Quantifying Contributions within Multimodal Fusion for Clinical Decisions

Allison Clement and Irina Voiculescu

Oxford University Department of Computer Science
allison.clement@cs.ox.ac.uk

Abstract. Many medical conditions are diagnosed using a combination of image findings and patient-specific clinical information. While medical images provide valuable diagnostic information, clinical decision-making also relies on additional clinical data, such as age or sex. Integrating imaging and patient-specific clinical data has the potential to improve classification beyond what either modality can achieve alone.
We propose a novel contribution-based fusion (CBF) method to combine image-derived measurements with clinical data for disease classification. Unlike many existing multimodal methods that rely on entangled latent representations and focus primarily on predictive performance, our CBF method explicitly quantifies the contribution to the final decision of each modality and its specific inputs. We demonstrate our method on the task of screening for Developmental Dysplasia of the Hip (DDH), where patients are classified as normal or abnormal.
Our proposed method achieves comparable classification performance to established multimodal fusion methods, while providing input feature-level (and therefore modality-level) contributions that are reliably associated with the model's decision-making process.
By reporting how much each modality (imaging vs. clinical data) influenced a decision, the proposed method enables direct quantification of how image and clinical data influence the final classification, providing information that may help clinicians trust the model's predictions.

Keywords: Contributions · Clinical Decisions · Explainability

1 Introduction

Clinical decision-making frequently relies on information from multiple sources. Medical images can provide detailed pathological or anatomical information (e.g. measurements), while clinical data (e.g. patient demographics, medical history) provide complementary context that influences diagnosis, prognosis, and treatment planning [19]. Effectively combining these heterogeneous data sources remains a challenge in the development of automating clinical decision-support systems.

Within medical imaging, landmark localisation is an established approach for deriving anatomical measurements that support clinical decision-making [6]. These landmark-based measurements often form the basis of diagnostic and

treatment decisions. In practice, clinicians often do not rely on imaging measurements alone and typically integrate these measurements with additional patient-specific information (clinical data) when making decisions [38,30,29,31]. This motivates the need for multimodal methods that can jointly utilise image-derived measurements and clinical data while maintaining explainability.

Recent multimodal methods have improved predictive performance by combining imaging and clinical data across many tasks [9,13,37]. However, fusion often relies on highly entangled latent representations, making it difficult to determine how individual modalities and features influence the predictions [15,11]. As a result these methods provide limited explainability, restricting clinicians' ability to verify model recommendations, hindering adoption [27,32].

In this work, the term *explainability* refers to the extent to which a model's reasoning and decision-making process can be understood. This includes both *clinical explainability*, which enables clinicians to interpret and assess model behaviour, and *model explainability*, which enables a model to provide inherent or post-hoc explanations for its predictions. All these properties are important in clinical decision support, where clinicians must be able to understand and trust model recommendations before incorporating them into practice.

This work develops a multimodal fusion framework that combines landmark-derived measurements and clinical data for diagnostic classification. The proposed method explicitly quantifies the contribution of each feature to the final prediction, enhancing model explainability and clinician understanding while maintaining predictive performance. By doing so, the framework aims to improve explainability to support building trust in clinical decisions.

Research Advances. In this work we develop a novel contribution-based fusion method that combines image-derived measurements and clinical data while explicitly quantifying a per-patient input feature contribution (and therefore modality-level) to the final classification. Specifically, we:

(i) Demonstrate that proposed contributions can be obtained without sacrificing **classification performance**.
(ii) Show the reliability of the contributions using **feature-deletion curves**.
(iii) Validate the contributions through **Shapley correlations**.

2 Related Work

2.1 Clinical Background

Motivation. Medical data that is used for the diagnosis and treatment planning varies according to the pathology and clinical objective. Depending on the anatomical structures of interest, clinical standards and protocols may require one or more imaging modalities (such as Computed Tomography, X-ray or Ultrasound). In orthopaedics, conditions such as Osteoarthritis (OA), Developmental Dysplasia of the Hip (DDH), and Adolescent Idiopathic Scoliosis (AIS), are influenced by image findings (e.g. measurements) as well as clinical data (e.g. age, sex) [38,30,29,31]. Clinical context is essential for accurate radiological

interpretation, with 85% of radiologists incorporating clinical data when evaluating images and formulating diagnoses [19], suggesting imaging alone cannot fully capture a patient's condition. Integrating patient-specific data with imaging findings can improve diagnostic accuracy [23]. Consequently, clinical assessment often integrates imaging findings with patient-specific clinical data.

