An XAI Clinical Decision Support System Approach For The Assessment Of Knee Osteoarthritis

Group H3

Timo Debono (13858750), Katrin Grunert (13914103), Andesh Haribhajan (10008454), Jonathan A. Harris, MSc. (13711180), Andreea Mateescu (12263664)
Data System Projects 2021
University of Amsterdam

ABSTRACT

Clinical decision support (CDS) systems have the potential to transform healthcare by synthesizing clinician experience with algorithmically learned medical knowledge. Current diagnoses, like the evaluation of knee osteoarthritis, rely heavily on medical imagining. Artificial intelligence (AI) may improve such manual image classification tasks. However, the adoption of AI-based CDS systems is limited by the lack of interpretability of the algorithms and complex, disruptive workflows. Explainable AI (XAI) approaches exist to identify pixels or image features that influence predictions, yet it is unknown how these methods can be incorporated into a working CDS application. As such, a survey of knee specialists was conducted and identified essential design requirements, including a simple user interface, the ability to preview and annotate the images, and heatmap visualizations with an explanation of model certainty. A CDS web application was successfully developed based on feedback and addressed the limitations mentioned above. Follow-up evaluation indicated that future iterations must be able to locate influential pixels and identify the physical characteristics of osteoarthritis within the highlighted regions.

Link to the CDS prototype: https://github.com/jah377/XAI_ImageClassification/tree/main

KEYWORDS

explainable artificial intelligence, neural networks, clinical decision support systems, digital health platforms

1 INTRODUCTION

The rise of digital information technology has transformed healthcare, resulting in the transition towards electronic health records (EHR) [20]. From 2008 to 2017, the percentage of hospitals in the US using EHR expanded from 9% to over 96%; a survey by the European Commission found similar usage trends [20, 34]. The increasing global digitization of health records has led to the development of computerized Clinical Decisions Support Systems (CDS) capable of supplementing the clinical-decision making process. Typically in the form of web apps, clinicians use CDS to reinforce their knowledge and experience with insights provided by the software [49]. These software systems can include but are not limited to tools for diagnostic evaluation, patient data reporting, clinical workflow, or patient-facing explanation [23, 38]. The use of CDS in the field of diagnostic imaging is of growing interest due to the necessity of arduous manual manipulation and interpretation and advancements in the field of deep learning – particularly in the area of osteoarthritis [3, 28, 45, 49].

The diagnosis of knee osteoarthritis is based on radiographic evaluation and scoring, making it an attractive application of CDS systems. Knee osteoarthritis (KOA) is a degenerative disease affecting over 654 million people worldwide [8]. This polymorphic disorder is defined by the presence and severity of osteophytes (painful boney protrusions), joint space width, subchondral sclerosis (brittle thickening of the bone), and general deformation [25, 57]. While several radiographic grading scales exist, Kellgren & Lawrence (1957) were the first, and now the KL Grade is commonly used to define the presence and severity of KOA [25, 27]. A planar X-ray image is captured, and a clinician provides a KL grade from 0 to 4, with four being the most severe osteoarthritis. Despite the prevalence of the KL grading system, inter- and intra-observer reliability may be poor, even among experienced graders [29, 42]. A review by Sheehy & Cooke (2015) suggests inconsistent interpretations of the KL scale across graders and studies exist [46].

Radiographic evaluation and scoring of KOA is a form of image classification – an image of a knee is classified based on features within the image. Deep learning algorithms have demonstrated high performance on image classification and have only recently been applied to osteoarthritis research [17, 26, 28]. Despite recent success, the perceived "black box" of these algorithms and the inability to visualize factors influencing the prediction limit their usefulness, especially in healthcare where rationale is a requirement for trust [10, 28, 49].

The lack of transparency and interpretability has led to the rise of "Explainable Artificial Intelligence" (XAI), a new area

of research that aims to help human users understand and interpret algorithmic predictions [10]. For example, classic multiple linear regression models return 'feature importance' to articulate which input variables influenced the prediction. As it relates to image classification, model-specific XAI approaches attempt to identify pixel content that affects image classification by exploiting aspects of the predictive model's architecture.

Factors affecting the adoption of CDS fall into two categories: (1) computational, such as algorithmic accuracy and "explainability"; or (2) human-related, such as software that disrupts clinician workflow, ignores human information processing, or requires a high degree of computer literacy [9, 36]. While several studies investigate the computational aspect of predicting KOA, to the author's knowledge, no study has attempted to build a CDS that considers both the "explainability" and human-component [28]. As such, the purpose of the present study is to determine what methods are there to implement, evaluate, and visualize XAI that takes a human-in-the-loop into account for the detection of knee osteoarthritis.

2 RELATED WORKS

Knee Osteoarthritis

The knee joint is the largest joint in the human body, defined by the complex articulation of the distal femur, patellar kneecap, and proximal tibia. Structural support of the knee depends on static capsular ligaments connecting the boney anatomy, while tendons distribute dynamic muscular forces affecting motion and loading; the cup-shaped meniscus separating the fibula and tibia guides rotation and stabilizes translation on the friction-less cartilage [14].

Knee osteoarthritis is a degenerative joint condition associated with cartilage loss and changes to the boney or structural anatomy [4]. Defining characteristics of KOA can include the presence of osteophytes, i.e. boney formations at the joint; loss of joint space width; subchondral sclerosis, i.e. thickening of the hard cortical bone beneath the cartilage; or overall joint deformation [25, 57]. Figure 1 provides a visual summary of structural changes associated with KOA progression [47]. The aetiology of KOA is not well understood; however, factors such as genetics, age-related physiological changes, and knee biomechanics may play a role. Clinical presentation of KOA commonly includes localized pain and stiffness of the joint, especially after physical activity [13].

Osteoarthritis is one of the top causes of impairment globally, affecting 4% of the world's population [37]; KOA accounts for 83% of OA-related diagnoses [27]. If conservative treatment plans fail, invasive total knee arthroplasty (TKA) is recommended to replace the joint. It is projected that the annual number of TKA procedures will grow by 85% (1.3 million) worldwide [15].

Figure 1: Radiographic image of knee osteoarthritis [47]. White arrows on left highlight the presence of osteophytes (abnormal bone protrusions); black arrows emphasize loss of joint space width; the white arrow on right points out a possible sclerosis (the brightened pixels indicate thick bone)

Kellgren-Lawrence Score

Planar radiography remains the most common method to diagnose KOA. Kellgren & Lawrence (1957) were the first to develop a grading system of KOA using X-rays [25] which was later adopted by the World Health Organization as the standard evaluation method [57]. The KL grade is a composite sum score of point evaluations for each defining characteristic previously highlighted in Figure 1. Table 1 describes the point system attributed to the severity of each characteristic. A normal, healthy knee is defined by a KL Grade of 0 (0 points) while higher grades of 1-4 define increased severity (1-2 points, 3-4 points, 5-9 points, and 10 points, respectively) [25, 57]. A KL grade ≥ 2 is commonly used as an indication for knee replacement surgery [57].

Computer Decision Support Systems

Background. CDS systems are a subset of digital health platforms that use EHR. These systems can be used on a wide array of electronic devices such as computers, tablets, or smartphones and are capable of creating output directly on these devices, or EHR databases [49]. In their review of CDS systems, Sutton et al. (2020) identify two categories:

Table 1: Kellgren-Lawrence grade of knee osteoarthritis based on evaluation of defining characteristics [25, 57]

Osteophyte formation	none: 0 definite: 1 large: 2
Joint space width	normal: 0 narrowing: 1 advanced: 2
Subchondral sclerosis	none: 0 discrete: 1 discrete w/ cyst: 2 severe w/ cyst: 3
Deformation	none: 0 discrete: 1 strong: 2

knowledge-based systems, where a program generates an action or output from physician-defined rules (i.e. If-Then statements); or non-knowledge-based systems, the decision relies on learning algorithms instead of user medical knowledge [49]. Presently, CDS software has been leveraged to improve patient safety [35], manage patient follow-up [32], contain medial costs [2], support administrative functions [18], and most relevant to the current study – diagnostics support [41]. The application of non-knowledge based CDS within diagnostic radiology has attracted research labs from IBM Watson Health, DeepMind, Google, and others. Several studies have highlighted the utility of CDS software for tumor detection [59], interpretation of medical imaging [22], diabetic-related deterioration of the eye [39], and Alzheimer's diagnosis [50].

Surgimap®. One CDS available on the market that significantly informed the design of the proposed application is Surgimap®(Nemaris Inc, New York, NY)[1]. Created by surgeons for surgeons (Figure 2), Surgimap®is a "free computer program that integrates radiographic measurements and tools for surgical planning with knowledge gained from published literature" [1]. The software provides digital measuring tools and corrective surgery simulation specifically modelled after clinicians' existing workflows. This CDS example emphasizes simplicity, ease-of-use, non-disruptive workflow, and features like the integration of patient information and image analysis into a single report.

