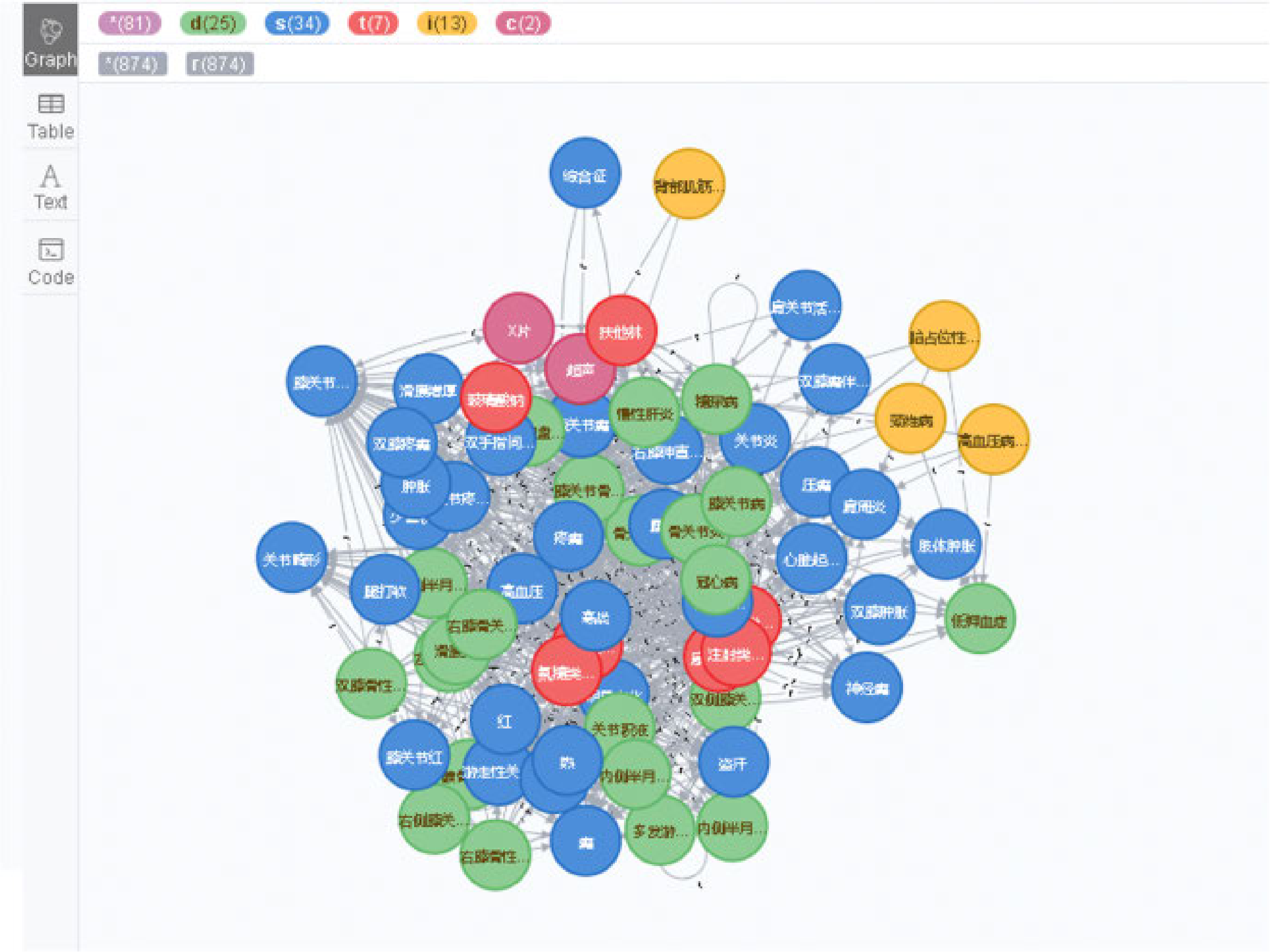


FIGURE 8. Example of knowledge graph for knee OA. Each circle represents a specific Chinese term associated with knee OA. Adapted from [74].

demographics and anthropometric factors had a significant influence on determining OA status, but acknowledged that these factors are not exclusive to OA and contribute to various clinical issues like pain and disability. Instead, the authors emphasized three other categories: comorbidity, blood measures, and physical activity measures. These categories were closely linked to the risk of experiencing side effects from analgesics in OA patients.

Despite the use of white-box mechanisms, relying solely on interpretable models may not provide sufficient explanation for complex models, particularly in scenarios with high-dimensional and heterogeneous data. To address this limitation, the application of regularization techniques becomes necessary during model training. Regularization helps control the number of relevant input features by introducing penalties or constraints, ensuring that the model focuses on the most important variables. For example, [70] used a robust methodology to process 707 features from multidisciplinary settings. They employed six feature selection techniques, including filter algorithms, wrapper approach, and embedded techniques, and ranked features based on a majority vote scheme. This process identified 40 relevant risk factors, resulting in a classification accuracy of 77.88% using logistic regression.

However, one limitation of the majority voting approach is that it treats all models in the ensemble equally, without considering the possibility of weak predictions. To address this limitation, [79] introduced a Fuzzy ensemble approach to optimize the model and improve decision-making by considering the reliability and uncertainty of individual predictions. Additionally, [73] demonstrated the effectiveness of recursive feature elimination (RFE), which considers the intrinsic characteristics of the data and model to select an optimal feature combination. RFE iteratively eliminates less relevant features, resulting in an informative subset that contributes significantly to the model's performance.

2) EXPLAINABILITY THROUGH ARCHITECTURAL ADJUSTMENTS

Attention mechanisms could introduce certain level of explainability and have revolutionized the utilization of DL algorithms [28], [77], [85], [109], [114], [115], [116]. Zhang et al. [28] utilized the convolutional block attention

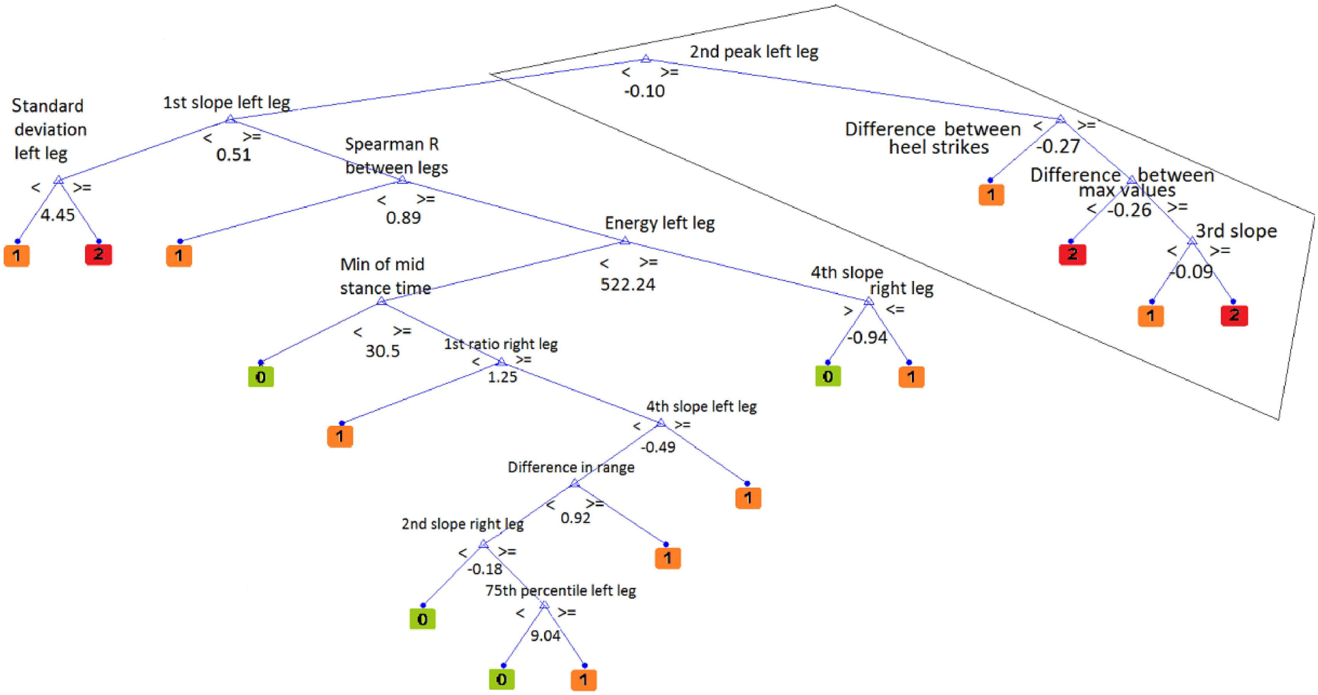


FIGURE 9. Example of regression tree. Adapted from [92].

module (CBAM) to implement an attention mechanism. Their CBAM consisted of both channel and spatial attention modules. By incorporating the module into ResNet34, the proposed approach identified the most relevant channel and spatial parts that contributed significantly to the final prediction and helped to enhance the model's performance by focusing on the most informative features in the input data during the training process. Despite the improved model performance in terms of accuracy, MSE, and Kappa coefficient, the integration of CBAM at the end of each residual block significantly increased the computational burden, making real-time deployment challenging. Jain et al. [85] addressed this issue by proposing a single integration of CBAM into a multi-resolution deep CNN, specifically the High-Resolution Network. This model effectively captured multi-scale information and enabled more efficient adaptive filtering of counterproductive features, thereby reducing the computational load while maintaining high performance.

In the study conducted by Feng et al. [115], the channel attention module within the CBAM was enhanced by incorporating additional non-linear layers after fusing the channel weights from dual branches. This modification increased the expressiveness of the CBAM network and improved the model's accuracy in detecting potential lesions in knee X-ray images. Similarly, [77] incorporated a gated attention mechanism to calculate attention scores for individual image slices, which can be interpreted as indicators of their importance. These scores were then utilized in the classification sub-model. Self-attention mechanism was implemented by Wang et al. [114] by integrating a visual transformer after their deep learning model. Their approach effectively captured the interrelationship among imaging features from multiple regions.