Developmental Dysplasia of the Hip (DDH). In this work, DDH screening serves as an example application combining image-derived measurements and clinical data for clinical decisions. DDH is characterised by inadequate femoral head alignment with the acetabulum, causing hip instability. Early diagnosis enables non-surgical treatment, whereas delayed detection may require surgery.

DDH screening uses a landmark-based workflow (Figure 1) in which clinicians identify landmarks on images to derive measurements that are compared against predefined diagnostic thresholds. These image-derived measurements are often considered alongside clinical data when making clinical decisions. The two methods most commonly used in clinical practice for DDH screening are: Femoral Head Coverage (FHC) [25] and the Graf method [12]. The FHC method quantifies the proportion of the femoral head covered by the acetabulum (healthy when $FHC{>}50\%$, unhealthy otherwise). The Graf method measures the alpha (α) angle between the iliac baseline and the bony roof line (healthy when $\alpha{\geq}60°$, unhealthy otherwise). As both methods use a threshold-based categorisation, small changes in landmark placement can influence the screening decision, particularly in cases near the diagnostic boundary.

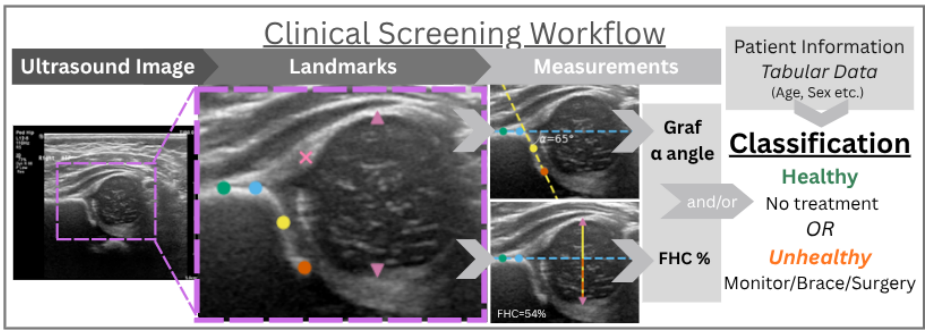

Fig. 1: DDH Screening. Landmarks on ultrasound images used to calculate the α angle and FHC %. The iliac points (green/blue), turning point (yellow), lower limb point (orange), labrum (pink), and upper and lower femoral head (purple) are shown. The α angle: measured between the iliac baseline (blue) and bony roof line (yellow). The FHC %: calculated as the acetabular depth (d, orange dashed) divided by the femoral head diameter (D, yellow solid). These measurements are combined with clinical data to classify patients (healthy/unhealthy).

2.2 Landmark-based Measurements in Medical Imaging

Evaluation of orthopaedic conditions such as OA, DDH, and AIS relies on landmark-based measurements from medical images. Automated methods can achieve clinician-level performance while producing measurement outputs that align with clinical workflows, improving the ability for clinicians to understand a model's predictions [6,4,17]. By following clinical workflows these models provide some explainability, which can increase clinician trust and support adoption.

Current automated landmark-based methods achieve performance within clinically acceptable inter-rater variability compared with experts [5]. As performance approaches clinician-level agreement, recent work has shifted towards confidence and uncertainty estimation rather than further improving accuracy [7], helping identify predictions that may require additional review. However, clinical diagnosis rarely relies on imaging alone, and integrating routinely collected clinical data with image-derived measurements therefore represents a clinically motivated extension of existing landmark-based methods.

