Computer Methods and Programs in Biomedicine 288 (2027) 109672

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

A radiomics-based machine learning framework for full-reference ultrasound image quality assessment across varying acquisition settings

Alessandro Piccolo [a,*], Vito Papapicco [b], Andrea Mariani [b], Arianna Menciassi [a,c], Andrea Cafarelli [a,c]

[a] The BioRobotics Institute, Scuola Superiore Sant'Anna, Viale Rinaldo Piaggio 34, 56025 Pontedera, Pisa, Italy
[b] Soundsafe Care Srl, Viale Tunisia 10, 20124, Milan, Italy
[c] Department of Excellence in Robotics & AI, Scuola Superiore Sant'Anna, Piazza Martiri della Libertà 33, 56127, Pisa, Italy

ARTICLE INFO

Keywords:
Image quality assessment
Ultrasound imaging
Radiomics
Machine learning
Full-reference metrics
Acquisition settings

ABSTRACT

Background and objective: Ultrasound image quality depends on acquisition conditions, while conventional full-reference image quality assessment (FR-IQA) metrics may not fully capture perceptually relevant intensity and texture changes. Radiomics has been explored for image-quality characterization in other medical modalities. This study developed a radiomics-based machine-learning framework for FR-IQA under controlled acquisition-setting variations.

Methods: Rabbit cadaver ultrasound data included two anatomies and two probe-scanner configurations, defining four reference groups. Varying frequency, gain, power, and dynamic range yielded 192 reference-test pairs. Four sonographers rated pairs. B-mode and wavelet radiomic descriptors formed six full-reference representations evaluated by reference-grouped nested model development. Performance was compared with eight classical FR-IQA methods and two learned perceptual metrics, with robustness and acquisition-setting analyses.

Results: Average-measure inter-observer agreement was high, with ICC(A,k) = 0.9602. Out-of-reference predictions achieved SROCC = 0.8930, PLCC = 0.8702, RMSE = 0.1462, and R^2 = 0.7196. On the primary region of interest (ROI), the framework showed higher observed aggregate performance than all classical FR-IQA methods, while learned perceptual metrics showed higher observed performance. No reference-aware pairwise comparison was significant after Holm correction. Predictions were stable under modest ROI and discretization perturbations. Gain had the largest median radiomic effect and was the dominant factor for subjective quality and predictions across reference groups, enabling feature- and prediction-level assessment of acquisition-setting sensitivity.

Conclusions: Radiomic full-reference representations provide an interpretable approach for characterizing acquisition-dependent ultrasound image quality, with potential utility for system evaluation and protocol optimization. Larger multi-system, in vivo studies with repeated acquisitions and independently defined references are needed to establish generalizability.

1. Introduction

Ultrasound imaging is a widely used diagnostic modality owing to its real-time capability, portability, relatively low cost, and absence of ionizing radiation. However, ultrasound image appearance is highly dependent on the acquisition process and may vary substantially across operators, transducers, scanners, anatomical targets, and system settings [1–4]. Parameters such as transmit frequency, gain, acoustic output, and dynamic range affect different characteristics of the displayed B-mode image, while probe positioning, contact pressure, and acoustic coupling introduce additional sources of variability [3,5,6]. Consequently, maintaining adequate and consistent image quality remains important for reliable image interpretation and for the development and evaluation of ultrasound imaging technologies [2,7].

Image quality assessment (IQA) can be performed subjectively or objectively. Subjective assessment relies on human observers and

* Corresponding author.
E-mail addresses: alessandro.piccolo@santannapisa.it (A. Piccolo), vito.papapicco@soundsafecare.com (V. Papapicco), andrea.mariani@soundsafecare.com (A. Mariani), arianna.menciassi@santannapisa.it (A. Menciassi), andrea.cafarelli@santannapisa.it (A. Cafarelli).

https://doi.org/10.1016/j.cmpb.2026.109672
Received 19 March 2026; Received in revised form 31 August 2026; Accepted 29 September 2026
Available online 29 September 2026
0169-2607/© 2026 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

provides a direct measure of perceived image quality, but it is resource-intensive, affected by observer variability, and difficult to scale [8,9]. Objective IQA methods aim instead to provide reproducible computational estimates and are conventionally classified as full-reference (FR), reduced-reference, or no-reference approaches according to the information available from a reference image [8]. Full-reference IQA (FR-IQA) compares a test image with a predefined reference and is therefore particularly suited to controlled settings in which corresponding reference data are available, such as imaging-system evaluation, acquisition-protocol development, or parameter optimization. Conversely, no-reference approaches are more directly applicable when a matched reference cannot be obtained. More broadly, recent studies have emphasized that the suitability and behavior of IQA measures in medical imaging can depend strongly on the imaging modality and assessment task [10,11].