Alternatively, [109] proposed a self-attention-based network, namely CSNet that has been designed in a layer-by-layer manner. Each layer incorporated patch convolution to extract local appearance features from individual vertices and graph convolution to facilitate communication among the vertices. The self-attention mechanism was employed in each layer to enhance the model's ability to capture information from the cartilage graph. The final assessment of knee cartilage defects was obtained by pooling information from all vertices in the graph, and the CSNet also allowed for easy 3D visualization of the defects, showcasing its interpretability. In contrast to previous approaches that did not take semantic information into account, a recent study by Huo et al. [116] introduced the use of an online class activation mapping (CAM) module to specifically direct the network's attention towards the cartilage regions.

C. POST-HOC INTERPRETABILITY

Post-hoc XAI methods were found to be more commonly employed than intrinsic XAI methods in those studies. These post-hoc methods provide an external explanation of the AI model's decisions after it has made predictions. It involves querying the trained model and constructing a white-box surrogate model to extract the underlying relationships the model has learned [40]. These methods could gain insights into the model's decision-making process by analyzing its predictions on specific instances, without altering the original model architecture. In contrast, intrinsic XAI focuses on designing AI models with inherent interpretability right from the model's architecture and design [44]. These models are built with specific structures or components that naturally provide transparency and understandability in their decision-making process.

1) ATTRIBUTION BASED

a: PERTURBATION APPROACH

Perturbation is a simple and effective method for computing the impact of changing input features on the output of an AI model [16]. It involves manipulating certain input features, running the forward pass, and measuring the difference from the original output. The importance of the input features can be ranked based on their effect on the output. Reference [76] demonstrated a region-wise method to visualize image areas that influenced predictions made by a neural network. To do this, the image regions are "masked" out by replacing them with a circle, and the value of the circle was set to the mean pixel value for the image. Gaussian smoothing was applied to prevent sharp boundaries. The neural network's predicted pain score was then compared between the masked image and the original image, and the absolute change in the predicted pain level was computed. This process was repeated for a 32 × 32 grid of regions that evenly tiled the 1024 × 1024-pixel image. This has allowed a heatmap analysis that revealed how much masking each region of the image affected the neural network's prediction.

Another perturbation approach is GNNExplainer. This technique is used to identify the nodes that are most critical to the performance of a Graph Neural Network (GNN). It operates by removing or altering nodes, features, edges, or subgraphs to determine which changes significantly impact the model's final decision and its confidence in that decision. For each node, GNNExplainer generates a score, S, ranging from 0 to 1, with higher values of S indicating greater node importance. Hu et al. [86] used GNNExplainer to assess the contribution of various MRI subregions. The analysis highlighted the importance of cartilage compared to other structures and revealed the dominance of the tibiofemoral joint over the patellofemoral joint for knee OA diagnosis. The proposed model also tracked changes in importance scores within compartments over time from baseline to 12 months and then to 24 months.

b: BACKPROPAGATION APPROACH

Backpropagation approach can be further divided into gradient-based approach, class-activation map, and gradient-weighted class activation map. Nearly half of the studies (34 out of 70) employed backpropagation XAI approach to explain the predictions of AI models as described in Supplementary Table 1.

Gradient-based approach focus solely on the gradient information when assessing the impact of modifying a specific pixel on the final prediction. Integrated Gradients is a specific technique within this approach. Another technique, namely SmoothGrad was introduced with the intention of reducing visual noise [117] and used by Tack et al. [83] for MRI study.

Class activation map (CAM) utilize global average pooling to compute the spatial average of feature maps in the final convolutional layer of a CNN [118]. Three studies [116], [119], [120] that focused on MRI data used CAM approach to analyze the prediction outcomes.

Gradient-weighted class activation map (GradCAM) is an extension of CAM, which is a technique that does not depend on a specific architecture. The principle of GradCAM is based on the concept of gradient-based CAM. It leverages the gradients of the final convolutional layer with respect to the predicted class to understand which parts of the input image are crucial for making that prediction. Around 34.3% of the studies (24 out of 70) utilized GradCAM for visualizing the final predictions. Another technique, namely GradCAM++ enhances GradCAM by substituting the globally averaged gradients with a weighted average of the gradients at the pixel level. This adaptation takes into account the significance of individual pixels in influencing the final prediction, resulting in more effective visual interpretations of CNN model predictions. GradCAM++ effectively overcomes the limitations of GradCAM, particularly in scenarios involving multiple instances of a class in an image.

Eigen class activation map (Eigen-CAM) is a variation of CAM that incorporates the use of principal components of the learned convolutional activations [121]. It offers more accurate localization of important regions in an image and provides a deeper understanding of the underlying features. Reference [122] employed Eigen-CAM as a tool for localizing osteoarthritis (OA) features in X-ray images, using the Kellgren Lawrence grading scheme. The application of Eigen-CAM revealed significant findings, specifically highlighting the medial and lateral margins of the knee joint. These highlighted regions correspond to joint space narrowing and osteophytes sign, offering valuable insights into the presence and severity of OA-related changes in the knee joint.

c: DEEPLIFT

DeepLIFT was used by Chan et al. [71] to quantitatively assess the contribution of each risk factor to the model's prediction. The assessment was carried out by computing the relative backpropagated gradients of the risk factors with regard to the model's prediction output. Their analysis revealed that for the prediction of knee OA onset, the medial JSN exhibited the highest DeepLIFT gradient, followed by history of injury. However, in the prediction of knee OA deterioration, diabetes and smoking habits showed the second and third highest gradients, respectively, alongside medial JSN, indicating their greater impact compared to injury.

2) GAME THEORY BASED

SHapley Additive Explanations (SHAP) is a widely favoured post-hoc approach for handling tabular data in machine learning models. This approach is rooted in game theory that provides local explanations for individual predictions in the models. By calculating Shapley Values, it assigns importance values to each feature based on their interactions and contributions to the prediction outcome. It enables a comprehensive understanding of the factors driving each prediction and facilitates interpretability by identifying the most influential features in the decision-making process. The findings of all eight studies that utilized SHAP on tabular data were summarized in Table 5.

3) CASE BASED

Case-Based approach is a knowledge-driven approach in which all relevant knowledge is pre-programmed and explicitly specified. Reference [82] demonstrated a case-based methodology for detecting knee osteoarthritis. Their proposed methodology integrated a logic programming approach to knowledge representation and reasoning with a case-based approach to computing, resulting in a comprehensive framework for effective problem-solving in OA field.

4) KNOWLEDGE EXTRACTION BASED

Knowledge distillation is the core of the knowledge extraction based approaches. Reference [116] demonstrated the use of dual-consistency mean teacher model (Figure 10) to discriminate cartilage damages. Both the teacher sub-model and the student sub-model shared common network architecture, but the teacher model utilized an exponential moving average (EMA) strategy for weight updates. This approach involved averaging the student network's weights across multiple training steps, enabling the teacher model to maintain consistent predictions and effectively guide the student network, particularly for unlabelled data. Recent study by [123] employed knowledge distillation to convey pixel and pair-wise information from a teacher network to a student network. The teacher network, built upon HRNet-W, featured a head convolution layer consisting of 64 filters and a 3×3 kernel, whereas the student network was equipped with 32 filters and 3×3 kernels. The student network was trained using pixel-wise knowledge extracted from heatmaps generated by the more complex teacher network with loss function as shown in Equation 3, enabling the student network to adopt a simpler and more compact architecture.

$$L_{pi} = \frac{\sum_{i\in\Re} KL(h_i^s||h_i^t)}{\hat{w} \times \hat{h}}, \quad \Re = 1, 2, \ldots, \hat{\mathrm{w}} \times \hat{\mathrm{h}} \tag{3}$$

where h_i^s represents the response of the pixel at *ith* position in student network, h_i^t represents the response by teacher network at *ith* position of pixel, *KL* represents the Kullback-Leibler exhibiting divergence among two heatmaps, and $\hat{w} \times \hat{h}$ represents feature map.

Reference [124] presented a novel two-stage method inspired by multiple instance learning. This method aimed to identify regions of high likelihood for pathologies by leveraging mixed-format data, which encompassed categorical and positional labels. Their approach incorporated a UNet network along with a morphological peak-finding algorithm to accurately localize defects. Prior to pathology detection, the images were automatically cropped around the anterior cruciate ligament or medial compartment cartilage. Additionally, they employed a deep reinforcement learning model to detect two anatomical landmarks, namely the intercondylar eminence and the fibular styloid, which were used to position a volume of interest in relation to the location of these landmarks.

5) NEURAL BASED

Neural-based techniques encompass methods that explain specific predictions, simplify neural networks, and visualize the features and concepts learned by the network. Reference [84] conducted feature important analysis on a pre-developed model based on random forest algorithm. Their findings demonstrated that cartilage and bone features, including the volume of femoral cartilage and patellar density, played a significant role in classifying the status of the knee, whether it was healthy, degenerative, or traumatic. Reference [9] implemented another neural based technique, namely layer-wise relevance propagation (LRP) as illustrated in Figure 11 to tackle important pixels by running a forward pass through the neural network. In addition, deep Taylor decomposition (DTD) was utilized to backpropagate the relevance $R_t^{(L)}$, allowing for the generation of a visualizable relevance map R_{LRP}.

VI. DISCUSSION

Knee OA is a chronic joint disease that causes disability. The heterogeneous nature of the disease makes the diagnosis challenging, and this has motivated researchers to study XAI for explainable OA diagnosis. Similar to general AI models, XAI models also require a robust evaluation protocol to ensure their effectiveness in addressing Co-12 properties proposed by Nauta et al. [45]. These properties include Correctness, Completeness, Consistency, Continuity, Contrastivity, Covariate complexity, Compactness, Composition, Confidence, Context, Coherence, and Controllability.