2.3 Multimodal Fusion

Medical data used in clinical decision-making is heterogeneous, varying in structure, representation, and content [2]. Clinical decisions are informed by imaging modalities containing pixel information alongside patient-specific clinical data (Section 2.1), represented as tabular data or text. These modalities differ in the information they encode. When combined, information may be redundant, complementary, or have complex interactions [21]. To address this, multimodal fusion has been proposed: early, intermediate, and late fusion [22,32]. These approaches differ in how multimodal information is combined, resulting in trade-offs.

Early fusion combines modalities at the input level, typically by concatenating image-derived and clinical features, and is currently the most common strategy for medical imaging fusion [14,8,9]. However, in early fusion the differences in modality dimensionality can cause one modality to dominate learning. Intermediate fusion learns modality-specific features and models their interactions within a shared latent space [10,15,11,20]. While often achieving strong performance, the resulting latent representations can become highly entangled, making modality contributions on the final prediction difficult to interpret. Late fusion trains separate models for each modality and combines their outputs at the decision level. Although simple and allowing direct assessment of modality influence, it fails to capture complex inter-modal relationships [14,24,32].

Reviews consistently conclude that no single fusion strategy is universally optimal; the most appropriate approach depends on the task, data characteristics, and clinical requirements [18,22]. Despite often improving predictive performance, many fusion methods remain difficult to interpret, limiting their use in clinical decision support where explainability is essential [1,8,32].

Explainability The terms encompassing *explainability* are used inconsistently in the literature, and it is beyond the scope of this paper to disambiguate them.

As such, for clarity, the following definitions are adopted. *Explainability* is defined as the extent to which the reasoning and decision-making process of a model can be understood. *Explainability* includes both: *model explainability*, the feasibility of the model to provide explanations for model decisions, and *clinician explainability*, which we will call *clinician understanding*, the degree to which a human user can comprehend the outputs and reasoning of a model. Together, these factors can influence clinicians' *trust* in the model: a clinician's belief in the model's ability to deliver accurate, reliable and safe outcomes [35,3,16].

Multimodal methods can report model explainability in several ways. First, some models exhibit model explainability by design, where the decision-making process can be directly followed from the model structure (e.g. decision trees, logistic regression). Second, models can provide clinically meaningful intermediates that align with clinical workflows. For example, providing a segmentation or relevant measurements alongside a classification allows clinicians to inspect the anatomical structures influencing the prediction. This improves clinician understanding compared to a classification output alone [35]. Third, post-hoc model explainability techniques can be applied to trained multimodal medical imaging models to provide additional insight into their decision-making processes. For example, saliency maps (e.g. Grad-CAM) are often used to identify image regions associated with a prediction [15]. Feature-attribution methods, such as Shapley (SHAP) values, can be used to report the influence of individual features on model outputs [33]. Attention has been incorporated into fusion to highlight modality and feature interactions [11]. Uncertainty-aware fusion has quantified modality-specific confidence in predictions [36].

All these methods can improve explainability in general. However, model explainability methods do not generally provide a direct decomposition of the model's prediction into quantified contributions from individual inputs. In multimodal settings, few fusion approaches explicitly quantify and validate such feature- and modality-level contributions to the final diagnostic decision.

3 Methods

Dataset. DICOM ultrasound scans (1,107) were acquired from Alder Hey Children's Hospital hospital, with a resolution of 1024×768 pixels. Annotations were collected from the same clinician for each image using `GIMP` software [34]. Landmarks were 25 pixels in diameter and were confirmed by two expert clinicians. Landmarks were used to calculate α and FHC %, with ground-truth classes derived from their combined classification [6].

The dataset was split into training (70%), validation (15%) and testing (15%). The dataset classes are the healthy (60.3%) vs. unhealthy (39.7%) groups. Clinical data included sex (58.0% female, 42.0% male), laterality (55.5% left, 44.5% right), and age (days: mean 64.7, SD 32.0). These class proportions and clinical data distributions were retained across the splits.