Conventional FR-IQA measures do not necessarily capture all perceptually relevant changes occurring in ultrasound B-mode images. Peak Signal-to-Noise Ratio (PSNR), for example, is based on pixel-wise error, whereas the Structural Similarity Index Measure (SSIM) compares local luminance, contrast, and structural information [12,13]. Although these properties can be informative, ultrasound image formation introduces additional complexity. Speckle is an intrinsic consequence of coherent echo formation and contributes to tissue texture, while acoustic attenuation and frequency-dependent resolution influence the spatial distribution and visibility of image information [4]. Furthermore, changes in gain alter signal amplification and overall image appearance, whereas dynamic-range adjustment modifies the mapping of echo amplitudes to displayed grayscale values [5,6]. Consequently, corresponding acquisitions of the same anatomy may exhibit substantial intensity and textural differences that are not fully characterized by direct pixel-wise or structural similarity. This does not imply that PSNR, SSIM, or related measures are intrinsically unsuitable for ultrasound; rather, their correspondence with perceived image quality needs to be established within the specific imaging context in which they are applied. Similar concerns have recently been raised more broadly in medical-image IQA, where metrics originally developed and validated mainly on natural images may behave differently across medical modalities and image-quality tasks [10,11].

The need for modality- and task-aware evaluation has motivated increasing research on perceptual IQA in medical imaging. Mason et al. compared several established objective image-quality measures with radiologists' assessments of diagnostic quality in magnetic resonance imaging, demonstrating that agreement with expert perception can vary markedly across metrics [14]. This type of study primarily examines the correspondence between established objective IQA measures and expert assessments rather than learning image-quality representations from paired reference and test acquisitions. More recently, perceptual IQA frameworks and dedicated datasets have been developed for low-dose computed tomography [15] and photoacoustic imaging [16], further emphasizing the importance of expert-rated medical-image benchmarks. In parallel, general IQA research has increasingly explored learned feature representations and perceptual models that move beyond direct pixel or handcrafted structural comparisons [17,18]. However, numerical performance values reported across different modalities, datasets, distortion processes, and subjective-rating tasks cannot be interpreted as directly comparable measures of methodological superiority.

Ultrasound-specific IQA has likewise evolved from conventional image statistics toward approaches incorporating perceptual, anatomical, and learned representations. Hemmsen et al. introduced a structured framework for clinical ultrasound image-quality evaluation based on relative subjective assessment by expert observers [19]. Other studies subsequently investigated automated ultrasound IQA using convolutional neural networks or structural-similarity-based learned representations [20,21]. More recently, degradation-aware deep learning has been proposed to separate quality-related degradation from semantic image content in ultrasound images [22]. Task-specific approaches have also incorporated radiomic and deep features to assess the quality of fetal ultrasound images for nuchal translucency measurement [23]. These approaches differ substantially in their definition of image quality, anatomical application, supervision strategy, and requirement for a reference image. Thus, assessing whether an ultrasound image meets the quality requirements of a specific clinical task and continuously quantifying its perceived quality relative to a corresponding reference image represent related but distinct IQA problems.

Radiomics provides a complementary representation of medical images through quantitative descriptors of intensity distributions and spatial texture. Radiomic analysis has been widely investigated in ultrasound for diagnostic and classification applications [24,25] and has also been explored for image-quality characterization in other imaging modalities, including PET and MRI [26–28]. Therefore, the use of radiomics for image-quality-related tasks is not itself unprecedented. Nevertheless, applying radiomics to ultrasound IQA raises specific methodological challenges because radiomic descriptors can be sensitive to image preprocessing, gray-level discretization, region-of-interest definition, and acquisition conditions [29,30]. Explicit specification of these factors and evaluation of feature and model robustness are therefore necessary when radiomic descriptors are used as quantitative indicators of image quality.

Against this background, the present study investigates a radiomics-based machine-learning strategy for full-reference assessment of ultrasound image quality under controlled variations in acquisition settings. Rather than relying on synthetically generated degradations, image-quality variability was produced through systematic changes in ultrasound acquisition parameters and evaluated against subjective expert assessments. Quantitative intensity and texture descriptors extracted from corresponding reference and test acquisitions were used to construct alternative full-reference representations, which were evaluated using a reference-grouped validation strategy designed to assess generalization beyond the reference acquisitions used for model development. The proposed approach was additionally examined through robustness and acquisition-setting sensitivity analyses and compared with both conventional FR-IQA measures and learned perceptual metrics under a consistent evaluation framework. The aim of this study is therefore to determine whether radiomic representations can provide a quantitative and interpretable characterization of perceived ultrasound image-quality variations while explicitly assessing their generalization, robustness, and limitations within the controlled experimental setting considered.

2. Materials and methods

2.1. Ultrasound image acquisition and dataset

The overall study workflow is summarized in Fig. 1.