In this survey, our observations indicate a relative lack of emphasis on evaluating explanations generated by XAI in existing research. Only three evaluation methods for XAI were identified: sensitivity analysis [76], rate of agreement with medical experts [119], and cross-validation [111]. These methods minimally address the Correctness and Confidence properties.

Pierson et al. [76] employed sensitivity analysis to determine the impact of altering input data, such as masking small image regions, on the AI model's predictions. This method helps assess the model's robustness and the relevance of different input regions. Chang et al. [119] engaged medical experts directly in the validation process to enhance the reliability of XAI techniques. In their approach, a musculoskeletal radiologist with extensive experience in interpreting knee MRI scans reviewed Class Activation Maps (CAMs) for the final 15% of cases. The radiologist compared the

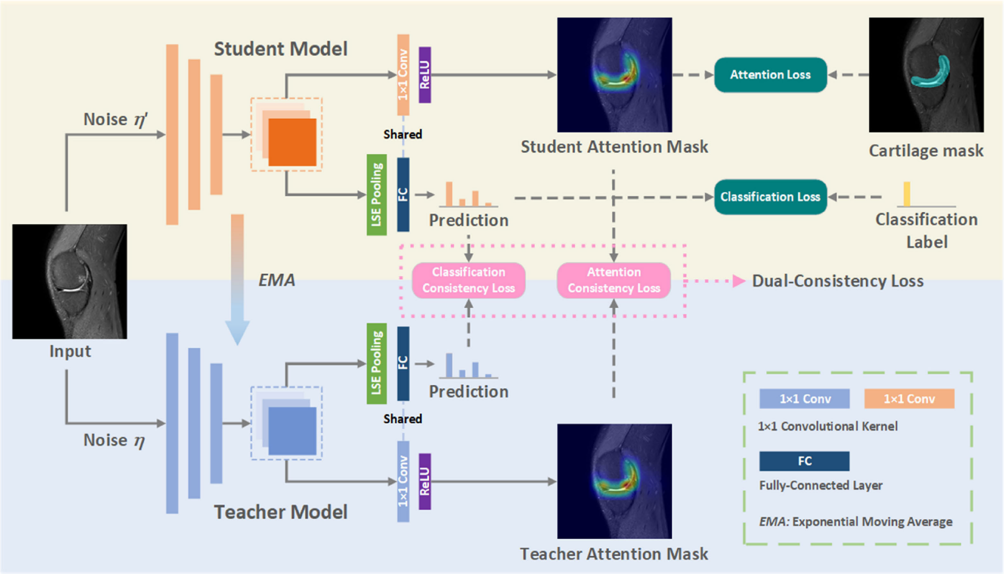


FIGURE 10. Example of knowledge distillation. Adapted from [116].

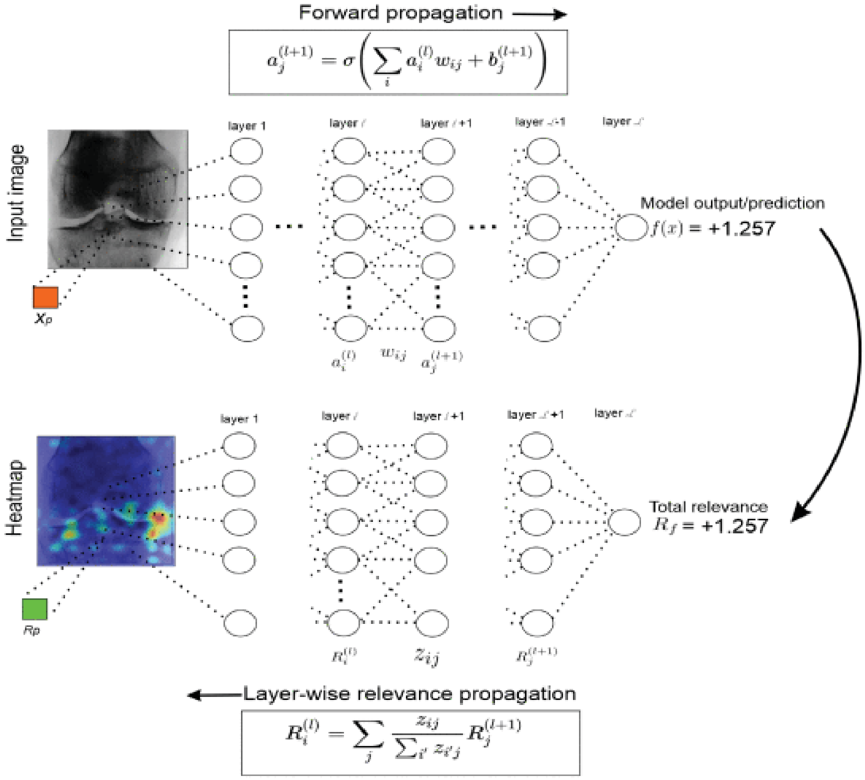


FIGURE 11. Example of LRP for knee OA detection. Adapted from [9].

CAM-highlighted areas with known abnormalities in the MRI images to assess whether the CAMs accurately pinpointed these abnormalities. This method not only provides a rigorous check on the accuracy of the CAMs but also emphasizes the crucial role of expert feedback in evaluating XAI systems. Incorporating insights from medical professionals ensures that the XAI techniques are aligned with clinical expertise, thereby improving the validity and relevance of the generated explanations and making them more applicable in real-world settings. Liu et al. [111] applied cross-validation to an interpretable model to identify key imaging and clinical features for early OA detection. By analyzing features that

TABLE 5. Summary of game-theory-based XAI techniques from included papers. JSN: joint space narrowing; SHAP: SHapley Additive exPlanations; XGBoost: eXtreme Gradient Boosting.

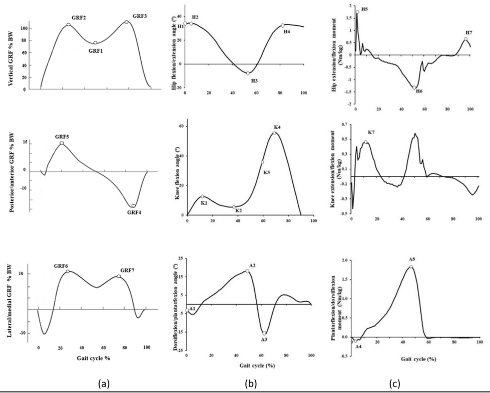
Paper	XAI method	Type of data	Evaluated model (performance)	Target	XAI findings
[70]	SHAP	Tabular clinical data	Logistic regression (77.88% accuracy with 40 risk factors)	Presence of OA (2 classes)	Top ten risk factors 1. Right knee symptoms: swelling, last 7 days (V00KSXRKN1) 2. Left knee symptoms: bend knee fully, last 7 days (V00KSXLKN5) 3. Either knee, history of knee surgery (P02KSURG) 4. Knee symptoms, risk factors, or both, status at initial eligible interview or screening visit (P02ELGRISK) 5. Baseline symptomatic knee OA status by person (P01SXKOA) 6. Right knee exam: patellofemoral crepitus present on exam (V00RKPFCRE) 7. Left knee exam: patellofemoral crepitus present on exam (V00LKPFCRE) 8. Left knee symptoms: swelling, last 7 days (V00KSXLKN1) 9. Average current scale weight in kg (P01WEIGHT) 10. Right knee baseline symptomatic OA status (P01RSXKOA)
[79]	SHAP	Tabular clinical data	Logistic regression (83.3% accuracy with 29 risk factors)	Prediction of JSN progression (2 classes)	The three most significant features were identified: lateral JSN on the right knee (P01SVRKJSL), lateral JSN on the left knee (P01SVLKJSL), and measure related to the percentage of foods marked as a small portion (V00PCTSMAL).
[78]	Kernel SHAP	Tabular clinical data	GenWrapper employing SVM (71.25% mean accuracy with 35 risk factors)	Prediction of OA progression (2 classes)	The most important variables that significantly influenced the prediction output were identified as lateral JSN on the left knee (P01SVLKOST), body mass index, average daily nutrients from vitamin supplements (V00SUPCA), and education level (V00EDCV).
[79]	SHAP	Tabular clinical data	XGBoost (78.14% average accuracy with 31 risk factors)	Prediction of JSN progression (2 classes)	Top five risk factors 1. Lateral JSN on the right knee (P01SVRKJSL) 2. Percentage of foods marked as large portion (V00PCTLARG) 3. Frequency of cream/half and half/non-dairy creamer in coffee or tea in the past 12 months (V00FFQ68) 4. Frequency of getting in and out of squatting position 10 or more times during a typical week in the past 30 days (V00PA530CV) 5. Lateral JSN on the left knee (P01SVLKOST)
[80]	SHAP	Tabular clinical data	Fuzzy feature selection and random forest (73.55% accuracy, 73.82% precision, 73.64% recall, 73.59 F1 with 21 risk factors)	Presence of OA (2 classes)	Top five risk factors 1. Knee symptoms (P02ELGRISK) 2. History of knee surgery (P02KSURG) 3. Age (V00AGE) 4. BMI (P01BMI) 5. KOOS quality of life score (V00KOOSQOL)
[98]	SHAP TreeExplainer	Biochemical markers	K-means clustering, random forest or KNN (F1 scores of 0.85 for C1 vs rest, 0.91 for C2 vs rest, and 0.88 for C3 vs rest)	OA dominant molecular endotypes (3 clusters)	**Cluster 1 - low tissue turnover:** This cluster demonstrated low repair and articular cartilage or subchondral bone turnover, and had the highest proportion of non-progressors. **Cluster 2 - structural damage:** This cluster demonstrated high bone formation/resorption, cartilage degradation and was mostly linked to longitudinal structural progression. **Cluster 3 - systemic inflammation:** This cluster demonstrated joint tissue degradation, inflammation, and cartilage degradation, and was linked to sustained or progressive pain.
[96]	SHAP	Tabular gait data	SVM (94.95% accuracy, 92.16–96.72% precision, 92.19–97.62% recall, and 93.07–96.47% F1 score)	Classification of anterior cruciate ligament injury status (3 classes)	The gait parameters K2, H4, A3, GRF4, GRF7, K1, A4, and GRF6 as illustrated in below figure were identified as the key factors that significantly influenced the model output, with mean SHAP values higher than 0.3. 