[1]www.surgimap.com

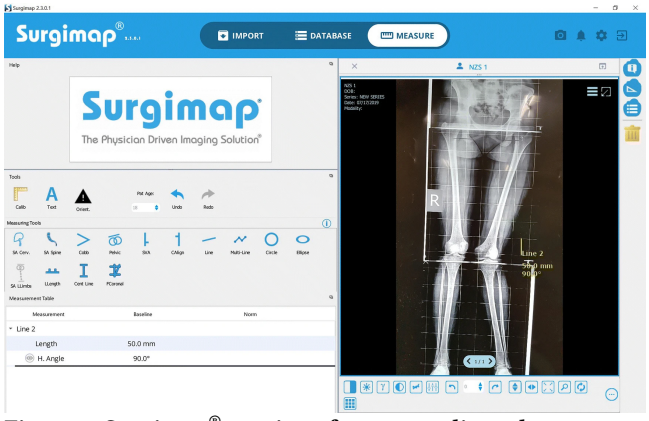

Figure 2: Surgimap®user interface streamlines the importation, preview and analysis, and reporting of patient-specific radiographic evaluation.

Explainability in Deep Learning

Taxonomy. Recent breakthroughs in deep learning for computer vision have produced results in which computers models outperformed humans on a variety of prediction tasks. Alexnet defeated the runner-up in the ImageNet Large Scale Visual Recognition Competition (ILSVRC) in 2012 [30]. Since then, more advanced approaches have enhanced Alexnet's accuracy. By 2015, Microsoft's ResNet had outperformed human prediction in the same competition [19]. Despite their greater performance, these complicated deep learning models have been slow to catch on, owing to poor model interpretability, the need for many annotated examples to train the models, and the need for advanced computational resources. The lack of model clarity in its predictions, in particular, has hampered regulatory approval for deployment in life-critical healthcare applications [5].

The goal of XAI methods is to facilitate understanding of the model predictions made by deep networks. Several taxonomies for classifying XAI methodologies have been developed in the literature. In general, these classifications are not absolute; they can vary greatly depending on the methods' properties, and they can be categorized into multiple overlapping classes simultaneously [48]. One such proposed taxonomy introduces a three-dimensional classification framework [58]:

- **Visualization methods**: Express an explanation by systematically highlighting characteristics of a given input that contributed the most to the model's prediction.
- **Model distillation methods**: Develop a separate, intrinsically explainable surrogate model to mimic the behaviour of the black-box model to identify relevant input features.

- **Intrinsic methods**: Created specifically to produce an explanation along with the predicted output.

Although we have implemented at least one method from each category (Visualization: Grad-CAM, Score-CAM, Integrated Gradients; Distillation: LIME, SHAP; Intrinsic: Attention Map), based on the methodology for selecting the most appropriate XAI approach for our solution introduced in the subsequent section, the focus hereafter shall lie on a backpropagation-based visualization method, Grad-CAM.

Gradient-weighted Class Activation Mapping (Grad-CAM). Grad-CAM is post-hoc explainability method, i.e. an approach for producing heatmaps applied to a convolutional neural network after training is complete and the parameters are fixed. It uses the gradients of any class flowing into the final convolutional layer to localize class-discriminative regions in an image for predicting a specific class, c, and creates a localization map thereof.

Formally, to generate such maps, the first step is to compute the gradient of y^c, the model output for class c before applying softmax to transform this score into a probability, with respect to the feature map activations A^k of the corresponding convolutional layer, i.e., $\frac{\partial y^c}{\partial A^k}$. The particular value of the gradient calculated here depends on the input image chosen [44]. Subsequently, the so-called neuron importance weights, α_k^c, are calculated. This is done by applying the global average pool operation, $g(x) = \frac{1}{Z}\sum_i \sum_j x_{ij}$, to the previously obtained gradients over the width (i) and height (j) dimensions, i.e.

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}.$$

Lastly, to obtain the final Grad-CAM map, each α_k^c is used as the weight of the corresponding feature map, A^k, and a weighted sum thereof is calculated. Then, the ReLU operation is applied to emphasize positive values and transform all negative values to zero, i.e.

$$L_{Grad-CAM}^c = ReLU(\sum_k \alpha_k^c A^k).$$

Variations of Grad-CAM, e.g. Score-CAM [56], differ primarily in combinations of gradients to represent α_k^c [44].

3 METHODOLOGY

Data Source

Data used in the present study was accessed from the "Knee Osteoarthritis Severity Grading Dataset" [7]. Available to the public, the dataset contains approximately 9,900 de-identified plain radiographs of the knee joint and ground truth KL grade labels. All five KL grades are represented. Normal knee radiographs (KL=0) represent 38% of the dataset, followed by 18%, 26% (KL=1 and KL=2 respectively). The worse degeneration classified as KL=4 and KL=5 represents 13% and 5% of the dataset, respectively.

Design Process

Interviews. The two stakeholders in the current study are a digital healthcare provider *Quin*[2] and clinicians. Quin is actively developing a DHP that supports AI-powered decisions in radiograph analysis. However, medical experts like clinicians will be the end-user of the proposed CDS. Essential design requirements from both corporate and clinical stakeholders must be identified.

The primary motivation of the corporate stakeholder was to adapt explainability tools for the medical domain in an attempt to demystify and explain medical AI decisions. Multiple rounds of interviews established three desired objectives of the CDS system: to develop a central platform with intuitive UI, to evaluate XAI components using quantitative or qualitative methods, and to design XAI visualizations tailored to the use and needs of relevant clinical specialists.

A total of six physicians specializing in KOA were surveyed to better understand existing clinical workflows and identify the critical design considerations from a clinical perspective. The formalized survey can be found in the appendix (Survey A.1). Surveys were conducted over Google Meet[3].

There was a consensus among physicians. Planar radiographs were ordered if the patient was under extreme discomfort; if images were taken, they would be presented and explained. All desired to inspect and interact with the radiographs, annotate, and summarize detailed findings into a report. Additionally, the procedural steps within the CDS application needed to resemble the diagnostic workflow; image upload, preview, analysis, and report stages were most appropriate. One physician stated that the ideal CDS system should aid the user in discovering observations that they would have otherwise missed. Lastly, all respondents emphasized simplicity, ease of use, the ability to revise. Perhaps most important, none were interested in the ability to select multiple AI models or a wide variety of XAI visualization techniques. Informal follow-up correspondence was conducted to identify preferred XAI visualization approaches but are discussed later.

Product Design. Following interviews of the corporate and clinical stakeholders, requirements of the front-facing user interface (UI) and behind-the-scenes modelling were distilled into crucial design requirements, as summarized in Figure 3.

The architecture in Figure 4 illustrates a high-level view of the prototype. It consists of a UI that provides interaction

[2]https://quin.md/en

[3]https://meet.google.com/

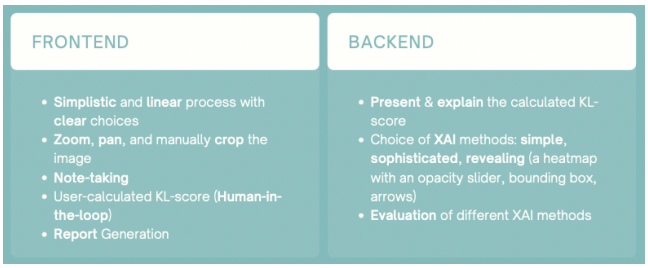

Figure 3: Summary of design requirements extracted from stakeholder interviews.

possibilities for the end-user, and the back-end, which inhabits the Deep Learning model and the XAI methods. UI and back-end are implementing a loosely coupled client-server-communication via REST API.

The UI was built on top of the market-leading web framework React[4]. React enables the use of reusable components and building blocks, facilitating a consistent user experience and design throughout the entire application. The back-end was built with Python[5], as it is the state-of-the-art technology for machine learning and artificial intelligence-powered applications.

The prototype implements a diagnosis flow that is divided into four stages. In the first step, which functions as a preparation step, the user can upload an X-ray and enter relevant patient information. From there, the X-ray can be inspected more closely by zooming and panning, and the user can enter their notes and evaluation of the X-ray. Next to their evaluation, a Deep Learning model analysis is displayed to the user's notes. The analysis consists of a KL-score prediction and visual explanations that the Deep Learning model calculates. The clinician can consider these results and potentially second guess their evaluation. It is to be emphasized that the model's decision is no definite conclusion but rather an evaluation that enhances the clinician's decision. Notes and information entered by the clinician and the results from the Neural Net are summarized and persisted in a report document which can be added to the patient's EHR.