Ultrasound images were acquired from a rabbit cadaver, targeting the left kidney and thigh musculature (Fig. 1a). Before imaging, the target areas were shaved and coupling gel was applied. Each anatomy was examined using two microconvex probe-scanner configurations: an Esaote mC 3–11 probe connected to a MyLab OmegaVET system (Esaote S.p.A., Genoa, Italy) and a Telemed MC10–4N12-A4 probe connected to an ArtUs EXT-1H scanner (UAB Telemed, Vilnius, Lithuania), yielding four reference groups defined by anatomy and probe-scanner configuration.

For each group, probe-specific acquisition settings were adjusted under the supervision of an experienced veterinary sonographer, and a reference image was selected by visual inspection based on perceived high image quality and minimal visible artifacts. No predefined quantitative acceptance criterion or independent multi-rater consensus procedure was used. The four reference images are shown in Fig. 2, with their acquisition settings reported in Supplementary Table S1.

The Esaote kidney and thigh images had matrix sizes of 841 × 841

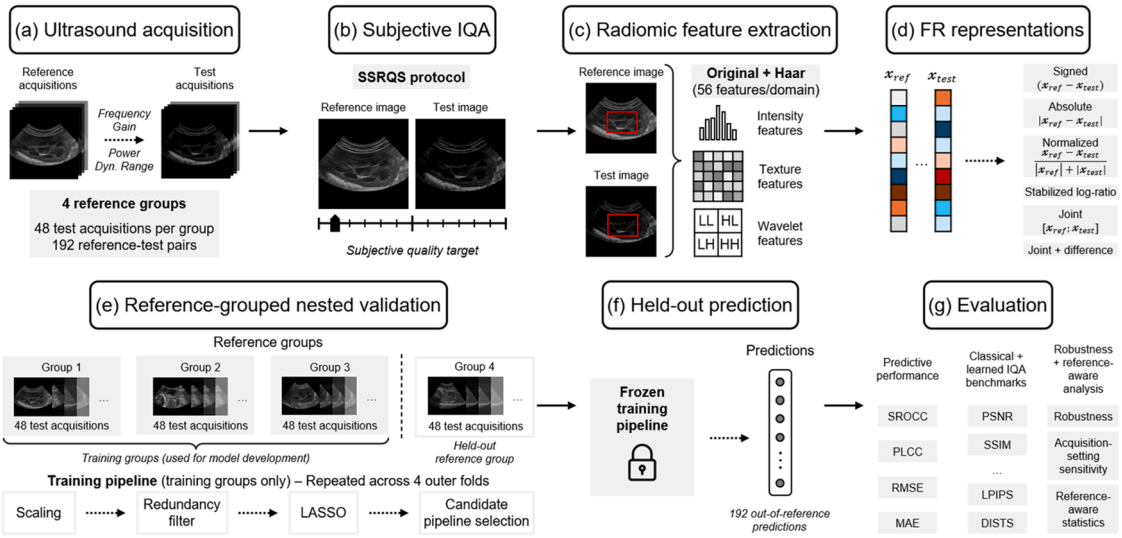


Fig. 1. Overview of the experimental and analytical workflow. Reference and test acquisitions obtained under controlled acquisition-setting variations (a) were evaluated subjectively (b), followed by radiomic feature extraction (c) and construction of six candidate full-reference radiomic representations (d). Reference-grouped nested validation was then used for training-only model development (e) and held-out prediction across all reference groups (f). The resulting out-of-reference predictions were used for performance evaluation, benchmarking, and robustness and reference-aware analyses (g).

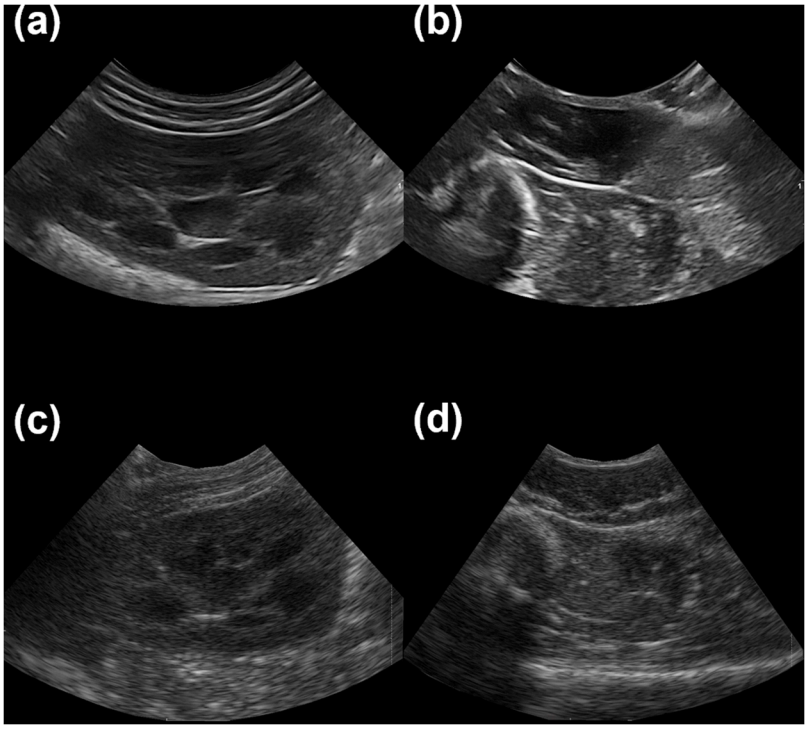


Fig. 2. Reference ultrasound images used in this study. (a) Left kidney acquired with the Esaote mC 3–11 microconvex probe. (b) Thigh musculature acquired with the Esaote mC 3–11 microconvex probe. (c) Left kidney acquired with the Telemed MC10–4N12-A4 microconvex probe. (d) Thigh musculature acquired with the Telemed MC10–4N12-A4 microconvex probe.