TABLE 5. *(Continued.)* **Summary of game-theory-based XAI techniques from included papers. JSN: joint space narrowing; SHAP: SHapley Additive exPlanations; XGBoost: eXtreme Gradient Boosting.**

[73]	SHAP	Tabular clinical data	XGBoost (0.741 AUC with 15 features)	Prediction of risk of developing venous thrombosis (2 classes)	KL grade, age, and hypertension emerged as the three pivotal variables in relation to the risk of venous thrombosis
[81]	Tree SHAP	Tabular clinical data and biomarkers	XGBoost (0.72 ROC-AUC)	Prediction of OA risk in five years (2 classes)	Top three predictors of increased OA risk 1. Older age 2. Higher BMI 3. Use of non-steroidal anti-inflammatory drugs (NSAIDs) in the year before diagnosis

consistently appeared in at least eight out of ten folds of the cross-validation, they determined which features were most significant. This approach ensures that the identified features are both robust and reliable, demonstrating stability and consistency across various data subsets.

The strengths and limitations of XAI methods are summarized in Table 6. Data interpretability plays an important role in understanding and characterizing input data. It is typically employed as a pre-processing model to extract representative features. Model interpretability is crucial for comprehending the decision-making process of AI models. However, due to technical limitations in constructing an effective intrinsic XAI model, most studies prefer a post-hoc model to visualize the reasoning behind predictions made by the trained model. Since most of the OA diagnosis models involve medical imaging, visual XAI dominates over attribution-based XAI. Visual XAI, such as GradCAM, has successfully identified certain OA features, such as bone spurs and narrowed joint spaces. However, the Correctness of visual XAI has not been fully validated.

The application of XAI gives rise to ethical concerns, including data privacy, patient safety, and risk of bias. As XAI models strive to elucidate human health status through the analysis of diverse data sources, there is a potential risk of patient information leakage to third parties, posing harm to the patient. Excessive reliance on XAI approaches may also predispose clinicians to biases. Thus, it is crucial to recognize that XAI is not all-powerful. Identifying the knowledge boundaries of XAI models is upmost important for users to have a clear understanding and make appropriate use of the tool. Users should be informed about the operational boundaries of the models and be able to discern when the models go beyond their knowledge limits, as this can potentially result in errors. While XAI-based systems have the potential to alleviate the workload of healthcare providers, they also raise concerns regarding legal responsibility in cases of unethical actions and errors. Therefore, the development of XAI models in healthcare should be approached with caution, balancing the potential for positive societal impact with the need for ethical consideration.

A. SUGGESTIONS FOR FUTURE WORK

Despite the limitations, XAI holds promise in revealing the crucial pathological patterns associated with the onset and progression of knee OA. Moreover, XAI has the potential to optimize the handling and organization of electronic health record data, resulting in streamlined clinical workflows and significantly reducing the physicians' spending time on making diagnosis, prognosis, and searching for pertinent patient information in electronic records. However, the current state of XAI has certain aspects that need to be addressed in future research.

1) DEVELOPMENT OF INTEGRATED EVALUATION METRICS FOR XAI

Our findings highlight a substantial gap in the evaluation of XAI-generated explanations. A significant number of studies (67 out of 70) have relied predominantly on qualitative assessments. While qualitative analysis can offer valuable insights, it often lacks the objective rigor required for evaluating the effectiveness of XAI techniques, particularly in medical contexts. This reliance on qualitative methods can introduce subjectivity and variability, impeding the objective assessment and comparison of XAI methods in knee OA diagnosis.

Qualitative analysis typically lacks quantifiable metrics, which are essential for objectively measuring the performance of XAI techniques. Without such metrics, comparing different techniques or assessing their performance over time becomes challenging. Furthermore, qualitative assessments often focus on specific cases or examples that may not be representative of broader populations. This limitation restricts the ability to determine the generalizability of XAI techniques across different datasets or real-world clinical settings. Additionally, qualitative analysis may inadvertently reinforce existing biases or preconceived notions about the effectiveness of XAI techniques, as researchers might emphasize positive outcomes while overlooking potential limitations or failures.

To address these challenges, the development and implementation of robust quantitative evaluation metrics are essential. Quantitative metrics provide objective measures and standardized benchmarks, which are crucial for several reasons. Firstly, they facilitate precise comparisons between different XAI techniques by offering clear, numerical criteria for evaluation. This objectivity allows for a more rigorous assessment of how well XAI techniques support decision-making processes in clinical settings. Secondly, quantitative metrics help in identifying the strengths and

TABLE 6. Comparison of XAI methods for OA diagnosis.

Category	Simple usage	Available evaluation methods	Strengths	Limitations
Data interpretability				
Feature extraction				
Exploratory data analysis	✓	-	Provides information about data distribution and patterns.	Requires domain expertise for meaningful interpretation.
Image descriptors	✓	-	Captures detailed information from imaging data, such as bone texture.	May struggle with interpretability of complex image features.
Dimensionality reduction	✓	-	Simplifies data while retaining essential features.	May lose some information during the reduction process.
Explainable feature engineering				
Domain-specific	-	-	Tailored to domain-specific knowledge for better relevance.	Limited generalizability to unspecified domains.
Model based	-	-	Incorporates knowledge from the underlying model.	Highly dependent on the model's interpretability.
Disentangled representation learning	-	-	Learns disentangled features, aiding interpretability.	May struggle with complex relationships between features.
Knowledge graphs	✓	-	Represents complex relationships in a structured manner.	Construction of knowledge graphs is resource-intensive.
Model interpretability				
Interpretable models	✓	✓	Easy to understand and explain.	May sacrifice predictive performance for interpretability.
Explainability through architectural adjustments	-	-	Offers control over interpretability features.	May require AI expertise in model architecture adjustments.
Post-hoc interpretability				
Attribution based	-	-	Provides feature importance scores for model decisions.	Susceptible to noise and lacks a unified evaluation metric.
Perturbation approach	-	-	Robust to noisy features and perturbations.	Computationally expensive for large datasets.
Backpropagation approach				
Gradient-based	✓	-	Directly links features to model decisions.	Sensitive to changes in input data.
CAM	✓	✓	Highlights regions contributing to predictions.	Limited to certain CNN architectures and may lack fine-grained details.
GradCAM	✓	✓	Improves upon CAM by providing more localized attention.	Limited to certain CNN architectures.
EigenCAM	-	-	Incorporates eigenvalues for improved attention maps.	Limited to certain CNN architectures and data types.
DeepLIFT	-	-	Addresses vanishing and exploding gradients.	Computationally expensive for deep neural networks.
Game theory based	✓	-	Captures interactions between features.	Limited research and application in comparison to other methods.
Case based	✓	-	Relies on specific instances for interpretability.	May lack generalizability.
Knowledge extraction based	-	-	Extracts explicit knowledge from the model.	May struggle with highly complex models.
Neural based	-	-	Utilizes neural networks for interpretability.	May lack transparency in complicated networks.

weaknesses of various models, enabling researchers and practitioners to pinpoint areas for improvement. This is crucial for advancing the field and ensuring that XAI techniques are both effective and reliable. Finally, the use of quantitative metrics enhances the reproducibility of results by establishing consistent evaluation standards, which is vital for validating the effectiveness of XAI techniques across different studies and datasets.

By integrating quantitative evaluation metrics alongside qualitative analyses, a more comprehensive and objective assessment of XAI techniques can be achieved. This dual approach ensures that evaluations are not only insightful but also rigorous, thereby enhancing the overall reliability and applicability of XAI methods in clinical settings. Such an integrated approach is necessary to fully understand and improve the impact of XAI on knee OA diagnosis.

2) INTEGRATION OF DOMAIN-SPECIFIC INFORMATION FROM STAKEHOLDERS AND CLINICAL VALIDATION

The deployment of XAI could lead to the real application of AI in healthcare, and overcome the lack of operator confidence in AI models. However, it is essential to understand how the application of these models in clinical tasks will be perceived, whether as a support or a substitute for medical expert's work, as well as the level of substitution. To achieve this, AI programmers must discern which explanations are valuable and which are not for medical professionals. Creating an XAI model that is deemed useless or difficult to comprehend may deter medical experts from utilizing it. Furthermore, patients play a significant role as stakeholders since the developed model aims to elucidate their health status. Therefore, their expectations and special needs should be taken into account and integrated into the process. To address the challenges, [15] implemented a qualitative co-design approach at an academic health center in Southern Alberta, which involved conducting focus groups with patients, physicians, researchers, and industry partners, as well as analyzing prioritization activities and a pre-post quality and satisfaction Kano survey. The structured co-design processes were developed based on the basis of shared concepts, language, power dynamics, rationale, mutual learning, and respect for diversity and differing opinions.

Further clinical validation is critical for assessing the real-world impact and feasibility of XAI techniques in knee OA diagnosis. Clinical validation ensures that XAI models are not only reliable and effective in controlled settings but also deliver consistent, interpretable results in actual medical practice. This process confirms that XAI techniques provide practical value and integrate well into everyday clinical scenarios.

Moreover, clinical validation enhances transparency and builds trust in AI-driven diagnostic systems. By demonstrating that XAI models offer accurate and understandable explanations, validation increases confidence among medical professionals and patients alike. It addresses reliability concerns and facilitates the integration of AI tools into standard clinical workflows. Robust validation also helps identify and mitigate potential risks and limitations, allowing for necessary refinements and improving the overall effectiveness of XAI in real-world applications.

In summary, integrating domain-specific information from stakeholders is essential for developing effective XAI models. However, robust clinical validation is equally crucial to ensure these models have a meaningful impact and are practical for widespread use in knee OA diagnosis. This dual approach ensures that XAI techniques meet both technical and clinical standards, ultimately enhancing patient care and outcomes.