Model Training. Despite not being the focus of this research, we trained a variety of different deep learning models using the *Keras*[6] library. We used the ImageDataGenerator class for image preprocessing to generate batches of tensor image data with real-time data augmentation (rescale, shear, zoom, rotation, flipping) for all training runs, regardless of the model type. We then proceeded with instantiating the model architectures directly from Keras. In cases where we do not use ImageNet weights, we loaded the models, including the classification head with randomly initialized weights.

[4]https://reactjs.org/
[5]https://www.python.org/
[6]https://keras.io/

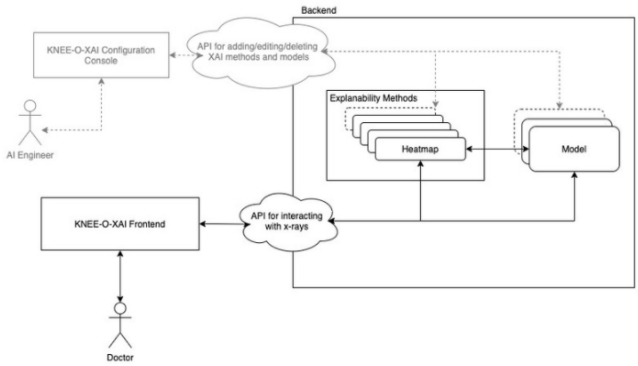

Figure 4: Rendered drawing of the software architecture of the proposed CDS system.

We consequently loaded the model architecture for the Visual Transformer where ImageNet weights are being used, excluding the classification head. Instead, we attached a customer classifier to the pre-trained base. In sequential order, this custom classification head consists of a flatten layer, a batch normalization layer, a dropout layer ($p = 0.2$), a dense layer ($n = 15$) with GeLU activation, another batch normalization layer, and, lastly, a dense layer ($n = 5$) with softmax activation. Irrespective of the model architecture, we specified a learning rate of $1e4$ and the number of training epochs as 75.

Moreover, we used Rectified Adam, a more robust implementation of the widespread Adam algorithm, as an optimization algorithm [33]. In addition, we used Sparse Categorical Crossentropy for a loss function, which computes the cross-entropy loss between labels and predictions. Lastly, we defined two callbacks to mitigate overfitting before initiating the training process. First, we defined a callback for reducing the learning rate upon hitting a plateau (*ReduceLROnPlateau*). Second, we defined an early stopping callback that terminates the training process when no significant improvements in validation loss have been identified over the past 15 epochs. To keep track of the different model specifications, we used the Weights & Biases[7] platform.

Model Evaluation. The primary criterion for evaluating and selecting the model connected to the final architecture was test-set accuracy. This metric captures the fraction of correctly classified images in the test set. We acknowledge that using this metric as a criterion is debatable, precisely given that we are dealing with a classification problem on an ordinary scale. An alternative that could have been considered here is the linearly weighted kappa score, which implicitly takes into account that a model prediction that is off by one is less severe than a prediction that is off by, e.g., three [31].

[7]https://wandb.ai/

XAI Method Selection. To select the primary XAI method among the range of approaches that were implemented, as briefly described in the background section, to display to the user, we employed a two-step process (Figure 5).

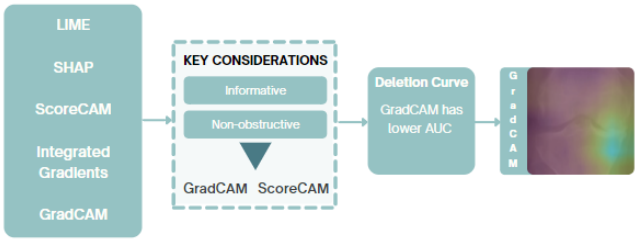

Figure 5: XAI method selection funnel.

In the first step, we drew on data gathered from the interviews with professionals in the medical domain. We found that the most frequently mentioned characteristics expected from an explainable visualization are informativeness and non-obstructiveness. With these two critical data points in mind, we narrowed the available options down to two CAM-based implementations: Grad-CAM and Score-CAM. As a final step, we calculated the area under the curve (AUC) for the deletion curve for each of the two remaining XAI methods and compared the results. In general, to derive the deletion curve, one pixel is iteratively deleted from the most pronounced region of the heatmap that was previously generated from an image. Subsequently, a prediction is made on the partially occluded image, and the predicted probability for the image's target class, as well as the percentage occlusion of the heatmap centre, are stored. This process is repeated until the heatmap centre is entirely deleted. The resulting values are averaged over a batch of 20 images. The AUC was then derived from plotting the average predicted target class probability against the percentage occlusion. A lower AUC implies that the respective XAI method captures important image regions more precisely. The specific results of this will be covered in the Results section of this paper.

Note that instead of simply displaying the heatmap in the final product, we went further and derived a bounding box and arrows from the most pronounced regions of the heatmap. In brief, to obtain the former, we analyzed the variance of pixel values above a certain threshold and then used this to isolate the corner points of the bounding box. We subsequently derived arrows pointing towards the region with the highest pixel intensity from these corner points.

Prototype Evaluation

Follow-up surveys were conducted to determine whether the CDS prototype met their requirements and identify possible deficiencies. Kappa scores evaluating the inter-observability of the predicted KL scores was not conducted as the present study's focus was to determine favourable XAI approaches and workflow options.

4 RESULTS

Model and XAI Method Analysis

In terms of test set accuracy, we achieve the following results for the different model architectures that were considered:

Table 2: Test accuracy for explored model architectures.

Model	Test accuracy
ViT (ImageNet)	0.715
ViT	0.679
DenseNet-169	0.637
ResNet-152	0.536
VGG-19	0.529

As is evident from Table 2, the Visual Transformer (ViT) using pre-trained ImageNet weights and then fine-tuning this architecture on the task at hand yields the best performance in terms of test accuracy (0.715).

Although this ViT was the best performing model, we ultimately decided to opt for the DenseNet-169 model. The reason behind this is that a Visual Transformer setup primarily allows for one architecture-specific visualization, i.e., attention maps (Figure 6).

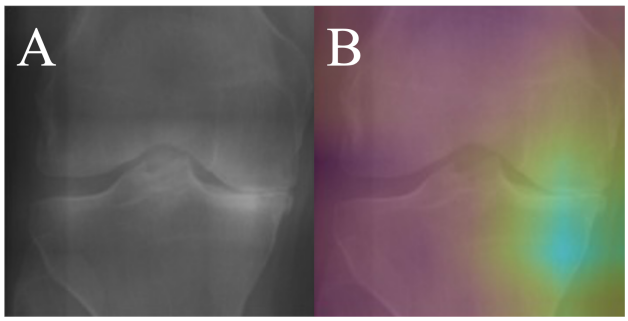

Figure 6: (1) Visual Transformer attention map and (2) Grad-CAM heatmap.

However, we a priori discarded this XAI method due to physicians' feedback that had ranked this visualization as less favourable. Consequently, this made us discard the transformer-based model architectures and choose the next-best option. To select the primary visualization that will ultimately be shown to the user, we follow the funnel logic introduced in Figure 5. As described in the methodologies section, we arrived at the final result by first evaluating the available approaches against the two critical criteria issued by physicians, i.e. *non-obstructiveness* and *informativeness*, and then calculating the AUC for the deletion curve of the remaining options. We find that the deletion curve for Grad-CAM (Figure 7) has the lowest AUC (0.0481).

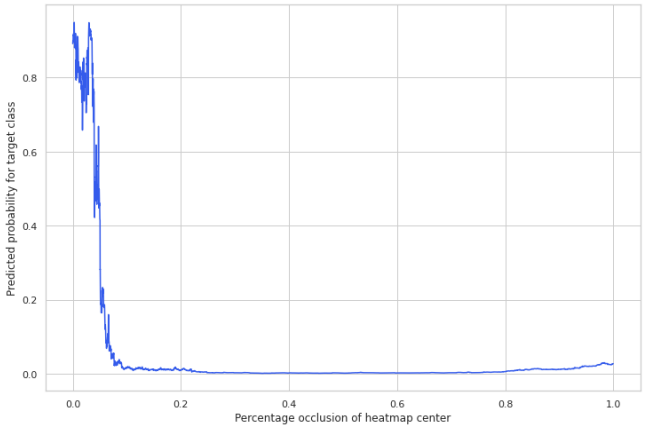

Figure 7: Grad-CAM Deletion Curve.

The final visualization using Grad-CAM, in addition to bounding box and arrow-based visualizations derived from the gradients (Figure 8).

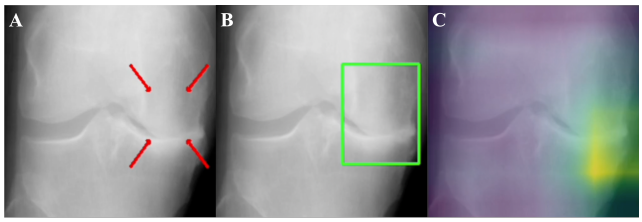

Figure 8: Representative images of XAI visualization methods (A) arrows, (B) bounding box, and (C) Grad-CAM heatmap.