Data was collected retrospectively from routine screening at Alder Hey Children's Hospital. All datasets were anonymised and stored within the Nuffield De-

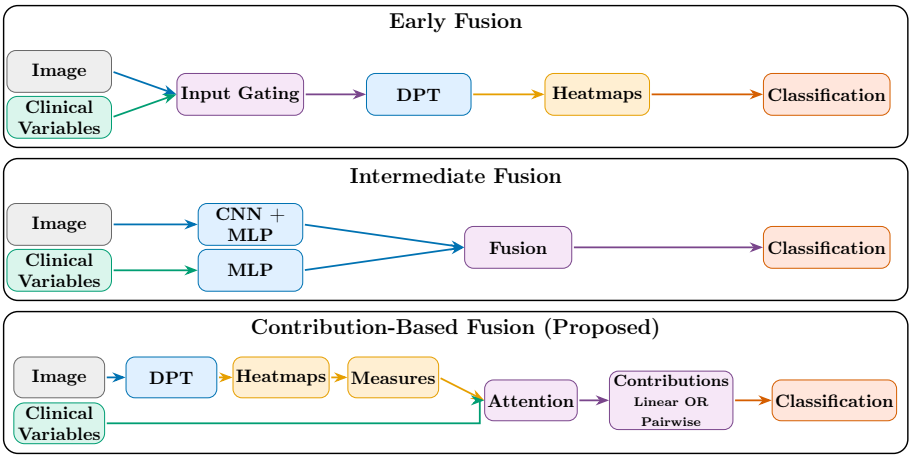

Fig. 2: Fusion. Grey: image-processing. Green: Clinical data. Blue: Model components (CNN, DPT, and MLP). Yellow: Landmark heatmaps and image-derived measurements. Purple: fusion/contribution. Orange: classification outputs. Early fusion integrates clinical data before image feature extraction. Intermediate fusion concatenates independently learned image and clinical data representations prior to classification. CBF predicts landmark-based measurements from the image and then combines with clinical data to allow for feature-level contributions.

partment of Orthopaedics. Ethical approval was obtained from the NHS Health Research Authority (ID 316325; REC exemption Ref 23/HRA/0966) and Oxford University. The authors acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work [28].

3.1 Model Architectures

Figure 2 summarises the architecture types compared in this study. We compared the proposed method with common early- and intermediate-fusion approaches, as well as image-only baselines. Late fusion was not included because the focus was on methods that combine individual imaging-derived and clinical inputs before the final decision. Evaluation against late fusion remains future work. Models are summarised below. The code is publicly available on `GitHub` (see multimodal-medical-confidence).

Image-Only. Two image-only methods were evaluated: a Dense Prediction Transformer (DPT) model [26] that predicted anatomical landmarks for classification, and a CNN that performed direct classification.

Early Fusion. As an early-fusion baseline, we used a metadata-gated Dense Prediction Transformer (DPT). Clinical variables (sex, laterality, and age) were

converted into patient-specific channel weights that modulated the image before feature extraction. The DPT predicted anatomical landmark heatmaps, from which measurements and the final classification were derived. Model explainability was provided through both the channel weights and predicted landmarks.

Intermediate Fusion. As an intermediate-fusion baseline, we used a CNN-linear model. Ultrasound images were encoded using a ResNet-50 backbone, while clinical variables were processed by a multilayer perceptron. The resulting image and clinical embeddings were concatenated and passed through a fusion layer to produce the final classification. This approach enabled joint learning of both imaging and clinical modalities while preserving modality-specific representations prior to fusion.

Contribution-based Fusion (Proposed). To retain explainability while allowing patient-specific contributions, we implemented a simple attention model that combined image-derived landmark measurements and clinical data. Image features were obtained using the same DPT landmark localisation model as the early fusion method. Landmark heatmaps were used to estimate α angle, FHC %, and their associated confidence estimates [7]. These features can be understood by clinicians, as they align with existing clinical workflows. Given an input vector $x \in \mathbb{R}^F$, the model computed patient-specific attention weights across all features. The attended features were then passed to an additive prediction head, which applied a small per-feature transformation before summing contributions (which are later detailed as the contributions in Figure 3) to produce the final logit. This design preserved a per-feature (and therefore modality-level) decomposition of the prediction enabling patient-specific feature contribution estimates. For visualisation, contributions were normalised per patient and expressed as percentages of the final prediction. Figure 3 illustrates this decomposition for an individual patient, showing how each image-derived measurement and clinical variable contributes towards the final classification.