3) EXPLORING PATIENT DISPARITIES AND POPULATION-SPECIFIC FACTORS TO ENHANCE GENERALIZABILITY

As highlighted by Pierson et al. [76], there are noticeable racial and socioeconomic disparities in OA data. By considering these disparities during the training of AI models, there is a potential to enhance accuracy. The study also revealed that patient-perceived OA symptoms vary based on factors such as education, culture, and geography. Considering these variations is crucial in developing AI models that accurately capture the diverse experiences and manifestations of OA among different patient populations.

However, a notable challenge is the limited representation of diverse populations in widely-used datasets. For instance, while large datasets like OAI are extensively utilized, with 64.3% of studies relying on them, they predominantly focus on the US population, with a marked emphasis on European American individuals. This overrepresentation can severely limit the generalizability of findings, as these datasets may not fully reflect the experiences of individuals from other racial, socioeconomic, or geographic backgrounds. As a result, AI models trained on these datasets may perform well for the overrepresented demographic but less effectively for underrepresented groups, potentially leading to biased results and less accurate clinical interpretations for those populations.

In addition, we observed that there is lack of well-organized open access data specifically for the Eastern population, despite the higher prevalence of OA issues in this population [125]. This highlights a significant gap in available resources for studying and addressing OA within the Eastern population. The limited availability of comprehensive and representative data from this specific demographic group hinders the development and evaluation of AI models tailored to their unique needs and characteristics.

4) EXPLORING ALTERNATIVE XAI TECHNIQUES FOR KNEE OA APPLICATIONS

In addition to the XAI applications discussed in the Section V, a prospective XAI technique in OA diagnosis could be image captioning. It is a process of generating a textual description of an image using AI algorithms. Medical imaging is an area where this technology could be particularly useful, as generating accurate and detailed descriptions of radiology and pathology images could help healthcare professionals to identify the specific areas of the knee that require treatment and make better-informed decisions about patient care. This area of research presents an exciting opportunity for the development of new XAI models that could have a significant impact on the future of musculoskeletal healthcare. Furthermore, the exploration of the counterfactual approach to XAI in the context of OA applications presents an additional avenue for research. This approach aims to enhance people understanding of AI systems by offering counterfactual explanations specific to target domain. Recent studies have shown that counterfactuals can provide richer information compared to causal explanations, as they encompass a broader range of possibilities in their mental representation [126].

VII. CONCLUSION

A substantial number of studies in the field of computer-aided knee OA diagnosis have sought to enhance the explainability of their deep learning models through the integration of XAI techniques. Despite the existing limitations in current XAI methods, such as the absence of a standardized evaluation method for measuring explanation quality, incompleteness, and the lack of established guidelines, the development of XAI in knee OA detection aligns with the trend of precision diagnosis, offering the potential to reduce the healthcare burden and promote preventive strategies for musculoskeletal diseases.

CONFLICT OF INTEREST

There is no conflict of interest reported.

REFERENCES

[1] J. D. Steinmetz, G. T. Culbreth, L. M. Haile, Q. Rafferty, J. Lo, K. G. Fukutaki, J. A. Cruz, A. E. Smith, S. E. Vollset, P. M. Brooks, and M. Cross, "Global, regional, and national burden of osteoarthritis, 1990–2020 and projections to 2050: A systematic analysis for the global burden of disease study 2021," *Lancet Rheumatol.*, vol. 5, no. 9, pp. e508–e522, 1990.

[2] Y. Ding, X. Liu, C. Chen, C. Yin, and X. Sun, "Global, regional, and national trends in osteoarthritis disability-adjusted life years (DALYs) from 1990 to 2019: A comprehensive analysis of the global burden of disease study," *Public Health*, vol. 226, pp. 261–272, Jan. 2024.

[3] E. H. Park and J. Fritz, "The role of imaging in osteoarthritis," *Best Pract. Res. Clin. Rheumatol.*, vol. 27, no. 2, 2023, Art. no. 101866.

[4] J.-B. Bouillon-Minois, C. Lambert, F. Dutheil, J. Raconnat, M. Benamor, B. Dalle, M. Laurent, O. J. Adeyemi, A. Lhoste-Trouilloud, and J. Schmidt, "Assessment of discordance between radiologists and emergency physicians of RADIOgraphs among discharged patients in an emergency department: The RADIO-ED study," *Emergency Radiol.*, vol. 31, no. 2, pp. 125–131, Jan. 2024.

[5] A. Tiulpin, S. Klein, S. M. A. Bierma-Zeinstra, J. Thevenot, E. Rahtu, J. V. Meurs, E. H. G. Oei, and S. Saarakkala, "Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data," *Sci. Rep.*, vol. 9, no. 1, p. 20038, Dec. 2019.

[6] C. Guida, M. Zhang, and J. Shan, "Improving knee osteoarthritis classification using multimodal intermediate fusion of X-ray, MRI, and clinical information," *Neural Comput. Appl.*, vol. 35, no. 13, pp. 9763–9772, May 2023.

[7] C. L. Piccolo, C. A. Mallio, F. Vaccarino, R. F. Grasso, and B. B. Zobel, "Imaging of knee osteoarthritis: A review of multimodal diagnostic approach," *Quant. Imag. Med. Surg.*, vol. 13, no. 11, pp. 7582–7595, Nov. 2023.

[8] A. Tiulpin, J. Thevenot, E. Rahtu, P. Lehenkari, and S. Saarakkala, "Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach," *Sci. Rep.*, vol. 8, no. 1, pp. 1–10, Jan. 2018.

[9] Md. R. Karim, J. Jiao, T. Döhmen, M. Cochez, O. Beyan, D. Rebholz-Schuhmann, and S. Decker, "DeepKneeExplainer: Explainable knee osteoarthritis diagnosis from radiographs and magnetic resonance imaging," *IEEE Access*, vol. 9, pp. 39757–39780, 2021.

[10] P. Liu, J. Zhang, S. Liu, T. Huo, J. He, M. Xue, Y. Fang, H. Wang, Y. Xie, M. Xie, D. Zhang, and Z. Ye, "Application of artificial intelligence technology in the field of orthopedics: A narrative review," *Artif. Intell. Rev.*, vol. 57, no. 1, p. 13, Jan. 2024.

[11] H. Hah and D. S. Goldin, "How clinicians perceive artificial intelligence–assisted technologies in diagnostic decision making: Mixed methods approach," *J. Med. Internet Res.*, vol. 23, no. 12, Dec. 2021, Art. no. e33540.

[12] J. Lee and S. W. Chung, "Deep learning for orthopedic disease based on medical image analysis: Present and future," *Appl. Sci.*, vol. 12, no. 2, p. 681, Jan. 2022, doi: 10.3390/app12020681.

[13] P. Kumar, S. Chauhan, and L. K. Awasthi, "Artificial intelligence in healthcare: Review, ethics, trust challenges & future research directions," *Eng. Appl. Artif. Intell.*, vol. 120, Apr. 2023, Art. no. 105894.

[14] L. Rubinger, A. Gazendam, S. Ekhtiari, and M. Bhandari, "Machine learning and artificial intelligence in research and healthcare," *Injury*, vol. 54, pp. S69–S73, May 2023.

[15] K. J. Mrklas, T. Barber, D. Campbell-Scherer, L. A. Green, L. C. Li, N. Marlett, J. Miller, B. Shewchuk, S. Teare, T. Wasylak, and D. A. Marshall, "Co-design in the development of a mobile health app for the management of knee osteoarthritis by patients and physicians: Qualitative study," *JMIR mHealth uHealth*, vol. 8, no. 7, Jul. 2020, Art. no. e17893.

[16] F. Giuste, W. Shi, Y. Zhu, T. Naren, M. Isgut, Y. Sha, L. Tong, M. Gupte, and M. D. Wang, "Explainable artificial intelligence methods in combating pandemics: A systematic review," *IEEE Rev. Biomed. Eng.*, vol. 16, pp. 5–21, 2023.

[17] 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, Jul. 2022, Art. no. 102470.

[18] A. M. Groen, R. Kraan, S. F. Amirkhan, J. G. Daams, and M. Maas, "A systematic review on the use of explainability in deep learning systems for computer aided diagnosis in radiology: Limited use of explainable AI?" *Eur. J. Radiol.*, vol. 157, Dec. 2022, Art. no. 110592.

[19] 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.

[20] S. Nazir, D. M. Dickson, and M. U. Akram, "Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks," *Comput. Biol. Med.*, vol. 156, Apr. 2023, Art. no. 106668.

[21] D. W. Joyce, A. Kormilitzin, K. A. Smith, and A. Cipriani, "Explainable artificial intelligence for mental health through transparency and interpretability for understandability," *NPJ Digit. Med.*, vol. 6, no. 1, p. 6, Jan. 2023.

[22] S. Bharati, M. R. H. Mondal, and P. Podder, "A review on explainable artificial intelligence for healthcare: Why, how, and when?" *IEEE Trans. Artif. Intell.*, vol. 5, no. 4, pp. 1429–1442, Apr. 2023.

[23] J. Antony, K. McGuinness, N. E. O'Connor, and K. Moran, "Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 1195–1200.

[24] M. W. Brejnebøl, P. Hansen, J. U. Nybing, R. Bachmann, U. Ratjen, I. V. Hansen, A. Lenskjold, M. Axelsen, M. Lundemann, and M. Boesen, ''External validation of an artificial intelligence tool for radiographic knee osteoarthritis severity classification," *Eur. J. Radiol.*, vol. 150, May 2022, Art. no. 110249.

[25] Z. Wang, A. Chetouani, and R. Jennane, "A confident labelling strategy based on deep learning for improving early detection of knee OsteoArthritis," 2023, *arXiv:2303.13203*.

[26] S. Moustakidis, N. I. Papandrianos, E. Christodolou, E. Papageorgiou, and D. Tsaopoulos, "Dense neural networks in knee osteoarthritis classification: A study on accuracy and fairness," *Neural Comput. Appl.*, vol. 35, no. 1, pp. 21–33, Jan. 2023.