Product Workflow

A video walk of the proposed CDS platform is presented[8]. The app opens with a landing page that provides a short description of the product and its use cases (Figure A.2). The bottom buttons lead to a tutorial page, a more detailed description of the app, and a contact form, respectively. On the upload page, relevant patient information can be entered, such as name, date of birth, date of appointment, name of the physician, and initial medical notes. Additionally, an X-ray image of the patient's knee can be uploaded to the front-end architecture (Figure A.3).

Once completed the upload stage, the user proceeds to the Preview Page (Figure A.4). In this step, the X-ray can be examined closer by zooming and panning, and the physician's initial evaluation, examining the four defining characteristics, osteophytes, joint space width, sclerosis, and deformation, which are used to calculate a KL-score. If needed, notes about the X-ray and evaluation can be entered.

Next, the image is sent via an API to the back-end architecture where the DenseNet-169 model analyzes the X-ray and the results are displayed to the user (Figure A.5). The results consist of the predicted KL-score and the probability distribution of the prediction. Furthermore, the user can explore different visualization methods by selecting a method from the dropdown menu (Arrows, Bounding Box, Heatmap as shown in Figure 8), and adjust the opacity of the overlayed explanation layer, or hiding the explanation, to view the X-ray unobstructed. Then, the Analysis page provides functionality to edit previously entered information, for example, adjusting the physician's initial evaluation of the KL-score and its characteristics or adding more notes.

Lastly, a report can be generated (see examples in Figures A.6, A.7, A.8, A.9). The report compiles all the user-defined information and the results of the AI into one PDF that can be downloaded. If needed, the user can return a previous stages in order to edit information that is displayed in the report.

Prototype Evaluation

While all six clinicians were contacted, only two responded by the submission deadline of this report. The physicians noted that the landing page identified the use case of the CDS system (Figure A.2). They also noted that the application was simple and easy to use, pointing out the marked header indicating their current location in the workflow and the coloured buttons to proceed to the following pages. Although the preview page enabled them to zoom and pan the image, they would prefer to have the option to view the image in full screen (Figure A.4). Still, the physician appreciated the need to evaluate the sub-components of KOA – e.g. select the grade for osteophytes – and calculate the KL score before the analysis was presented.

As it relates to the analysis page, there was some confusion about the bar graph (Figure A.5); the physician asked if the numbers represented *certainty* or *probability* of the class prediction. Previously, clinicians voiced their approval of the heatmap, bounding box, and arrow-based XAI visualizations. The interviewed physician was pleased that they could effortlessly select the visualization and opacity. Furthermore, they commented that the visualization and generated report would be a helpful education tool when discussing the results with patients.

When asked why this prototype may not be helpful in its current form, one physician commented on the incompleteness of the XAI visualization. Although the XAI approaches informed them which regions influenced the predicted KL score, the system cannot identify the presenting medical conditions – osteophytes, sclerosis, joint space width, or deformity. Ideally, the system would locate and quantify the aforementioned structural irregularities. One unavoidable limitation of the existing prototype is that the recent concern

[8]https://vimeo.com/673296028

in radiation exposure to the patient has resulted in few X-rays; a responding physicians noted that radiographic images are only ordered if the patient presents extreme discomfort.

5 DISCUSSION

Clinician Decision Support Systems

There exists a paucity in research evaluating CDS system recommendations. Nehrer et al. (2019) assessed the effect of a CDS that evaluated KL grade, joint space narrowing, osteophytes and sclerosis [43]. Researchers compared analyses between two groups: plain radiographs and plain radiographs supplemented with a report generated from their CDS system. Unsurprisingly, they found that CDS increased the accuracy and consistency between physicians evaluating KOA. It is uncertain whether the increase in consistency was due to physicians defaulting to the results from the CDS. That uncertainty represents a significant gap in their research methodology.

Saban et al. (2022) conducted a pilot study investigating the acceptance and reliability of the European Society of Radiology's (ESR) iGuide, a CDS system that makes imaging referral recommendations (i.e. which patients should receive which medical imaging test) [43]. Researchers asked four experts (two radiologists, two physicians) to evaluate 40 simulated clinical cases on a 5-level scale based on their agreement with the ESR iGuide recommendations. Overall, there was an agreement between the experts 77.3% of the time, with an average iGuide agreement rating of 4.2±0.7 where 5 represents complete agreement with the CDS system.

The methodology presented by Saban et al. should be commended and represents a standard as further investigations evaluate new CDS systems. The present study similarly attempted to quantify the acceptance and reliability of the proposed CDS system evaluating radiographs for the detection of KOA. However, evaluation was limited to qualitative survey and assessment instead of a quantitative 5-scale system. The lack of physician response to follow-up inquiry in the present study precluded thorough evaluation. Only one physician provided a qualitative evaluation of the current CDS prototype, primarily focusing on the UX aspect of the program.

Explainable AI

Despite numerous studies investigating image classification or XAI model approaches, few are within the context of knee osteoarthritis evaluation. Tiulpin et al. (2018) first proposed a ResNet34 model to predict a KL grade and explained their results using GradCAM attention maps, resulting in an average multi-class accuracy of 66.7% [54]. Similarly, Thomas et al. (2020) found that deeper CNN-based models improved accuracy to 0.71, whereas the physician cohort achieved an accuracy of 0.60 [52].

Karim et al. (2021) is the most relevant example with the creation of `DeepKneeExplainer` [24]. The researchers focused on the 'computational' aspect of CDS by proposing a diagnosis pipeline that predicts a KL grade using magnetic resonance imaging (MRI) and planar radiographs. Following preprocessing, novel VVGNet, ResNet, and DenseNet CNN-based models were used to define regions of interest in corresponding radiographs, and MRI and KL grades were predicted; Grad-CAM, Grad-CAM++, and LRP attention maps were also compared. The study found that DenseNet and VGG architectures achieve 91% accuracy. Perhaps most importantly, their survey of orthopaedic surgeons identified the XAI visualization approach, Grad-CAM++, as providing the most reliable interpretable explanation.

The present study found that DenseNet-169 and the Visual Transformer model achieved an overall test accuracy of 63.7% and 71.5%, respectively. While these results are lower than published literature [24, 52, 54], model accuracy was not the sole focus of the study. In theory, any of the models proposed in the literature may be applied within our CDS framework. Compared with transformer-based models, the surveyed physicians preferred the CNN-based GradCam XAI approach over attention maps. These results are in agreement with Karim et al. (2021) [24].

Legal and Ethical Considerations

Clinical decision support systems necessitate infrastructure that enables physicians to query patients-generated EHRs, and software to process EHR and inform subsequent medical decisions. As these infrastructures and systems are designed and implemented to improve the quality of care, legal oversight and ethical implications must be considered.

In the United States, the Food and Drug Administration (FDA) regulates software considered a 'medical device' – programs or algorithms intended to affect the human body (e.g. a cardiac pacemaker) or used in the diagnosis, cure, or prevention of disease [21]. The broad scope of CDS systems functions obfuscates clear, actionable regulatory pathways. For example, the use case and inherent risk of the proposed CDS are fundamentally different from algorithms automating the detection of breast cancer. This nuance motivated the '21st Century Cures Act' passed in 2016, updating previous regulatory oversights to emphasize a risk-based approach. As a result, low-risk clinical decision support technologies are now excluded, provided the manufacturer provides sufficient evidence.

The intended use of the proposed CDS system is for the evaluation of knee osteoarthritis and should be defined and regulated as a medical device. Nevertheless, the CDS system should be considered low-risk as the workflow emphasizes

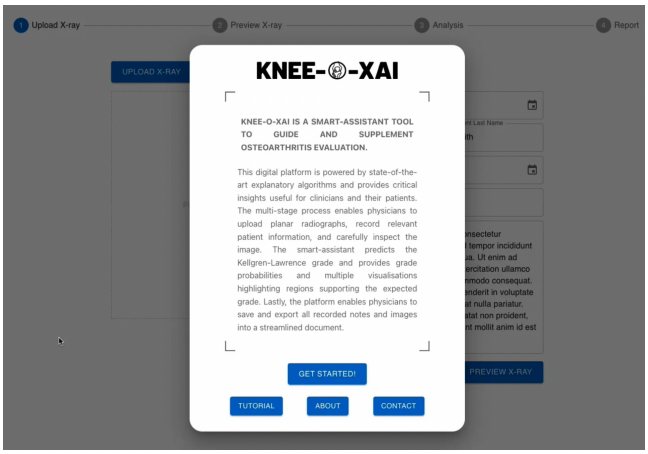

Figure 9: Image of the landing page articulating conditions of use of the proposed CDS system.