Training. Images were padded and resized to 384×384. Augmentations were randomly selected and applied to each image at training time. All models were trained using an Adam optimizer (10-50 epochs, early stopping). Loss functions were selected according to the task.

3.2 Evaluation

We evaluated the proposed method from three perspectives: model performance, the validity of the estimated feature contributions, and agreement between the proposed explanations and those produced by established explanation methods.

(i) Classification Performance. We compared the proposed model against other fusion and baseline methods using standard classification metrics, including accuracy, precision, recall, and F1-score.

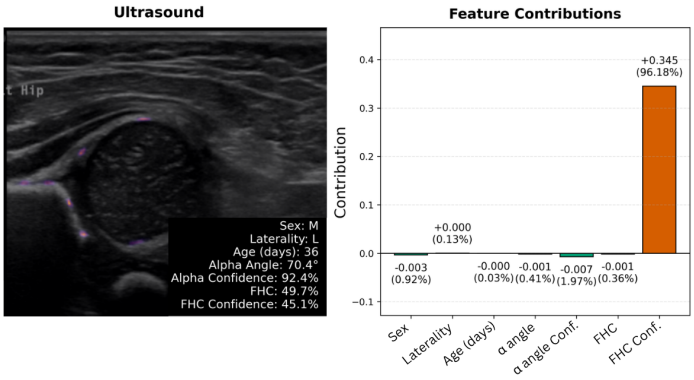

Fig. 3: Left: Ultrasound image (and landmark heatmaps) of a 'borderline' case, in which α and FHC % are both close to their decision boundaries and disagree on patient classification individually (if one says unhealthy, the decision unhealthy). Patient-specific features, including clinical data (sex, age, and laterality) and image-derived features (α, FHC %, and their confidence estimates), are shown in the lower right-hand corner. Right: Patient-specific feature contributions. Red indicates positive contributions toward unhealthy classification, where green indicates contributions to a healthy classification.

(ii) Feature-deletion Curves. To verify that the contribution estimates reflect the model's decision process, we created a feature-deletion curve. Features were ranked by their contributions and each feature was then progressively removed, with model performance re-evaluated after feature removal. Reliable contributions should show a rapid decline in performance when the highest-ranked features are removed.

(iii) Shapley Correlations. SHAP values are widely used to estimate feature importance and explain model predictions in machine learning [3,33,35]. We compared the proposed contribution estimates with SHAP values to assess whether both methods identify similar patterns of feature importance.

4 Results and Discussion

Fig. 3 shows a borderline case (both the predicted α angle and FHC % lie close to their thresholds with conflicting classes). The prediction is driven primarily by FHC confidence, which strongly contributes towards the unhealthy class. The α confidence provides the largest opposing contribution towards a healthy classification, as expected in this borderline case where the two measurements provide conflicting evidence. Sex provides the third largest towards a healthy class, consistent with the higher prevalence of DDH in females, while laterality contributes towards the unhealthy class, reflecting the higher prevalence of left-sided DDH [30].

Table 1: Method Comparison. Explainability and performance. Abbreviations: Accuracy - Acc., Precision - Prec., Recall - Rec., Landmark heatmaps - heatmaps.

Input	**Model**	**Acc. (%)**	**Prec. (%)**	**Rec. (%)**	**F1 (%)**	**Explainability**
Image Only	DPT	95.7	96.3	95.0	95.6	Heatmaps
	CNN	87.6	90.7	83.5	86.7	Grad-CAM
Image & Clinical	Early fusion	96.3	**98.7**	93.8	96.2	Heatmaps/attention
	Intermediate fusion	95.0	96.0	94.1	93.3	Grad-CAM
	Contribution-based	**96.4**	96.3	**96.3**	**96.3**	Direct contributions

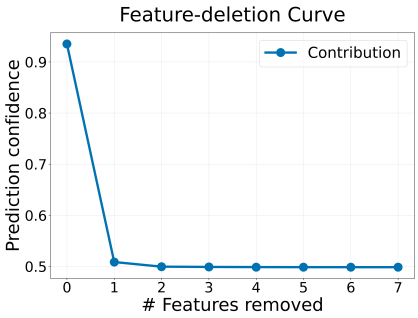

Fig. 4: Feature-deletion curves.