[27] P. Chen, L. Gao, X. Shi, K. Allen, and L. Yang, "Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss," *Computerized Med. Imag. Graph.*, vol. 75, pp. 84–92, Jul. 2019.

[28] B. Zhang, J. Tan, K. Cho, G. Chang, and C. M. Deniz, "Attention-based CNN for KL grade classification: Data from the osteoarthritis initiative," in *Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2020, pp. 731–735.

[29] K. A. Thomas, Ł. Kidziński, E. Halilaj, S. L. Fleming, G. R. Venkataraman, E. H. G. Oei, G. E. Gold, and S. L. Delp, "Automated classification of radiographic knee osteoarthritis severity using deep neural networks," *Radiol., Artif. Intell.*, vol. 2, no. 2, Mar. 2020, Art. no. e190065.

[30] V. V. Kishore, V. Kalpana, and G. H. Kumar, "Evaluating the efficacy of deep learning models for knee osteoarthritis prediction based on Kellgren–Lawrence grading system," *e-Prime, Adv. Electr. Eng., Electron. Energy*, vol. 5, Sep. 2023, Art. no. 100266.

[31] Y. X. Teoh, A. Othmani, K. W. Lai, S. L. Goh, and J. Usman, "Stratifying knee osteoarthritis features through multitask deep hybrid learning: Data from the osteoarthritis initiative," *Comput. Methods Programs Biomed.*, vol. 242, Dec. 2023, Art. no. 107807.

[32] M. Binvignat, V. Pedoia, A. J. Butte, K. Louati, D. Klatzmann, F. Berenbaum, E. Mariotti-Ferrandiz, and J. Sellam, "Use of machine learning in osteoarthritis research: A systematic literature review," *Osteoarthritis Cartilage*, vol. 30, p. 81, Apr. 2022.

[33] D. Saini, T. Chand, D. K. Chouhan, and M. Prakash, "A comparative analysis of automatic classification and grading methods for knee osteoarthritis focussing on X-ray images," *Biocybern. Biomed. Eng.*, vol. 41, no. 2, pp. 419–444, Apr. 2021.

[34] R. Kijowski, J. Fritz, and C. M. Deniz, "Deep learning applications in osteoarthritis imaging," *Skeletal Radiol.*, vol. 52, no. 11, pp. 2225–2238, Nov. 2023.

[35] L. Arbeeva, M. C. Minnig, K. A. Yates, and A. E. Nelson, "Machine learning approaches to the prediction of osteoarthritis phenotypes and outcomes," *Current Rheumatol. Rep.*, vol. 25, no. 11, pp. 213–225, Nov. 2023.

[36] C. Wang, B. Huang, N. Thogiti, W. Zhu, C. Chang, J. Pao, and F. Lai, "Successful real-world application of an osteoarthritis classification deep-learning model using 9210 knees—An orthopedic surgeon's view," *J. Orthopaedic Res.*, vol. 41, no. 4, pp. 737–746, Apr. 2023.

[37] S. Buijsman, "Defining explanation and explanatory depth in XAI," *Minds Mach.*, vol. 32, no. 3, pp. 563–584, Sep. 2022.

[38] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, Dec. 2020.

[39] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera, "Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence," *Inf. Fusion*, vol. 99, Nov. 2023, Art. no. 101805.

[40] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 44, pp. 22071–22080, Oct. 2019.

[41] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, ''Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,'' *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.

[42] G. Schwalbe and B. Finzel, ''A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts,'' *Data Mining Knowl. Discovery*, pp. 1–59, Jan. 2023.

[43] A. Chaddad, J. Peng, J. Xu, and A. Bouridane, ''Survey of explainable AI techniques in healthcare,'' *Sensors*, vol. 23, no. 2, p. 634, Jan. 2023.

[44] M. Du, N. Liu, and X. Hu, ''Techniques for interpretable machine learning,'' *Commun. ACM*, vol. 63, no. 1, pp. 68–77, Dec. 2019.

[45] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, and C. Seifert, ''From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI,'' *ACM Comput. Surv.*, vol. 55, no. 13s, pp. 1–42, Dec. 2023, doi: 10.1145/3583558.

[46] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim, ''Designing theory-driven user-centric explainable AI,'' in *Proc. CHI Conf. Human Factors Comput. Syst.*, May 2019, pp. 1–15.

[47] T. A. J. Schoonderwoerd, W. Jorritsma, M. A. Neerincx, and K. van den Bosch, ''Human-centered XAI: Developing design patterns for explanations of clinical decision support systems,'' *Int. J. Hum.-Comput. Stud.*, vol. 154, Oct. 2021, Art. no. 102684.

[48] A. Hleg, ''Ethics guidelines for trustworthy AI,'' B-1049, Brussels, Belgium, 2019, doi: 10.2759/346720.

[49] A. Kumar, T. Braud, S. Tarkoma, and P. Hui, ''Trustworthy AI in the age of pervasive computing and big data,'' in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops (PerCom Workshops)*, Mar. 2020, pp. 1–6.

[50] C. F. Mondschein and C. Monda, ''The EU's general data protection regulation (GDPR) in a research context,'' in *Fundamentals of Clinical Data Science*. Cham, Switzerland: Springer, 2019, pp. 55–71.

[51] R. Hamon, H. Junklewitz, I. Sanchez, G. Malgieri, and P. De Hert, ''Bridging the gap between AI and explainability in the GDPR: Towards trustworthiness-by-design in automated decision-making,'' *IEEE Comput. Intell. Mag.*, vol. 17, no. 1, pp. 72–85, Feb. 2022.

[52] M. Kop, ''EU artificial intelligence act: The European approach to AI,'' Stanford Law School, Stanford-Vienna Transatlantic Technol. Law Forum, Transatlantic Antitrust IPR Develop., Stanford Univ., 2021, no. 2/2021. [Online]. Available: https://law.stanford.edu/publications/eu-artificial-intelligence-act-the-european-approach-to-ai/

[53] H. Van Kolfschooten, ''EU regulation of artificial intelligence: Challenges for patients' rights,'' *Common Market Law Rev.*, vol. 59, no. 1, pp. 81–112, 2022.

[54] D. Gunning and D. W. Aha, ''DARPA's explainable artificial intelligence program,'' *AI Mag*, vol. 40, no. 2, p. 44, 2019.

[55] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, ''The false hope of current approaches to explainable artificial intelligence in health care,'' *Lancet Digit. Health*, vol. 3, no. 11, pp. e745–e750, Nov. 2021.

[56] B. Kitchenham, ''Procedures for performing systematic reviews,'' *Keele, U.K., Keele Univ.*, vol. 33, no. 2004, pp. 1–26, 2004.

[57] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, ''Grad-CAM: Visual explanations from deep networks via gradient-based localization,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[58] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, ''Attention is all you need,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[59] N. J. van Eck and L. Waltman, ''Software survey: VOSviewer, a computer program for bibliometric mapping,'' *Scientometrics*, vol. 84, no. 2, pp. 523–538, Aug. 2010.

[60] United States Department of Health and Human Services. (2023). *Osteoarthritis Initiative*. [Online]. Available: https://nda.nih.gov/oai/

[61] N. A. Segal, M. C. Nevitt, K. D. Gross, J. Hietpas, N. A. Glass, C. E. Lewis, and J. C. Torner, ''The multicenter osteoarthritis study: Opportunities for rehabilitation research,'' *PM R*, vol. 5, no. 8, pp. 647–654, Aug. 2013.

[62] N. Bien et al., ''Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet,'' *PLOS Med.*, vol. 15, no. 11, Nov. 2018, Art. no. e1002699.

[63] R. Zhao, B. Yaman, Y. Zhang, R. Stewart, A. Dixon, F. Knoll, Z. Huang, Y. W. Lui, M. S. Hansen, and M. P. Lungren, ''FastMRI+, clinical pathology annotations for knee and brain fully sampled magnetic resonance imaging data,'' *Scientific Data*, vol. 9, no. 1, p. 152, Apr. 2022.

[64] F. Knoll et al., ''FastMRI: A publicly available raw k-space and DICOM dataset of knee images for accelerated MR image reconstruction using machine learning,'' *Radiol., Artif. Intell.*, vol. 2, no. 1, Jan. 2020, Art. no. e190007.

[65] J. Wesseling, M. Boers, M. A. Viergever, W. K. Hilberdink, F. P. Lafeber, J. Dekker, and J. W. Bijlsma, ''Cohort profile: Cohort hip and cohort knee (CHECK) study,'' *Int. J. Epidemiol.*, vol. 45, no. 1, pp. 36–44, Feb. 2016.

[66] Q. Wang et al., ''A machine learning approach reveals features related to clinicians' diagnosis of clinically relevant knee osteoarthritis,'' *Rheumatology*, vol. 62, no. 8, pp. 2732–2739, 2022.

[67] S. Olsson, E. Akbarian, A. Lind, A. S. Razavian, and M. Gordon, ''Automating classification of osteoarthritis according to kellgren-lawrence in the knee using deep learning in an unfiltered adult population,'' *BMC Musculoskeletal Disorders*, vol. 22, no. 1, pp. 1–8, Dec. 2021.

[68] L. Liu, Y. Yu, Z. Fei, M. Li, F.-X. Wu, H.-D. Li, Y. Pan, and J. Wang, ''An interpretable boosting model to predict side effects of analgesics for osteoarthritis,'' *BMC Syst. Biol.*, vol. 12, no. S6, pp. 29–38, Nov. 2018.

[69] D. H. Kim, K. J. Lee, D. Choi, J. I. Lee, H. G. Choi, and Y. S. Lee, ''Can additional patient information improve the diagnostic performance of deep learning for the interpretation of knee osteoarthritis severity,'' *J. Clin. Med.*, vol. 9, no. 10, p. 3341, Oct. 2020.