and 840 × 840 pixels, respectively, and were stored as 24-bit Bitmap (BMP), whereas the corresponding Telemed kidney and thigh images measured 931 × 931 and 930 × 930 pixels, respectively, and were stored as 24-bit Joint Photographic Experts Group (JPEG). These group-specific dimensions and formats were retained for the corresponding test acquisitions. Because each system used a different storage format, image format and probe-scanner configuration were not varied independently.

For each reference group, 48 test images were acquired by systematically varying frequency (3 levels), gain (2), power (2), and dynamic range (4), corresponding to the complete 3 × 2 × 2 × 4 factorial design and yielding 192 reference-test pairs. No artificial post-processing or synthetic degradation was used to generate test images. Within each anatomy and probe-scanner configuration, imaging depth, view area, focal depth, and all other acquisition parameters were kept fixed. Complete probe-specific parameter values are provided in Supplementary Table S2. Fig. 3 shows the Esaote kidney reference and five representative test acquisitions, with settings reported in Supplementary Table S3. An analogous example for the Telemed thigh group is shown in

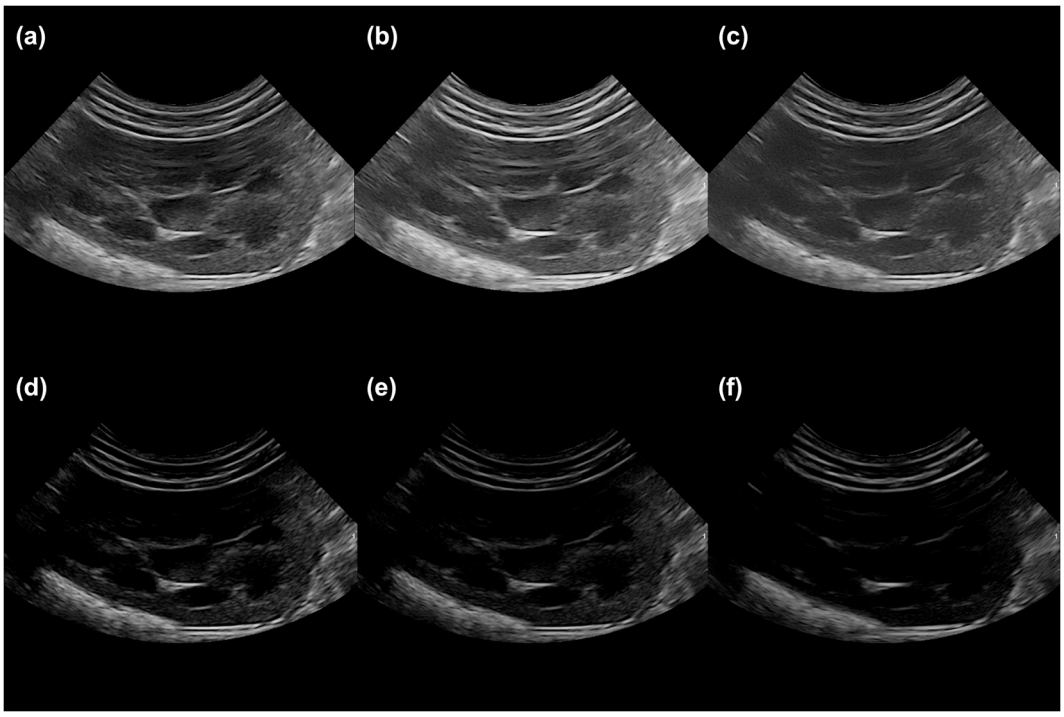


Fig. 3. Representative example of acquisition-setting variation. (a) Reference image of the left kidney acquired with the Esaote mC 3–11 microconvex probe. (b–f) Five representative test acquisitions obtained by varying frequency, gain, power, and dynamic range. The corresponding acquisition settings are reported in Supplementary Table S3.

Supplementary Figure S1, with settings reported in Supplementary Table S4.

Frequency, gain, power, and dynamic range were selected because they affect complementary aspects of the displayed B-mode image. Frequency influences the resolution-penetration trade-off, gain controls amplification of received echoes, power regulates transmitted acoustic output, and dynamic range affects the mapping of echo amplitudes to grayscale contrast [5,6]. Scanner-specific parameter scales were retained; in particular, Esaote frequency and dynamic-range settings use proprietary labels for which corresponding physical units are not disclosed by the manufacturer.