Table 2: Correlation between SHAP values and feature contributions.

Feature	Pearson r	Spearman ρ
Sex	0.920	0.848
Laterality	0.667	0.609
Age (days)	0.965	0.891
Alpha conf.	0.916	0.795
Alpha mean	0.956	0.782
FHC conf.	0.875	0.715
FHC mean	0.996	0.993

(i) Classification Performance. Performance was comparable across the fusion methods (Table 1). The proposed CBF achieved the highest accuracy (96.4%), recall (96.3%), and F1-score (96.3%), while early fusion achieved the highest precision (98.7%). These results demonstrate that explicit contribution estimates can be obtained without sacrificing predictive performance.

The main advantage of the proposed method lies in its explainability. Conventional fusion approaches provide indirect explanations through attention weights, landmark heatmaps, or Grad-CAM visualisations, but do not directly quantify how individual inputs contribute to a prediction. In contrast, the proposed method yields explicit per-patient feature-level contributions, quantifying the influence of individual image-derived measurements and clinical data points for each patient. By combining comparable performance with patient-specific explanations, the proposed method may promote clinician trust and support adoption.

(ii) Feature-deletion Curves. The deletion curve shows a sharp drop in performance after the highest-contributing feature is removed, followed by a smaller decrease after the second feature. This supports the reliability of the feature contribution analysis, and suggests that the model's predictions are often mainly driven by a single dominant feature. This is plausible for DDH screening, where

many cases are unambiguous, whereas borderline cases are likely to rely on a more balanced combination of features.

(iii) Shapley Correlations. Table 2 compares the proposed feature contributions with SHAP values. Strong agreement was observed across most features (Pearson r=0.667−0.996; Spearman ρ=0.609−0.993), indicating that the proposed contributions capture similar patterns of feature importance to SHAP, a widely used model-explainability method. This supports the validity of the proposed contribution estimates while providing patient-specific explanations of how individual features influence each prediction. Such transparent explanations may improve clinician understanding and trust in AI-assisted decision support.

5 Conclusion

This work presents a CBF multimodal framework that combines image-derived measurements with clinical data and, at the same time, it explicitly quantifies the contribution of individual patient-specific features to the final prediction. Using DDH screening as a representative clinical application, the proposed CBF method has been compared with established fusion strategies.

The proposed method achieves strong performance while providing explicit feature-level contributions. Unlike common fusion approaches that rely on attention weights, heatmaps, or complex latent representations, the framework directly quantifies how patient-specific measurements and clinical data influence individual patient predictions. Feature-deletion analysis demonstrates that features assigned larger contributions have a greater impact on model performance when removed, while strong agreement with SHAP values supported the validity of the contributions. By quantifying the influence of individual features and, in turn, modalities, the proposed framework offers a practical method for improving explainability. Whilst clinician trust was not evaluated directly in this study, the method also has the potential to support clinician trust in multimodal decision-support systems.

References

1. Acosta, J.N., Falcone, G.J., Rajpurkar, P., Topol, E.J.: Multimodal biomedical AI. Nature Medicine **28**(9), 1773–1784 (2022)
2. Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: A survey and taxonomy. IEEE TPAMI **41**(2), 423–443 (2018)
3. Borys, K., Schmitt, Y.A., Nauta, M., Seifert, C., Krämer, N., Friedrich, C.M., Nensa, F.: Explainable AI in medical imaging: An overview for clinical practitioners–beyond saliency-based XAI approaches. European journal of radiology **162**, 110786 (2023)
4. Chen, B., Xu, Q., Wang, L., Leung, S., Chung, J., Li, S.: An automated and accurate spine curve analysis system. IEEE Access **7**, 124596–124605 (2019)