[70] C. Kokkotis, S. Moustakidis, E. Papageorgiou, G. Giakas, and D. Tsaopoulos, ''A machine learning workflow for diagnosis of knee osteoarthritis with a focus on post-hoc explainability,'' in *Proc. 11th Int. Conf. Inf., Intell., Syst. Appl. (IISA*, Jul. 2020, pp. 1–7.

[71] L. C. Chan, H. H. T. Li, P. K. Chan, and C. Wen, ''A machine learning-based approach to decipher multi-etiology of knee osteoarthritis onset and deterioration,'' *Osteoarthritis Cartilage Open*, vol. 3, no. 1, Mar. 2021, Art. no. 100135.

[72] A. E. Nelson, T. H. Keefe, T. A. Schwartz, L. F. Callahan, R. F. Loeser, Y. M. Golightly, L. Arbeeva, and J. S. Marron, ''Biclustering reveals potential knee OA phenotypes in exploratory analyses: Data from the osteoarthritis initiative,'' *PLoS ONE*, vol. 17, no. 5, May 2022, Art. no. e0266964.

[73] C. Lu, J. Song, H. Li, W. Yu, Y. Hao, K. Xu, and P. Xu, ''Predicting venous thrombosis in osteoarthritis using a machine learning algorithm: A population-based cohort study,'' *J. Personalized Med.*, vol. 12, no. 1, p. 114, Jan. 2022.

[74] X. Li, H. Liu, X. Zhao, G. Zhang, and C. Xing, ''Automatic approach for constructing a knowledge graph of knee osteoarthritis in Chinese,'' *Health Inf. Sci. Syst.*, vol. 8, no. 1, pp. 1–8, Dec. 2020.

[75] C. Ntakolia, C. Kokkotis, S. Moustakidis, and D. Tsaopoulos, ''Prediction of joint space narrowing progression in knee osteoarthritis patients,'' *Diagnostics*, vol. 11, no. 2, p. 285, Feb. 2021.

[76] E. Pierson, D. M. Cutler, J. Leskovec, S. Mullainathan, and Z. Obermeyer, ''An algorithmic approach to reducing unexplained pain disparities in underserved populations,'' *Nature Med.*, vol. 27, no. 1, pp. 136–140, Jan. 2021.

[77] J.-B. Schiratti, R. Dubois, P. Herent, D. Cahané, J. Dachary, T. Clozel, G. Wainrib, F. Keime-Guibert, A. Lalande, M. Pueyo, R. Guillier, C. Gabarroca, and P. Moingeon, ''A deep learning method for predicting knee osteoarthritis radiographic progression from MRI,'' *Arthritis Res. Therapy*, vol. 23, no. 1, pp. 1–10, Dec. 2021.

[78] C. Kokkotis, S. Moustakidis, V. Baltzopoulos, G. Giakas, and D. Tsaopoulos, ''Identifying robust risk factors for knee osteoarthritis progression: An evolutionary machine learning approach,'' *Healthcare*, vol. 9, no. 3, p. 260, Mar. 2021.

[79] C. Ntakolia, C. Kokkotis, S. Moustakidis, and D. Tsaopoulos, ''Identification of most important features based on a fuzzy ensemble technique: Evaluation on joint space narrowing progression in knee osteoarthritis patients,'' *Int. J. Med. Informat.*, vol. 156, Dec. 2021, Art. no. 104614.

[80] C. Kokkotis, C. Ntakolia, S. Moustakidis, G. Giakas, and D. Tsaopoulos, ''Explainable machine learning for knee osteoarthritis diagnosis based on a novel fuzzy feature selection methodology,'' *Phys. Eng. Sci. Med.*, vol. 45, no. 1, pp. 219–229, Mar. 2022.

[81] R. L. Nielsen, T. Monfeuga, R. R. Kitchen, L. Egerod, L. G. Leal, A. T. H. Schreyer, F. S. Gade, C. Sun, M. Helenius, L. Simonsen, M. Willert, A. A. Tahrani, Z. McVey, and R. Gupta, ''Data-driven identification of predictive risk biomarkers for subgroups of osteoarthritis using interpretable machine learning,'' *Nature Commun.*, vol. 15, no. 1, p. 2817, Apr. 2024.

[82] M. Esteves, H. Vicente, J. Machado, V. Alves, and J. Neves, "A case based methodology for problem solving aiming at knee osteoarthritis detection," in *Proc. 2nd Int. Conf. Soft Comput. Data Mining (SCDM)*, Bandung, Indonesia. Cham, Switzerland: Springer, Aug. 2016, pp. 274–284.

[83] A. Tack, A. Shestakov, D. Lüdke, and S. Zachow, "A multi-task deep learning method for detection of meniscal tears in MRI data from the osteoarthritis initiative database," *Frontiers Bioeng. Biotechnol.*, vol. 9, Dec. 2021, Art. no. 747217.

[84] F. K. Ciliberti, L. Guerrini, A. E. Gunnarsson, M. Recenti, D. Jacob, V. Cangiano, Y. A. Tesfahunegn, A. S. Islind, F. Tortorella, M. Tsirilaki, H. Jónsson, P. Gargiulo, and R. Aubonnet, "CT- and MRI-based 3D reconstruction of knee joint to assess cartilage and bone," *Diagnostics*, vol. 12, no. 2, p. 279, Jan. 2022.

[85] R. K. Jain, P. K. Sharma, S. Gaj, A. Sur, and P. Ghosh, "Knee osteoarthritis severity prediction using an attentive multi-scale deep convolutional neural network," *Multimedia Tools Appl.*, vol. 83, no. 3, pp. 6925–6942, Jan. 2024.

[86] C. J. Hanrahan, "Editorial for 'Associating knee osteoarthritis progression with temporal-regional graph convolutional network analysis on MR Images," *J. Magn. Reson. Imag.*, May 2024, doi: 10.1002/jmri.29440.

[87] J. Hu, C. Zheng, Q. Yu, L. Zhong, K. Yu, Y. Chen, Z. Wang, B. Zhang, Q. Dou, and X. Zhang, "DeepKOA: A deep-learning model for predicting progression in knee osteoarthritis using multimodal magnetic resonance images from the osteoarthritis initiative," *Quant. Imag. Med. Surg.*, vol. 13, no. 8, pp. 4852–4866, Aug. 2023.

[88] K. Fatema, M. A. H. Rony, S. Azam, M. S. H. Mukta, A. Karim, M. Z. Hasan, and M. Jonkman, "Development of an automated optimal distance feature-based decision system for diagnosing knee osteoarthritis using segmented X-ray images," *Heliyon*, vol. 9, no. 11, Nov. 2023, Art. no. e21703.

[89] S.-W. Pi, B.-D. Lee, M. S. Lee, and H. J. Lee, "Ensemble deep-learning networks for automated osteoarthritis grading in knee X-ray images," *Sci. Rep.*, vol. 13, no. 1, p. 22887, Dec. 2023.

[90] F. Prezja, L. Annala, S. Kiiskinen, and T. Ojala, "Exploring the efficacy of base data augmentation methods in deep learning-based radiograph classification of knee joint osteoarthritis," *Algorithms*, vol. 17, no. 1, p. 8, Dec. 2023.

[91] C. Nichols, H. T. Crane, D. Ewart, and O. T. Inan, "Combining knee acoustic emissions, patient-reported measures, and machine learning to assess osteoarthritis severity," in *Proc. IEEE 19th Int. Conf. Body Sensor Netw. (BSN)*, Oct. 2023, pp. 1–4.

[92] M. Kotti, L. D. Duffell, A. A. Faisal, and A. H. McGregor, "Detecting knee osteoarthritis and its discriminating parameters using random forests," *Med. Eng. Phys.*, vol. 43, pp. 19–29, May 2017.

[93] G. Leporace, F. Gonzalez, L. Metsavaht, M. Motta, F. P. Carpes, J. Chahla, and M. Luzo, "Are there different gait profiles in patients with advanced knee osteoarthritis? A machine learning approach," *Clin. Biomech.*, vol. 88, Aug. 2021, Art. no. 105447.

[94] G. C. Ozmen, A. H. Gazi, S. Gharehbaghi, K. L. Richardson, M. Safaei, D. C. Whittingslow, S. Prahalad, J. L. Hunnicutt, J. W. Xerogeanes, T. K. Snow, and O. T. Inan, "An interpretable experimental data augmentation method to improve knee health classification using joint acoustic emissions," *Ann. Biomed. Eng.*, vol. 49, no. 9, pp. 2399–2411, Sep. 2021.

[95] X. Zeng, Y. Zhu, Z. Xie, G. Zhong, W. Huang, L. Ma, Y. Zhang, and C. Mao, "3D knee kinematic parameters effectively diagnose knee osteoarthritis and assess its therapeutic strategy," *Adv. Intell. Syst.*, vol. 4, no. 6, Jun. 2022, Art. no. 2100161.

[96] C. Kokkotis, S. Moustakidis, T. Tsatalas, C. Ntakolia, G. Chalatsis, S. Konstadakos, M. E. Hantes, G. Giakas, and D. Tsaopoulos, "Leveraging explainable machine learning to identify gait biomechanical parameters associated with anterior cruciate ligament injury," *Sci. Rep.*, vol. 12, no. 1, p. 6647, Apr. 2022.

[97] K. L. Young-Shand, P. C. Roy, M. J. Dunbar, S. S. R. Abidi, and J. L. A. Wilson, "Gait biomechanics phenotypes among total knee arthroplasty candidates by machine learning cluster analysis," *J. Orthopaedic Res.*, vol. 41, no. 2, pp. 335–344, Feb. 2023.

[98] F. Angelini, P. Widera, A. Mobasheri, J. Blair, A. Struglics, M. Uebelhoer, Y. Henrotin, A. C. Marijnissen, M. Kloppenburg, F. J. Blanco, I. K. Haugen, F. Berenbaum, C. Ladel, J. Larkin, A. C. Bay-Jensen, and J. Bacardit, "Osteoarthritis endotype discovery via clustering of biochemical marker data," *Ann. Rheumatic Diseases*, vol. 81, no. 5, pp. 666–675, May 2022.