2.2. Subjective image quality assessment

Subjective image quality was evaluated using a customized protocol derived from the Simultaneous Double Stimulus for Continuous Evaluation and Double Stimulus Continuous Quality Scale methods described in Recommendation BT.500 of the International Telecommunication Union Radiocommunication Sector (ITU-R) [31]. This protocol, termed the Simultaneous Stimulus Relative Quality Scale (SSRQS) [19], presented each reference-test pair side by side (Fig. 1b). Each of the 192 pairs was shown twice with reversed left-right positioning and randomized throughout the session to reduce memorization and positional bias, yielding 384 comparisons per session (Supplementary Figure S2). Observers were blinded to the designated reference and judged the relative perceived quality of the two images rather than absolute diagnostic adequacy. Before testing, they completed 24 training comparisons using images not included in the testing phase. Ratings were provided on a continuous signed Visual Analog Scale (VAS; Supplementary Figure S3) ranging from −100 to +100, with 0 indicating equal perceived quality and the sign indicating which displayed image was judged better. Scale labels were shown only during the initial instructions to limit categorical response behavior.

Four sonographers performed the assessment independently using EchoQA, a custom application developed in MATLAB R2024b App Designer (The MathWorks, Inc., Natick, MA, USA), whose interface is shown in Fig. 4. A representative sequence is provided in Supplementary Video S1. Ambient illumination and display settings were kept constant, whereas viewing distance was unconstrained to allow each observer's preferred position, consistent with the application-dependent treatment of viewing distance in ITU-R BT.500 [31].

For analysis, the absolute value of each VAS response was used, yielding a 0–100 relative quality-difference score. For each observer and image pair, the scores from the two mirrored presentations were averaged, with lower values indicating a smaller perceived quality difference. The mean opinion score (MOS) was then calculated across the four observers and transformed to a higher-is-better subjective target as $Y = (100 - MOS)/100$, yielding one value in [0,1] for each of the 192 pairs.

To describe coverage of the perceptual quality range [9], subjective scores were expressed on the equivalent 0–100 scale and visualized using fixed 10-point histogram intervals. Compatibility with a uniform distribution was assessed separately using a Sturges-based binning scheme. Following Sturges' rule [32], the bin width was

$$w = \frac{\max(Y_{100}) - \min(Y_{100})}{1 + \log_2(N)}, \tag{1}$$

where $Y_{100} = 100Y$ and $N = 192$. For the chi-square goodness-of-fit test, the observed frequencies in the Sturges-based bins were compared with equal expected frequencies N/k, where k denotes the number of bins, with $df = k - 1$ and $\alpha = 0.05$.

Reliability was assessed before observer scores were averaged. Within-observer repeatability between mirrored presentations was quantified separately for each observer using the concordance correlation coefficient (CCC) [33]. Inter-observer absolute agreement was assessed using a two-way random-effects, absolute-agreement intraclass correlation coefficient, reporting ICC(A,1) and ICC(A,k) for k = 4 with 95% confidence intervals [34,35]. Observer influence on the aggregate target was further examined by recomputing the target after excluding each observer in turn and comparing each three-observer target with the complete four-observer target using CCC. Statistical analyses in this