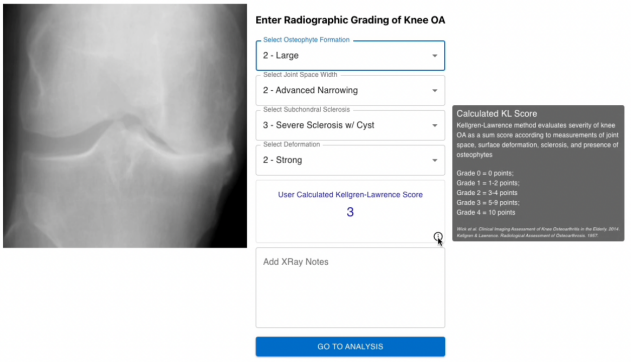

Figure 10: Example of supplemental information (grey box) added to proposed CDS system to better inform user interpretation of output.

human-in-the-loop components requiring the physician to preview and annotate the images prior to any algorithmic prediction or analysis – unlike the detection of breast cancer example. Furthermore, the proposed system only evaluates knee structure; clinical diagnosis of knee osteoarthritis is based on physical examination, medical history, in addition to radiographic evaluation. The multi-modal dependencies of KOA diagnosis reduce the overall risk posed to the patient by the CDS system.

Irrespective of legal definitions, CDS-related risk may not concern harm to the human body but rather what a clinician does with the information provided by the software or algorithms. As such, Evans & Whicher (2018) outline relevant ethical and regulatory oversight that aims to ensure appropriate use and interpretation of CDS output, validation of algorithms and data, and sufficient privacy protections [11].

Critical design decisions were made that aim to follow their recommended guidelines, as well as more recent guidance outlined by the FDA [55]. First, a landing page on the web app was created to articulate the appropriate conditions for using the software, as shown in Figure 9.

Furthermore, supplemental information concerning KL score calculations (Figure 10), software features, and certainty of the score predictions was included to inform the interpretation of the output report. Validation of algorithms and data was also considered, as the architecture enables an AI engineer to update or modify AI models and XAI methods (Figure 4).

The privacy and protection of sensitive patient EHR are critical for healthcare professionals as well as software designers [51]. The design of the proposed CDS system did not specifically address patient privacy but was considered indirectly. Currently, the CDS system is hosted on a single server where the model is stored. Only the de-identified image is transferred to the back-end architecture while sensitive patient information remains on the UI side and is not stored during the creation of the report. In theory, the application may be hosted on a local, encrypted server at the hospital; communication to the back-end architecture does not occur on the cloud. Additionally, identity and access management (IAM) to the patient EHR are inherent in using the web application; thus, EHR infrastructure was not considered in the present study. Similarly, model building only required de-identified source images and labelled KL scores – remote updates to the model training protocol could be performed assuming approval of the patients, physicians, and institutions.

These indirect attempts to address patient privacy may be naive. Rahulamathavan et al. (2013) conceptualized privacy-preserving CDS systems that make predictions on encrypted data, thus eliminating the concern of raw patient data transmission to third-party remote servers [40]. However, these approaches were not explored in the current study. Development of multi-modal classification models that predict KOA using raw radiographic data, clinical examination results and previous medical history may further complicate concerns about privacy protection; and present new ethical issues about model fairness and unintended predictors related to race or gender [53].

Automation Bias and Knowledge Erosion

The integration of technology into nearly all aspects of human life has ignited the concern that an over-reliance may lead to knowledge erosion. The classic example of this debate is the concern that the use of calculators negatively impacts mental math development in children. In 2014, the UK government went as so far as to ban calculators in primary school classrooms [6]. Like the previous example, 'automation bias' is also a concern for CDS system use [16], where the clinician may become dependent on the CDS program

for a task. While these concerns are valid, they may not be relevant to the proposed CDS system. The ultimate aim of the current platform is to supplement the evaluation of radiographic images used to assess KOA. This use case is educational, and human-in-the-loop components may reinforce expertise instead of creating over-reliance. The proposed system first asks users to evaluate and annotate the image; then, the automated analyses are presented. Furthermore, the XAI visualization components provide an additional opportunity where the user can compare their interpretation to the interpretation of the AI model.

Limitations

While the present study successfully created a CDS system that considers both the "explainability" and human element for detecting knee osteoarthritis, it is not without limitation. The reduced number of surveyed physicians precluded statistical and inter-rater reliability analyses from thoroughly evaluating the CDS prototype. By the submission of this paper, only one physician responded to the follow-up questions. Additionally, and perhaps most significantly, the lack of data granularity minimizes the utility of the web application. The KL grade represents a composite sum score representing the presence and severity of joint space width loss, osteophytes, sclerosis, and deformation. Explainable AI can identify pixels that influence the prediction of the KL score, yet those pixels represent something physical – the characteristics above of KOA. Without granular data describing these components, a more robust explanation is impossible. The Osteoarthritis Initiative (OAI)[9] database contains the granularity required for such analyses. Available for public access, the database includes quantitative measurements of joint space width and sclerosis, number of osteophytes, and qualitative assessment of deformity and KL score – for both the medial and lateral knee joint. However, the database contains terabytes of EHR, often unorganized, which made it unfeasible to use in the current study. Nevertheless, the ability of AI to identify individual characteristics would make the program more valuable to physicians, especially considering that the severity of some features instead of others may impact the decision to perform invasive knee arthroplasty [12].

6 CONCLUSION

The present study proposed and developed an end-to-end clinical decision support system that leverages plain radiographs and out-of-the-box AI/XAI algorithms to assess knee osteoarthritis. Surveys conducted identified vital design requirements of physician end-users. Experiments using publically available radiographic data yielded between 63.7% and 71.5% accuracy. Regions of interest found in the explainable AI visualizations suggest that inference of KL scores were based on medically-relevant features of knee osteoarthritis. Nevertheless, the inability of XAI methods to identify the presenting conditions contained in these regions precisely limit the adoption of these systems.

[9]https://data-archive.nimh.nih.gov/oai/

ACKNOWLEDGEMENT

The authors would like to acknowledge Drs. Jaggoe-Haribhajan G., Drs. Ooft. JR, Drs. Lierop Van M.G., Bersaoui M., MSc, van der Schouw M., DC, and Dr. A. Abu-Awwad, MD, PhD, for providing invaluable clinical perspective during the ideation and evaluation phase of the study. Additionally, the authors would like to acknowledge Dr. Cristian Rivero (UvA) and Majd Zreik (Quin) for their mentorship and critical input throughout the study.

REFERENCES

[1] Michael Akbar, Jamie Terran, Christopher P Ames, Virginie Lafage, and Frank Schwab. 2013. Use of Surgimap Spine in sagittal plane analysis, osteotomy planning, and correction calculation. *Neurosurgery Clinics* 24, 2 (2013), 163–172.

[2] Claudia A Algaze, Matthew Wood, Natalie M Pageler, Paul J Sharek, Christopher A Longhurst, and Andrew Y Shin. 2016. Use of a checklist and clinical decision support tool reduces laboratory use and improves cost. *Pediatrics* 137, 1 (2016).

[3] Bibb Allen, Robert Gish, and Keith Dreyer. 2019. The role of an artificial intelligence ecosystem in radiology. In *Artificial Intelligence in Medical Imaging*. Springer, 291–327.

[4] Roy Altman, E Asch, D Bloch, G Bole, D Borenstein, K Brandt, W Christy, TD Cooke, R Greenwald, Mea Hochberg, et al. 1986. Development of criteria for the classification and reporting of osteoarthritis: classification of osteoarthritis of the knee. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology* 29, 8 (1986), 1039–1049.

[5] Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, and Vince I. Madai. 2020. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making* 20 (12 2020), 310. Issue 1. https://doi.org/10.1186/s12911-020-01332-6

[6] Judith Burns. 2012. Government bans calculators from primary maths tests. *BCC* (Nov 2012). https://www.bbc.com/news/education-20259382

[7] Pingjun Chen. 2018. Knee osteoarthritis severity grading dataset. https://data.mendeley.com/datasets/56rmx5bjcr/1

[8] Aiyong Cui, Huizi Li, Dawei Wang, Junlong Zhong, Yufeng Chen, and Huading Lu. 2020. Global, regional prevalence, incidence and risk factors of knee osteoarthritis in population-based studies. *EClinicalMedicine* 29 (2020), 100587.

[9] Srikant Devaraj, Sushil K Sharma, Dyan J Fausto, Sara Viernes, Hadi Kharrazi, et al. 2014. Barriers and facilitators to clinical decision support systems adoption: A systematic review. *Journal of Business Administration Research* 3, 2 (2014), 36.

[10] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. 2018. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE, 0210–0215.