5. Clement, A., Singh, A., Perry, D., Voiculescu, I.: Improving automated ultrasound infant hip screening using an integrated clinical classification loss. In: MICCAI. vol. 14860, p. 382–397. Springer (2024)
6. Clement, A., Singh, A., Voiculescu, I.: Landmark-based screening: Femoral head coverage and graf classification in infant developmental dysplasia of the hip. In: European Conference on Computer Vision. pp. 1–14. Springer (2024)
7. Clement, A., Willoughby, J., Voiculescu, I.: Confidence in Angle Predictions for Clinical Decision Support . In: MICCAI. vol. LNCS 15974, pp. 116 – 124. Springer Nature Switzerland (2025)
8. Cui, C., Yang, H., Wang, Y., Zhao, S., Asad, Z., Coburn, L.A., Wilson, K.T., Landman, B.A., Huo, Y.: Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: a review. Progress in Biomedical Engineering **5**(2), 022001 (2023)
9. Fang, C., Bai, S., Chen, Q., Zhou, Y., Xia, L., Qin, L., Gong, S., Xie, X., Zhou, C., Tu, D., et al.: Deep learning for predicting COVID-19 malignant progression. Medical image analysis **72**, 102096 (2021)
10. Fiterau, M., Bhooshan, S., Fries, J., Bournhonesque, C., Hicks, J., Halilaj, E., Ré, C., Delp, S.: Shortfuse: Biomedical time series representations in the presence of structured information. In: Machine Learning for Healthcare Conference. pp. 59–74. PMLR (2017)
11. Golovanevsky, M., Eickhoff, C., Singh, R.: Multimodal attention-based deep learning for Alzheimer's disease diagnosis. Journal of the American Medical Informatics Association **29**(12), 2014–2022 (2022)
12. Graf, R.: Fundamentals of sonographic diagnosis of infant hip dysplasia. Journal of Pediatric Orthopaedics **4**(6), 735–740 (1984)
13. Grant, D., Papież, B.W., Parsons, G., Tarassenko, L., Mahdi, A.: Deep learning classification of cardiomegaly using combined imaging and non-imaging ICU data. In: Medical Image Understanding and Analysis. pp. 547–558. Springer (2021)
14. Huang, S.C., Pareek, A., Seyyedi, S., Banerjee, I., Lungren, M.P.: Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. npj Digital Medicine **3**(1), 136 (2020)
15. Huang, S.C., Pareek, A., Zamanian, R., Banerjee, I., Lungren, M.P.: Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection. Scientific reports **10**(1), 22147 (2020)
16. Jin, W., Li, X., Hamarneh, G.: Evaluating explainable AI on a multi-modal medical imaging task: Can existing algorithms fulfill clinical requirements? In: The AAAI conference on artificial intelligence. vol. 36, pp. 11945–11953 (2022)
17. Jo, C., Hwang, D., Ko, S., Yang, M.H., Lee, M.C., Han, H.S., Ro, D.H.: Deep learning-based landmark recognition and angle measurement of full-leg plain radiographs can be adopted to assess lower extremity alignment. Knee Surgery, Sports Traumatology, Arthroscopy **31**(4), 1388–1397 (2023)
18. Krones, F., Marikkar, U., Parsons, G., Szmul, A., Mahdi, A.: Review of multimodal machine learning approaches in healthcare. Information Fusion **114**, 102690 (2025)
19. Leslie, A., Jones, A., Goddard, P.: The influence of clinical information on the reporting of CT by radiologists. The British journal of radiology **73**(874), 1052–1055 (2000)
20. Li, T.Z., Still, J.M., Xu, K., Lee, H.H., Cai, L.Y., Krishnan, A.R., Gao, R., Khan, M.S., Antic, S., Kammer, M., et al.: Longitudinal multimodal transformer integrating imaging and latent clinical signatures from routine EHRs for pulmonary nodule classification. In: MICCAI. pp. 649–659. Springer (2023)