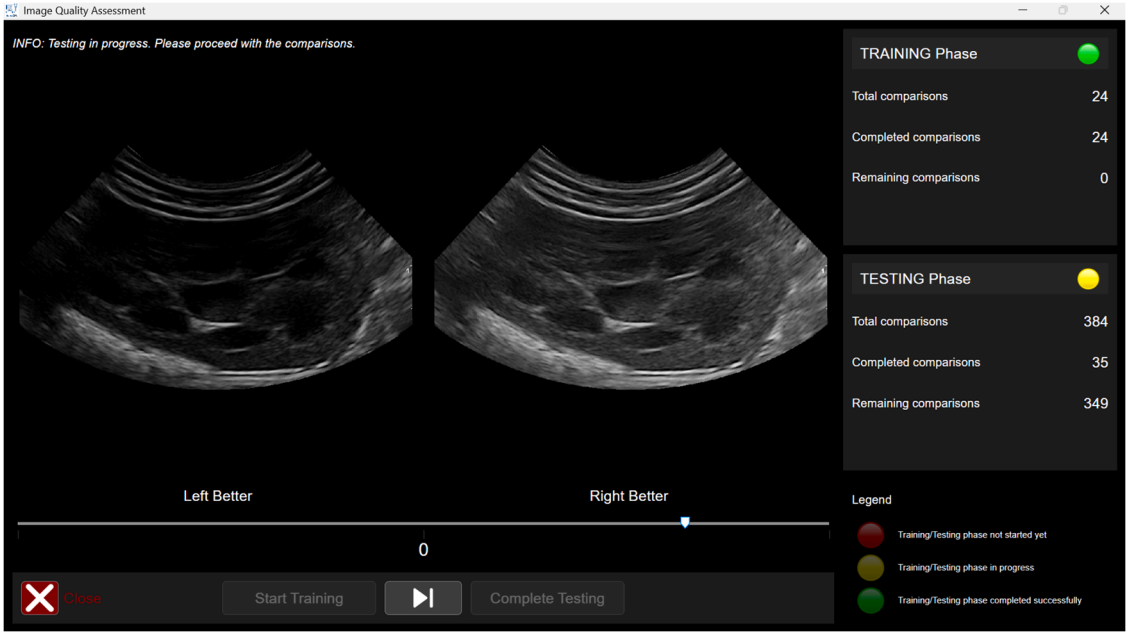


Fig. 4. Graphical user interface of EchoQA, the custom MATLAB application used for subjective ultrasound image-quality assessment with the Simultaneous Stimulus Relative Quality Scale protocol. Paired ultrasound images were displayed side by side and rated using a continuous Visual Analog Scale.

subsection were performed in MATLAB R2026a.

2.3. ROI definition and radiomic feature extraction

Radiomic analysis was performed within one reference-specific rectangular region of interest (ROI) for each of the four reference groups. ROIs were manually defined on the reference images, centrally within the acoustic window, under the supervision of an experienced veterinary sonographer, to provide a consistent tissue region for paired radiomic analysis. The same coordinates were applied unchanged to all 48 corresponding test acquisitions. ROI matrix sizes were 210 × 329 pixels for the Esaote kidney group, 247 × 362 for the Esaote thigh group, 271 × 351 for the Telemed kidney group, and 250 × 346 for the Telemed thigh group, reported as rows × columns. The radiomics pipeline was therefore ROI-based, whereas subjective assessment was performed on the complete displayed images. Accordingly, benchmark methods were evaluated both within the same reference-specific ROIs and on the complete displayed images, as described in Section 2.6.

Three-channel images were converted to grayscale before ROI extraction. Original B-mode ROIs were analyzed on their native grayscale intensity scale without geometric resampling, interpolation, or intensity resegmentation. Each ROI was also converted from its native 8-bit grayscale scale to double-precision values in [0,1] and decomposed using a single-level two-dimensional Haar discrete wavelet transform into low-low (LL), low-high (LH), high-low (HL), and high-high (HH) subbands [36,37]. The fixed 8-bit scale, rather than per-image min-max normalization, was used for this conversion, and signed wavelet coefficients were retained without further subband-wise normalization or remapping.

Radiomic features were extracted in MATLAB R2026a using the Medical Imaging Toolbox, with definitions aligned with Image Biomarker Standardisation Initiative (IBSI) terminology [30] (Fig. 1c). Discretization-dependent features were computed independently within each image domain using a Fixed Bin Number (FBN) scheme with 32 bins, without additional intensity normalization or resegmentation. The base set comprised 56 descriptors: 17 intensity features (15 intensity-based statistical and 2 intensity histogram features), 23 Gray Level Co-occurrence Matrix (GLCM) features, and 16 Gray Level Run Length Matrix (GLRLM) features [38–42]. GLCM and GLRLM features were computed over 0°, 45°, 90°, and 135° directions, at a Chebyshev distance of one pixel for GLCM, and averaged across directions. The same 56 features were extracted independently from the original B-mode ROI and each of the four wavelet subbands, yielding 280 descriptors per image. Reference and test images were processed identically. Complete feature lists and mathematical definitions are provided in Supplementary Tables S5–S7.

2.4. Full-reference radiomic representations

Let x_{ref} and x_{test} denote the 280-dimensional radiomic feature vectors of the reference and test images. Six full-reference representations were constructed for each pair (Fig. 1d). Signed and absolute differences were defined as $r_{\Delta} = x_{ref} - x_{test}$ and $r_{|\Delta|} = |x_{ref} - x_{test}|$. The normalized difference was computed element-wise as $r_{N,j} = (x_{ref,j} - x_{test,j})/(|x_{ref,j}| + |x_{test,j}|)$, with $r_{N,j} = 0$ when $|x_{ref,j}| + |x_{test,j}| = 0$. To accommodate zero or negative radiomic values, a stabilized log-ratio was also constructed. For each feature j, reference and test values from the relevant training partition were pooled into z_j. A positive margin $\varepsilon_j = 10^{-6}\, max\{IQR(z_j), median(|z_j|), 1\}$ and shift $s_j = \max(0, -m_j) + \varepsilon_j$ were defined, where $m_j = \min(z_j)$. Stabilized values were $a_{ref,j} = \max(x_{ref,j} + s_j, \varepsilon_j)$ and $a_{test,j} = \max(x_{test,j} + s_j, \varepsilon_j)$, yielding $r_{L,j} = \ln(a_{ref,j}) - \ln(a_{test,j})$. Margins and shifts were estimated exclusively from the corresponding training partition and applied unchanged to validation or held-out data.