[99] S. Gornale and P. Patravali, 2020, "Digital knee X-ray images," Mendeley Data, doi: 10.17632/t9ndx37v5h.1.

[100] A. Patron, L. Annala, O. Lainiala, J. Paloneva, and S. Äyrämö, "An automatic method for assessing spiking of tibial tubercles associated with knee osteoarthritis," *Diagnostics*, vol. 12, no. 11, p. 2603, Oct. 2022.

[101] A. G. Morales, J. J. Lee, F. Caliva, C. Iriondo, F. Liu, S. Majumdar, and V. Pedoia, "Uncovering associations between data-driven learned qMRI biomarkers and chronic pain," *Sci. Rep.*, vol. 11, no. 1, p. 21989, Nov. 2021.

[102] K. Sahoo, A. K. Samal, J. Pramanik, and S. K. Pani, "Exploratory data analysis using Python," *Int. J. Innov. Technol. Exploring Eng.*, vol. 8, no. 12, pp. 4727–4735, 2019.

[103] L. Jakaite, V. Schetinin, J. Hladuvka, S. Minaev, A. Ambia, and W. Krzanowski, "Deep learning for early detection of pathological changes in X-ray bone microstructures: Case of osteoarthritis," *Sci. Rep.*, vol. 11, no. 1, pp. 1–9, Jan. 2021.

[104] N. Bayramoglu, A. Tiulpin, J. Hirvasniemi, M. T. Nieminen, and S. Saarakkala, "Adaptive segmentation of knee radiographs for selecting the optimal ROI in texture analysis," *Osteoarthritis Cartilage*, vol. 28, no. 7, pp. 941–952, Jul. 2020.

[105] N. Farajzadeh, N. Sadeghzadeh, and M. Hashemzadeh, "IJES-OA net: A residual neural network to classify knee osteoarthritis from radiographic images based on the edges of the intra-joint spaces," *Med. Eng. Phys.*, vol. 113, Mar. 2023, Art. no. 103957.

[106] W. Li, Z. Xiao, J. Liu, J. Feng, D. Zhu, J. Liao, W. Yu, B. Qian, X. Chen, Y. Fang, and S. Li, "Deep learning-assisted knee osteoarthritis automatic grading on plain radiographs: The value of multiview X-ray images and prior knowledge," *Quant. Imag. Med. Surg.*, vol. 13, no. 6, pp. 3587–3601, Jun. 2023.

[107] Z. Wang, A. Chetouani, M. Jarraya, D. Hans, and R. Jennane, "Transformer with selective shuffled position embedding and key-patch exchange strategy for early detection of knee osteoarthritis," 2023, *arXiv:2304.08364*.

[108] Y. Du, R. Almajalid, J. Shan, and M. Zhang, "A novel method to predict knee osteoarthritis progression on MRI using machine learning methods," *IEEE Trans. Nanobiosci.*, vol. 17, no. 3, pp. 228–236, Jul. 2018.

[109] Z. Zhuang, L. Si, S. Wang, K. Xuan, X. Ouyang, Y. Zhan, Z. Xue, L. Zhang, D. Shen, W. Yao, and Q. Wang, "Knee cartilage defect assessment by graph representation and surface convolution," *IEEE Trans. Med. Imag.*, vol. 42, no. 2, pp. 368–379, Feb. 2023.

[110] H.-D. Moon, H.-G. Choi, K.-J. Lee, D.-J. Choi, H.-J. Yoo, and Y.-S. Lee, "Can deep learning using weight bearing knee anterio-posterior radiograph alone replace a whole-leg radiograph in the interpretation of weight bearing line ratio?" *J. Clin. Med.*, vol. 10, no. 8, p. 1772, Apr. 2021.

[111] L. Liu, J. Chang, P. Zhang, Q. Ma, H. Zhang, T. Sun, and H. Qiao, "A joint multi-modal learning method for early-stage knee osteoarthritis disease classification," *Heliyon*, vol. 9, no. 4, Apr. 2023, Art. no. e15461.

[112] F. Prezja, J. Paloneva, I. Pölönen, E. Niinimäki, and S. Äyrämö, "DeepFake knee osteoarthritis X-rays from generative adversarial neural networks deceive medical experts and offer augmentation potential to automatic classification," *Sci. Rep.*, vol. 12, no. 1, p. 18573, Nov. 2022.

[113] Z. Wang, A. Chetouani, and R. Jennane, "Key-exchange convolutional auto-encoder for data augmentation in early knee OsteoArthritis classification," 2023, *arXiv:2302.13336*.

[114] Y. Wang, X. Wang, T. Gao, L. Du, and W. Liu, "An automatic knee osteoarthritis diagnosis method based on deep learning: Data from the osteoarthritis initiative," *J. Healthcare Eng.*, vol. 2021, pp. 1–10, Sep. 2021.

[115] Y. Feng, J. Liu, H. Zhang, and D. Qiu, "Automated grading of knee osteoarthritis X-ray images based on attention mechanism," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2021, pp. 1927–1932.

[116] J. Huo, X. Ouyang, L. Si, K. Xuan, S. Wang, W. Yao, Y. Liu, J. Xu, D. Qian, Z. Xue, Q. Wang, D. Shen, and L. Zhang, "Automatic grading assessments for knee MRI cartilage defects via self-ensembling semi-supervised learning with dual-consistency," *Med. Image Anal.*, vol. 80, Aug. 2022, Art. no. 102508.

[117] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: Removing noise by adding noise," 2017, *arXiv:1706.03825*.

[118] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[119] G. H. Chang, D. T. Felson, S. Qiu, A. Guermazi, T. D. Capellini, and V. B. Kolachalama, "Assessment of knee pain from MR imaging using a convolutional Siamese network," *Eur. Radiol.*, vol. 30, no. 6, pp. 3538–3548, Jun. 2020.

[120] M. Dunnhofer, N. Martinel, and C. Micheloni, "Deep convolutional feature details for better knee disorder diagnoses in magnetic resonance images," *Comput. Med. Imag. Graph.*, vol. 102, Dec. 2022, Art. no. 102142.

[121] M. B. Muhammad and M. Yeasin, "Eigen-CAM: Class activation map using principal components," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–7.

[122] M. Bany Muhammad and M. Yeasin, "Interpretable and parameter optimized ensemble model for knee osteoarthritis assessment using radiographs," *Sci. Rep.*, vol. 11, no. 1, p. 14348, Jul. 2021.

[123] S. Aladhadh and R. Mahum, "Knee osteoarthritis detection using an improved CenterNet with pixel-wise voting scheme," *IEEE Access*, vol. 11, pp. 22283–22296, 2023.

[124] M. Kornreich, J. Park, J. Braun, J. Pawar, J. Browning, R. Herzog, B. Odry, and L. Zhang, "Combining mixed-format labels for AI-based pathology detection pipeline in a large-scale knee MRI study," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2022*. Springer: Springer, Sep. 2022, pp. 183–192.

[125] K. Inoue, "Prevalence of large-joint osteoarthritis in Asian and Caucasian skeletal populations," *Rheumatology*, vol. 40, no. 1, pp. 70–73, Jan. 2001.

[126] L. Celar and R. M. J. Byrne, "How people reason with counterfactual and causal explanations for artificial intelligence decisions in familiar and unfamiliar domains," *Memory Cognition*, vol. 51, no. 7, pp. 1481–1496, Oct. 2023.

YUN XIN TEOH received the B.Eng. degree (Hons.) in biomedical engineering (prosthetics and orthotics) from Universiti Malaya, Malaysia, where she is currently pursuing the Ph.D. degree. She is with the Laboratoire Images, Signaux et Systèmes Intelligents (LISSI), Université Paris-Est Créteil, France, for a mobility research stay. Her research interests include medical image analysis, artificial intelligent solutions for clinical problems, and rehabilitation engineering.

ALICE OTHMANI has been an Associate Professor with Université Paris-Est Créteil, France, since 2017. She has been working in several international institutions, such as the École Normale Supérieure de Paris, Collége de France, and Agency for Science, Technology and Research (A*STAR), Singapore. Her research interests include developing computer vision and artificial intelligence solutions for healthcare, emotional intelligence, and psychiatry.

SIEW LI GOH received the Ph.D. degree from the University of Nottingham. Before that, she was involved in network meta-analysis in knee and hip osteoarthritis. She is currently a Clinician and Medical Lecturer in sports medicine at Universiti Malaya. She is leading the Biomechanics Interest Group, under the newly formed Sports and Exercise Medicine Research and Education Group (SEMREG), to advance the research agenda of sports medicine. Her research interests include evidence-based medicine and biomechanics of lower limb.

JULIANA USMAN is currently a Senior Lecturer with the Department of Biomedical Engineering, Universiti Malaya (UM); a member of the Centre for Applied Biomechanics, UM; and an appointed Secretary with Malaysian Society of Biomechanics. Her research interests include sports biomechanics in terms of injury prevention and performance enhancement and motion analysis. She is a member of the Institute of Engineering and Technology, U.K., and the Board of Engineers, Malaysia. She is a certified Chartered Engineer (CEng) from the Engineering Council of U.K.

KHIN WEE LAI (Senior Member, IEEE) received the B.Eng. degree (Hons.) from Universiti Teknologi Malaysia (UTM), Malaysia, and the Ph.D. degree from Technische Universität Ilmenau, Germany, and UTM, under the DAAD Ph.D. Sandwich Programme. He is currently an Associate Professor with the Faculty of Engineering, Universiti Malaya. His research interests include machine learning, medical image processing, and healthcare analytics. He is a registered Professional Engineer with a Practicing Certificate (Ir) of the Board of Engineers Malaysia, a fellow of the Engineers Australia (FIEAust), the Institute of Engineers Malaysia, CEng, U.K., APEC Engineer IntPE, Australia, and a Chartered Professional Engineer (CPEng.) at NER, Australia. He is currently an Associate Editor of *IET Image Processing*.