21. Liang, P.P., Cheng, Y., Fan, X., Ling, C.K., Nie, S., Chen, R.J., Deng, Z., Allen, N., Auerbach, R., Mahmood, F., et al.: Quantifying & modeling multimodal interactions: An information decomposition framework. In: Conference on Neural Information Processing Systems (2023)
22. Liu, C., Ye, F.: A review of multimodal medical data fusion techniques for personalized medicine. In: International Conference on Biomedical and Intelligent Systems. pp. 338–347 (2025)
23. Loy, C.T., Irwig, L.: Accuracy of diagnostic tests read with and without clinical information: a systematic review. Jama **292**(13), 1602–1609 (2004)
24. Mohsen, F., Ali, H., El Hajj, N., Shah, Z.: AI-based methods for fusion of electronic health records and imaging data. Scientific Reports **12**(1), 17981 (2022)
25. Morin, C., Harcke, H., MacEwen, G.: The infant hip: real-time us assessment of acetabular development. Radiology **157**(3), 673–677 (1985)
26. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: The IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12179–12188 (2021)
27. Reyes, M., Meier, R., Pereira, S., Silva, C.A., Dahlweid, F.M., Tengg-Kobligk, H.v., Summers, R.M., Wiest, R.: On the interpretability of AI in radiology: challenges and opportunities. Radiology: artificial intelligence **2**(3) (2020)
28. Richards, A.: University of Oxford Advanced Research Computing (2015). `https://doi.org/10.5281/zenodo.22558`
29. Roposch, A., Liu, L.Q., Hefti, F., Clarke, N.M., Wedge, J.H.: Standardized diagnostic criteria for developmental dysplasia of the hip in early infancy. Clinical Orthopaedics and Related Research **469**, 3451–3461 (2011)
30. Roposch, A., Protopapa, E., Malaga-Shaw, O., Gelfer, Y., Humphries, P., Ridout, D., Wedge, J.H.: Predicting developmental dysplasia of the hip in at-risk newborns. BMC musculoskeletal disorders **21**, 1–8 (2020)
31. Sanders, J.O., Browne, R.H., McConnell, S.J., Margraf, S.A., Cooney, T.E., Finegold, D.N.: Maturity assessment and curve progression in girls with idiopathic scoliosis. JBJS **89**(1), 64–73 (2007)
32. Stahlschmidt, S.R., Ulfenborg, B., Synnergren, J.: Multimodal deep learning for biomedical data fusion: a review. Briefings in Bioinformatics **23**(2) (2022)
33. Tariq, A., Tang, S., Sakhi, H., Celi, L.A., Newsome, J.M., Rubin, D.L., Trivedi, H., Gichoya, J.W., Patel, B.N., Banerjee, I.: Graph-based fusion of imaging and non-imaging data for disease trajectory prediction. In: 2023 11th International IEEE/EMBS Conference on Neural Engineering (NER). IEEE (2023)
34. The GIMP Development Team: GIMP: GNU Image Manipulation Program. `https://www.gimp.org` (2024), computer software
35. Wang, A.Q., Karaman, B.K., Kim, H., Rosenthal, J., Saluja, R., Young, S.I., Sabuncu, M.R.: A framework for interpretability in machine learning for medical imaging. IEEE Access **12**, 53277–53292 (2024)
36. Wang, H., Subramanian, V., Syeda-Mahmood, T.: Modeling uncertainty in multimodal fusion for lung cancer survival analysis. In: IEEE International symposium on biomedical imaging (ISBI). pp. 1169–1172. IEEE (2021)
37. Yoo, Y., Tang, L.Y., Li, D.K., Metz, L., Kolind, S., Traboulsee, A.L., Tam, R.C.: Deep learning of brain lesion patterns, and user-defined clinical & MRI features for predicting conversion to multiple sclerosis from clinically isolated syndrome. Comp Meth in Biomech & Biomed Eng: Imaging & Visualization **7**(3), 250–259 (2019)
38. Zhang, Y., Jordan, J.M.: Epidemiology of osteoarthritis. Clinics in geriatric medicine **26**(3), 355–369 (2010)