[11] Emily L Evans and Danielle Whicher. 2018. What should oversight of clinical decision support systems look like? *AMA journal of ethics* 20, 9 (2018), 857–863.

[12] DT Felson, DR Gale, M Elon Gale, J Niu, DJ Hunter, J Goggins, and MP Lavalley. 2005. Osteophytes and progression of knee osteoarthritis. *Rheumatology* 44, 1 (2005), 100–104.

[13] David T Felson. 2009. Developments in the clinical understanding of osteoarthritis. *Arthritis Research Therapy* 11 (2009), 203. Issue 1. https://doi.org/10.1186/ar2531

[14] Fred Flandry and Gabriel Hommel. 2011. Normal anatomy and biomechanics of the knee. *Sports medicine and arthroscopy review* 19, 2 (2011), 82–92.

[15] Jiaxiang Gao, Dan Xing, Shengjie Dong, and Jianhao Lin. 2020. The primary total knee arthroplasty: a global analysis. *Journal of orthopaedic surgery and research* 15 (2020), 1–12.

[16] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. 2011. Automation Bias–A Hidden Issue for Clinical Decision Support System Use. *International Perspectives In Health Informatics* (2011), 17–22.

[17] Tianmei Guo, Jiwen Dong, Henjian Li, and Yunxing Gao. 2017. Simple convolutional neural network on image classification. In *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*. IEEE, 721–724.

[18] Shoshana Haberman, Joseph Feldman, Zaher O Merhi, Glenn Markenson, Wayne Cohen, and Howard Minkoff. 2009. Effect of clinical-decision support on documentation compliance in an electronic medical record. *Obstetrics & Gynecology* 114, 2 Part 1 (2009), 311–317.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *CoRR* abs/1512.03385 (2015). arXiv:1512.03385 http://arxiv.org/abs/1512.03385

[20] Jeff Hecht. 2019. The future of electronic health records. *Nature* 573, 7775 (2019), S114–S114.

[21] Gail H Javitt. 2018. Regulatory Landscape for Clinical Decision Support Technology. *Anesthesiology* 128, 2 (2018), 247–249.

[22] Kwang Nam Jin, Eun Young Kim, Young Jae Kim, Gi Pyo Lee, Hyungjin Kim, Sohee Oh, Yong Suk Kim, Ju Hyuck Han, and Young Jun Cho. 2022. Diagnostic effect of artificial intelligence solution for referable thoracic abnormalities on chest radiography: a multicenter respiratory outpatient diagnostic cohort study. *European radiology* (2022), 1–11.

[23] Stefane M Kabene. 2010. *Healthcare and the Effect of Technology: Developments, Challenges, and Advancements*. IGI Global (701 E. Chocolate Avenue, Hershey, Pennsylvania, 17033, USA).

[24] Md Rezaul Karim, Jiao Jiao, Till Doehmen, Michael Cochez, Oya Beyan, Dietrich Rebholz-Schuhmann, and Stefan Decker. 2021. DeepKneeExplainer: Explainable Knee Osteoarthritis Diagnosis From Radiographs and Magnetic Resonance Imaging. *IEEE Access* 9 (2021), 39757–39780.

[25] Jonas H Kellgren and JS1006995 Lawrence. 1957. Radiological assessment of osteo-arthrosis. *Annals of the rheumatic diseases* 16, 4 (1957), 494.

[26] S Kluzek and TA Mattei. 2019. Machine-learning for osteoarthritis research. *Osteoarthritis and cartilage* 27, 7 (2019), 977–978.

[27] Mark D Kohn, Adam A Sassoon, and Navin D Fernando. 2016. Classifications in brief: Kellgren-Lawrence classification of osteoarthritis. *Clinical Orthopaedics and Related Research®* 474, 8 (2016), 1886–1893.

[28] C Kokkotis, S Moustakidis, E Papageorgiou, G Giakas, and DE Tsaopoulos. 2020. Machine learning in knee osteoarthritis: A review. *Osteoarthritis and Cartilage Open* 2, 3 (2020), 100069.

[29] Özkan Köse, Baver Acar, Fatif Çay, Baris Yilmaz, Ferhat Güler, and Halil Yalçın Yüksel. 2018. Inter-and intraobserver reliabilities of four different radiographic grading scales of osteoarthritis of the knee joint. *The journal of knee surgery* 31, 03 (2018), 247–253.

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.), Vol. 25. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[31] Tarald O. Kvålseth. 2018. An Alternative Interpretation of the Linearly Weighted Kappa Coefficients for Ordinal Data. *Psychometrika* 83 (2018), 618–627.

[32] Raymond Kwok, Michael Dinh, David Dinh, and Matthew Chu. 2009. Improving adherence to asthma clinical guidelines and discharge documentation from emergency departments: implementation of a dynamic and integrated electronic decision support system. *Emergency Medicine Australasia* 21, 1 (2009), 31–37.

[33] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2019. On the Variance of the Adaptive Learning Rate and Beyond. *CoRR* abs/1908.03265 (2019). arXiv:1908.03265 http://arxiv.org/abs/1908.03265

[34] F Lupiañez Villanueva, F Folkvord, and C Fauli. 2018. Benchmarking deployment of eHealth among general practitioners. *RAND. org* (2018).

[35] Charles D Mahoney, Christine M Berard-Collins, Reid Coleman, Joseph F Amaral, and Carole M Cotter. 2007. Effects of an integrated clinical information system on medication safety in a multi-hospital setting. *American Journal of Health-System Pharmacy* 64, 18 (2007), 1969–1977.

[36] B Middleton, DF Sittig, and A Wright. 2016. Clinical decision support: a 25 year retrospective and a 25 year vision. *Yearbook of medical informatics* 25, S 01 (2016), S103–S116.

[37] Juan C Mora, Rene Przkora, and Yenisel Cruz-Almeida. 2018. Knee osteoarthritis: pathophysiology and current treatment modalities. *Journal of Pain Research* Volume 11 (10 2018), 2189–2196. https://doi.org/10.2147/JPR.S154002

[38] Z Omididan and AM Hadianfar. 2011. The role of clinical decision support systems in healthcare (1980-2010): A systematic review study. *Jentashapir Sceintific-Research Quarterly* 2, 3 (2011), 125–34.

[39] Jerome A Osheroff, Jonathan M Teich, Donald Levick, Luis Saldana, Ferdinand T Velasco, Dean F Sittig, Kendall M Rogers, and Robert A Jenders. 2012. *Improving outcomes with clinical decision support: an implementer's guide*. Himss Publishing.

[40] Yogachandran Rahulamathavan, Suresh Veluru, Raphael C-W Phan, Jonathon A Chambers, and Muttukrishnan Rajarajan. 2013. Privacy-preserving clinical decision support system using gaussian kernel-based classification. *IEEE journal of biomedical and health informatics* 18, 1 (2013), 56–66.

[41] Salman Razzaki, Adam Baker, Yura Perov, Katherine Middleton, Janie Baxter, Daniel Mullarkey, Davinder Sangar, Michael Taliercio, Mobasher Butt, Azeem Majeed, et al. 2018. A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis. *arXiv preprint arXiv:1806.10698* (2018).

[42] Daniel L Riddle, William A Jiranek, and Jason R Hull. 2013. Validity and reliability of radiographic knee osteoarthritis measures by arthroplasty surgeons. *Orthopedics* 36, 1 (2013), e25–e32.

[43] Mor Saban, Jacob Sosna, Clara Singer, Sharona Vaknin, Vicki Myers, Dorit Shaham, Jacob Assaf, Alon Hershko, Paula Feder-Bubis, Rachel Wilf-Miron, et al. 2022. Clinical decision support system recommendations: how often do radiologists and clinicians accept them? *European radiology* (2022), 1–7.

[44] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. 2016. Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization. *CoRR* abs/1610.02391 (2016). arXiv:1610.02391 http://arxiv.org/abs/1610.02391

[45] Puneet Sharma, Michael Suehling, Thomas Flohr, and Dorin Comaniciu. 2020. Artificial intelligence in diagnostic imaging: status quo, challenges, and future opportunities. *Journal of thoracic imaging* 35 (2020), S11–S16.

[46] Lisa Sheehy and T Derek V Cooke. 2015. Radiographic assessment of leg alignment and grading of knee osteoarthritis: A critical review.

World 2 (2015).

[47] Milena Simic, Alison R Harmer, Maria Agaliotis, Lillias Nairn, Lisa Bridgett, Lyn March, Milana Votrubec, John Edmonds, Mark Woodward, Richard Day, et al. 2021. Clinical risk factors associated with radiographic osteoarthritis progression among people with knee pain: a longitudinal study. *Arthritis research & therapy* 23, 1 (2021), 1–10.