Two joint representations were defined as $r_J = [x_{ref}; x_{test}]$ and $r_{J\Delta} = [x_{ref}; x_{test}; x_{ref} - x_{test}]$, where semicolons denote feature-vector concatenation. Signed difference, absolute difference, normalized difference, and stabilized log-ratio each contained 280 features; the joint and joint-plus-difference representations contained 560 and 840 features, respectively. All six were candidate inputs to the predictive pipeline, with representation selection performed exclusively within the training reference groups as described in Section 2.5.

2.5. Reference-grouped nested model development

Model development and validation used a reference-grouped nested scheme (Fig. 1e). The four reference groups defined in Section 2.1 served as the grouping units for cross-validation. In each of four outer folds, one

complete group (48 reference-test pairs) was held out, while the remaining three groups (144 pairs) were used for model development. The held-out group was excluded from representation fitting, preprocessing, feature selection, model selection, and hyperparameter optimization.

Within each outer-training set, three grouped inner folds were used for model selection. One training reference group served as validation in each fold, yielding 96 training and 48 validation pairs. Representation-specific parameters and preprocessing were estimated only from the corresponding inner-training data and applied unchanged to inner-validation data. Zero-variance features were removed, and the remaining features were standardized using the training-set mean and standard deviation. Feature redundancy was controlled with a deterministic Spearman rank-correlation filter: features were processed in fixed input order, and a later feature was removed when its absolute correlation with an already retained feature exceeded 0.90.

Least Absolute Shrinkage and Selection Operator (LASSO) regression was then used for feature selection [43]. Eight prespecified regularization parameters were evaluated, retaining only features with non-zero coefficients. Five regression families were considered: Multiple Linear Regression, Ridge Regression, Decision Tree Regression, Gradient Boosting Regression, and radial-basis-function Support Vector Regression (RBF-SVR). The LASSO values and model-specific hyperparameter grids are reported in Supplementary Table S8. Across six full-reference representations, eight LASSO values, and 44 regression-model configurations, 2112 candidate pipelines were enumerated per outer fold. Pipelines for which LASSO retained no features in at least one inner-training fold were excluded from ranking.

Predictions from the three inner-validation folds were pooled across the 144 outer-training pairs for each candidate. Selection was based primarily on the lowest pooled root mean square error (RMSE), with lower mean absolute error (MAE), higher coefficient of determination (R^2), and a fixed deterministic candidate order used sequentially as tie-breakers. The selected pipeline, including representation, preprocessing, selected features, regression family, and hyperparameters, was then refitted on all 144 outer-training pairs.

The refitted pipeline was applied once to the 48 pairs in the held-out group. Repeating this procedure for all four groups yielded one out-of-reference prediction for each of the 192 pairs (Fig. 1f). No post-hoc clipping or calibration based on held-out targets was applied. Model development was performed in Python 3.13.9 (64-bit) using scikit-learn 1.9.0 [44], with a fixed master random seed for stochastic model components.

2.6. Benchmark full-reference image quality assessment methods

The proposed radiomics-based framework was benchmarked against eight classical and two learned/perceptual FR-IQA methods (Fig. 1g): Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [12], Multi-Scale SSIM (MS-SSIM) [45], Information Content Weighted SSIM (IW-SSIM) [46], Feature Similarity Index Measure (FSIM) [47], Visual Signal-to-Noise Ratio (VSNR) [48], Information Fidelity Criterion (IFC) [49], Visual Information Fidelity (VIF) [50], Learned Perceptual Image Patch Similarity (LPIPS) with an AlexNet backbone and version 0.1 wt [51], and Deep Image Structure and Texture Similarity (DISTS) [52]. All benchmarks were evaluated on the 192 reference-test pairs using two spatial supports: the reference-specific ROIs used by the radiomics framework as the primary analysis and the complete displayed grayscale image frame as a secondary analysis. Reference and test images were evaluated on matched spatial supports, without resizing, interpolation, padding, registration, or additional alignment.

Classical FR-IQA methods were evaluated in MATLAB R2026a. PSNR used the corresponding built-in MATLAB implementation, whereas the remaining classical metrics were computed using the original source code provided by their respective authors. For MS-SSIM, IFC, and VIF, native 8-bit grayscale intensities were converted to double precision on a fixed [0,1] scale; PSNR, SSIM, IW-SSIM, FSIM, and VSNR used the native grayscale scale. The implemented VSNR score was expressed as $100 - VSNR$ so that higher values indicated greater similarity.

LPIPS and DISTS were evaluated in Python 3.13.9 (64-bit) using the original implementations released by their respective authors. Grayscale images were converted to 32-bit floating-point values in [0,1] and replicated across three channels. LPIPS inputs were then mapped to [−1,1], whereas DISTS inputs remained in [0,1]. Both pretrained models were used in evaluation mode without study-specific training or fine-tuning. Inference was performed with PyTorch 2.9.1 on the CPU, using batch size 1 and disabled gradient computation. Because native LPIPS and DISTS outputs are distances, their scores were multiplied by −1 for rank-based evaluation so that higher oriented scores indicated better image quality.