[48] Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. 2020. Explainable deep learning models in medical image analysis. *CoRR* abs/2005.13799 (2020). arXiv:2005.13799 https://arxiv.org/abs/2005.13799

[49] Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine* 3, 1 (2020), 1–10.

[50] Kenji Suzuki and Yisong Chen. 2018. *Artificial intelligence in decision support systems for diagnosis in medical imaging*. Vol. 140. Springer.

[51] Rayhan A Tariq and Pamela B Hackert. 2018. Patient confidentiality. (2018).

[52] Kevin A Thomas, Łukasz Kidziński, Eni Halilaj, Scott L Fleming, Guhan R Venkataraman, Edwin HG Oei, Garry E Gold, and Scott L Delp. 2020. Automated classification of radiographic knee osteoarthritis severity using deep neural networks. *Radiology: Artificial Intelligence* 2, 2 (2020), e190065.

[53] Aleksei Tiulpin, Stefan Klein, Sita Bierma-Zeinstra, Jérôme Thevenot, Esa Rahtu, Joyce van Meurs, Edwin HG Oei, and Simo Saarakkala. 2019. Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data. *Scientific reports* 9, 1 (2019), 1–11.

[54] Aleksei Tiulpin, Jérôme Thevenot, Esa Rahtu, Petri Lehenkari, and Simo Saarakkala. 2018. Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. *Scientific reports* 8, 1 (2018), 1–10.

[55] U.S. Food Drug Administration. 2019. Clinical Decision Support Software: Draft Guidance for Industry and Food and Drug Administration Staff. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software.

[56] Haofan Wang, Mengnan Du, Fan Yang, and Zijian Zhang. 2019. Score-CAM: Improved Visual Explanations Via Score-Weighted Class Activation Mapping. *CoRR* abs/1910.01279 (2019). arXiv:1910.01279 http://arxiv.org/abs/1910.01279

[57] Marius C Wick, Martin Kastlunger, and Rüdiger J Weiss. 2014. Clinical imaging assessments of knee osteoarthritis in the elderly: a mini-review. *Gerontology* 60, 5 (2014), 386–394.

[58] Ning Xie, Gabrielle Ras, Marcel van Gerven, and Derek Doran. 2020. Explainable Deep Learning: A Field Guide for the Uninitiated. *CoRR* abs/2004.14545 (2020). arXiv:2004.14545 https://arxiv.org/abs/2004.14545

[59] Xiaoyao Zhao, Yinbin Zhang, Xingcong Ma, Yinxi Chen, Junfeng Xi, Xiaoran Yin, Huafeng Kang, Haitao Guan, Zijun Dai, Di Liu, et al. 2020. Concordance between treatment recommendations provided by IBM Watson for Oncology and a multidisciplinary tumor board for breast cancer in China. *Japanese journal of clinical oncology* 50, 8 (2020), 852–858.

APPENDIX

Survey A.1 Design Interview

(1) What is the series of decisions/evaluations/diagnoses that occur from when a patient comes to them complaining about their knee to surgery?
(2) What imaging modalities are most prevalent?
    (a) Under what circumstances do you choose one over the other?
(3) How does radiography inform your decision-process? What measurements do you take?
    (a) What measurements do you take?
(4) What criteria are there for categorising an X-ray image on the KL-scale?
    (a) How rigid is the 0-4 scale; would a binary categorization (0-1 & 2-4) suffice?
(5) What does the process of explaining an X-ray image to a patient following a diagnosis (OA vs. no-OA) look like?
(6) Are there currently any technological tools being used to support the decision-making process when diagnosing knee OA?
(7) Assuming you had a tool that could automate the classification process and explain its classification decision:
    (a) What factors would you want to have explained?
    (b) How would you want to view the explanations?
        (i) Would you prefer a purely visual explanation (e.g. area highlighting)?
        (ii) Would you prefer a heat-map, bounding box, or arrows to identify influential regions?
(8) To what extent would you, as expert and practitioner, want to be involved in the decision-making or classification process?

Survey A.2 CDS Prototype Evaluation

Concept validation (to be answered before testing)

(1) Have you previously seen or used other tools to aid the diagnostics process?
    (a) How have the previous solutions fail?
(2) Do you think you need a product to aid you in the knee osteoarthritis diagnostics process?
(3) What do you expect to be able to do with the proposed system?

Prototype evaluation (to be answered after testing)

(1) Do you understand what our product does?
(2) How easy/difficult is the product to use?
(3) How closely do the procedural steps in the application mirror your existing workflow?
(4) Are there any features you wish were included in the prototype?
(5) Are there any features that were unnecessary?
(6) What features did you like or dislike?
(7) Do you think the visualisations are sufficiently informative?

(8) What about the current design limit your adoption of the product? What is the added value of this CDS in your current diagnostic process?

Figures

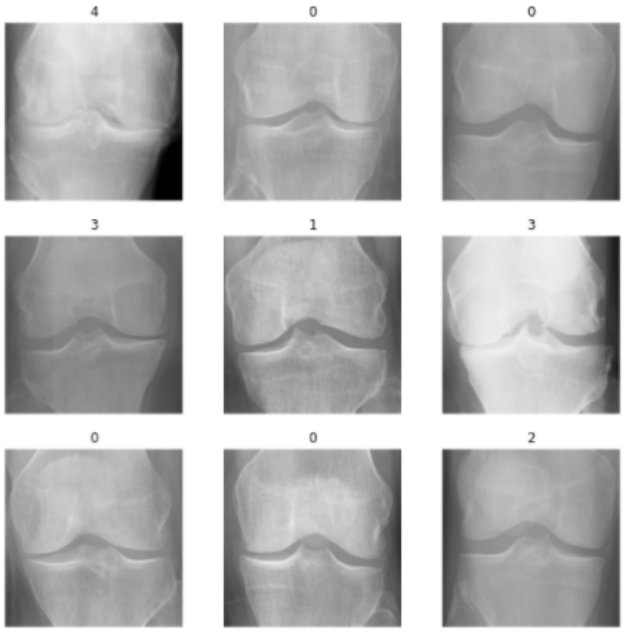

Figure A.1: Representative X-ray images with corresponding ground truth KL grades.

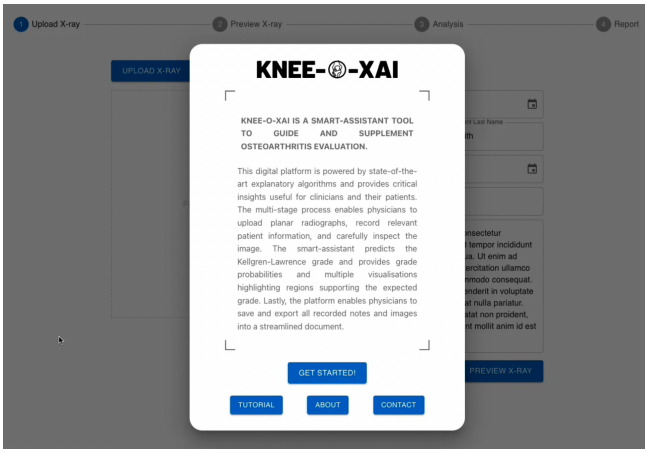

Figure A.2: Image capture of landing page.

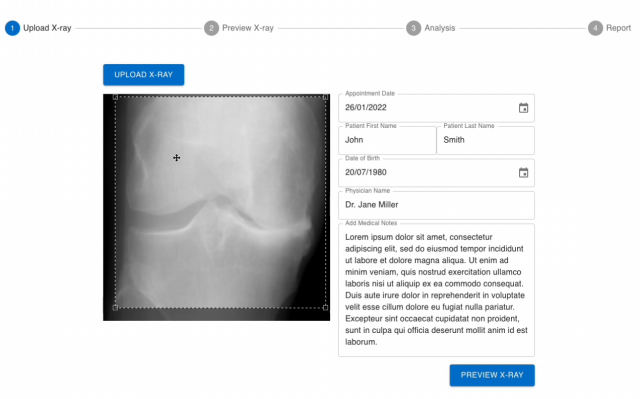

Figure A.3: Image capture of upload page.

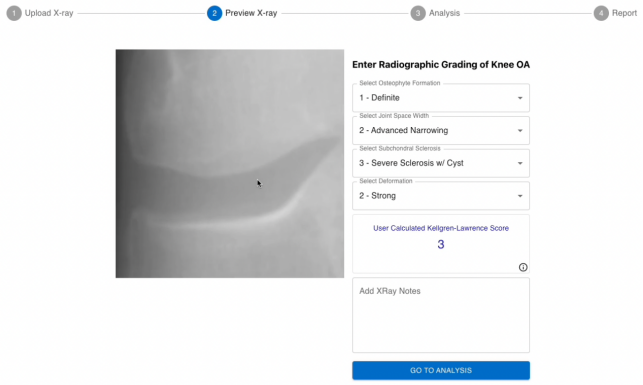

Figure A.4: Image capture of preview page. Note that the user has the ability to inspect the image via panning and zooming.

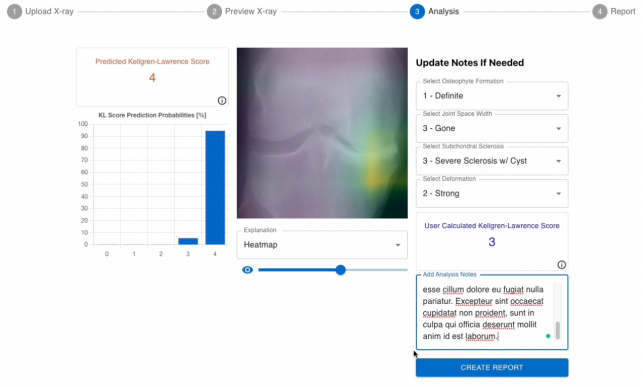

Figure A.5: Image capture of analysis page with GradCAM heatmap XAI visualization.

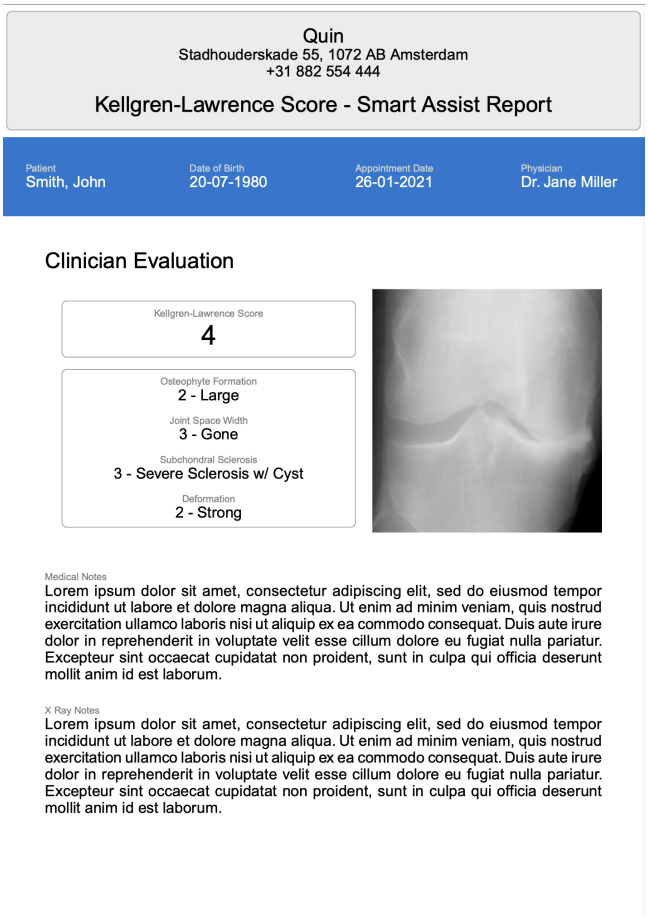

Figure A.6: Page 1 - Clinician evaluation- of an example generated report

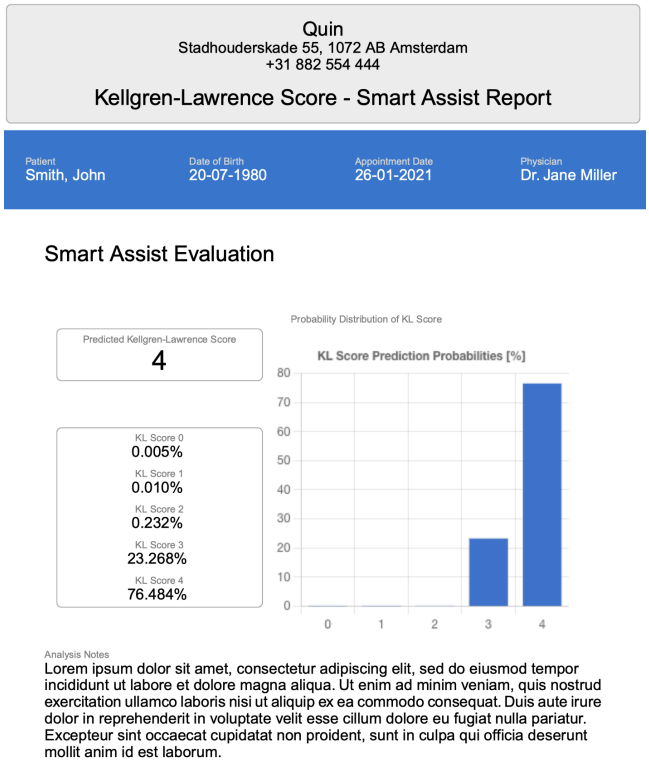

Figure A.7: Page 2 - Smart Assist Evaluation - of an example generated report

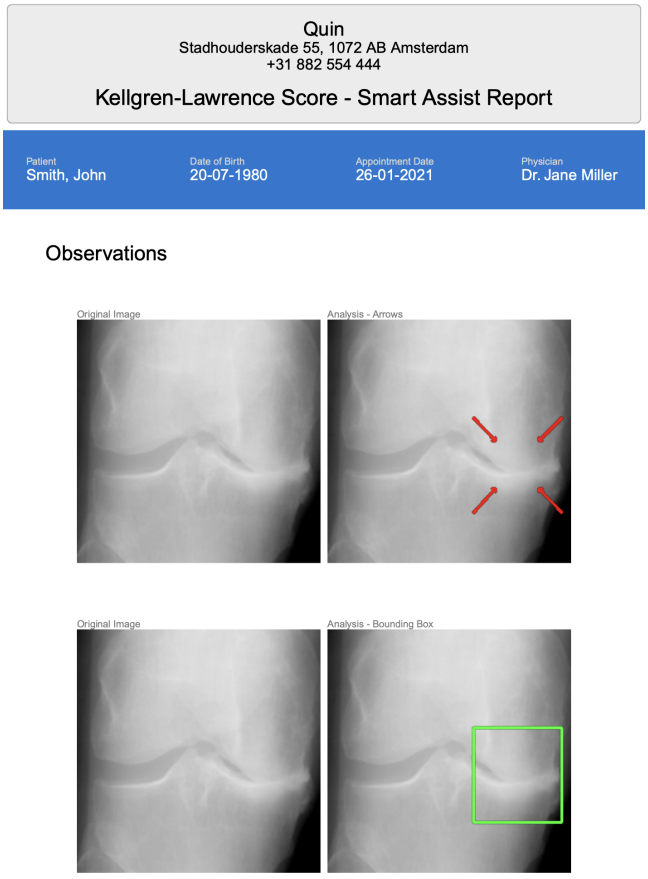

Figure A.8: Page 3 - Smart Assists Observations - of an example generated report

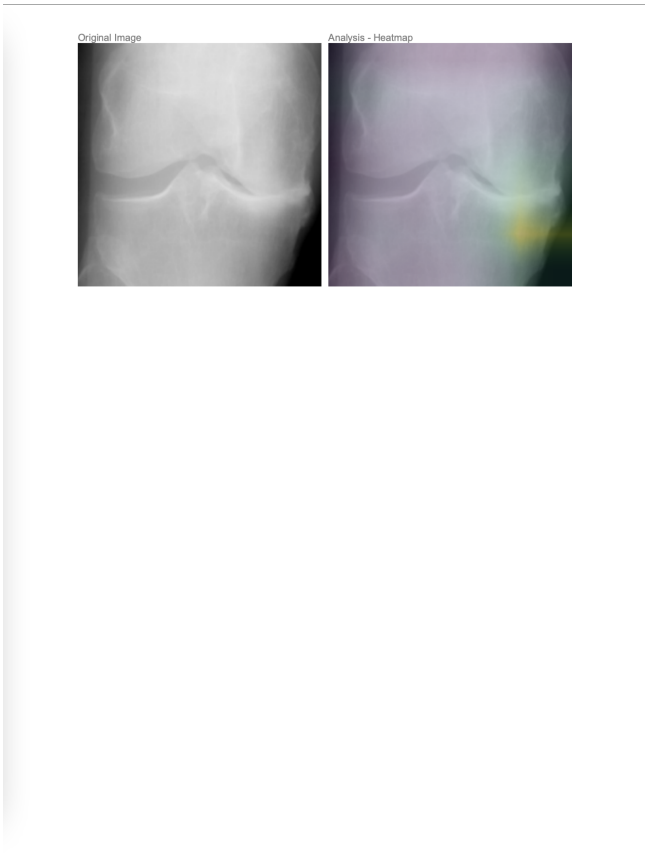

Figure A.9: Page 3 cont. - Smart Assists Observations - of an example generated report