For performance measures requiring predictions on the subjective target scale, each benchmark was calibrated using the same four outer reference-group folds defined in Section 2.5. Within each fold, raw scores from the 144 training pairs were standardized using the corresponding training mean and standard deviation, and a five-parameter logistic mapping [13] was fitted exclusively to the standardized training scores and subjective targets:

$$f(z) = \beta_1 \left(0.5 - \frac{1}{1 + \exp[\beta_2 (z - \beta_3)]} \right) + \beta_4 z + \beta_5, \tag{2}$$

where z is the standardized metric score and $\beta_1, \ldots, \beta_5$ are calibration parameters. The mapping was fitted using a deterministic profiled least-squares procedure with multiple initializations to improve numerical robustness and was applied unchanged to the 48 held-out pairs. Calibration was performed independently for each benchmark and spatial support. Held-out targets did not influence standardization or calibration, and calibrated predictions were not clipped or otherwise adjusted. The Spearman rank-order correlation coefficient (SROCC) was computed from consistently oriented raw benchmark scores, whereas the Pearson linear correlation coefficient (PLCC), MAE, RMSE, and R^2 were computed from the corresponding out-of-reference calibrated predictions. Performance-measure definitions and reference-aware statistical comparisons are described in Section 2.9.

2.7. Robustness and sensitivity analyses

Robustness of the radiomics framework was evaluated under controlled perturbations of ROI geometry, gray-level discretization, and image noise. For ROI sensitivity, each reference-specific ROI and its corresponding test ROI were jointly translated by 5 pixels in eight directions or isotropically scaled about the ROI center to 95% or 105% of its original size. Radiomic features were then re-extracted using the procedure described in Section 2.3. Discretization sensitivity was assessed by repeating feature extraction with FBNs of 16 and 64, using FBN = 32 as the baseline and leaving all other extraction settings unchanged. Noise sensitivity was evaluated by adding zero-mean Gaussian noise with standard deviations of 0.01, 0.02, and 0.05 relative to the full grayscale range. Three independent realizations were generated at each noise level, with independent perturbations applied to reference and test images.

Feature- and representation-level stability relative to the unperturbed baseline was quantified using CCC and SROCC for the test-image radiomic features and for all six full-reference radiomic representations defined in Section 2.4. For each outer fold, any training-dependent representation parameters were estimated from the unperturbed outer-training data and kept fixed across perturbation conditions.

Prediction-level robustness was assessed by applying the four outer-fold pipelines obtained in Section 2.5 unchanged to each perturbed feature set, without retraining, preprocessing refitting, representation refitting, or feature reselection. Stability relative to the corresponding

unperturbed out-of-reference predictions was quantified using CCC and SROCC, while prediction drift was quantified using the RMSE and MAE of the prediction differences. For noise perturbations, these metrics were summarized across the three realizations. These analyses quantified sensitivity to controlled post-acquisition perturbations.

2.8. Model and dataset diagnostic analyses

Additional diagnostic analyses examined feature-selection consistency, sensitivity to the number of training reference groups, dependence on acquisition settings, and reference-test content stability.

Feature-selection consistency across the four outer folds was assessed after mapping representation-specific selected features to their underlying radiomic feature identities. Recurrence across folds and pairwise Jaccard indices between the corresponding selected-feature sets were then computed.

Sensitivity to the amount of grouped training data was examined using a fixed-configuration diagnostic with one, two, or three training reference groups, corresponding to 48, 96, and 144 training pairs, respectively. For the one- and two-reference conditions, grouped training-validation combinations were constructed within the corresponding outer-training reference sets; the three-reference condition corresponded to the outer held-reference evaluation described in Section 2.5. The fold-specific full-reference representation, selected feature set, regression family, and model hyperparameters were kept fixed, with no additional model selection or hyperparameter optimization. RMSE was summarized across the valid grouped evaluations. This analysis was descriptive and was not used for model selection or primary performance estimation.

Dependence on acquisition settings was evaluated separately within each reference group using the balanced $3 \times 2 \times 2 \times 4$ factorial design. Frequency, gain, power, and dynamic range were treated as categorical factors, and their effects on the subjective target, frozen out-of-reference predictions, and selected radiomic features were quantified using eta-squared (η^2), defined as the factor sum of squares divided by the total sum of squares. For the subjective target and model predictions, the four-factor effect profiles were compared by identifying the dominant factor and the maximum absolute difference between corresponding η^2 values. In a further sensitivity analysis, features retained by each outer-fold pipeline and showing strong association with gain or dynamic range ($\eta^2 \geq 0.80$) in the corresponding outer-training reference groups were removed. The same regression family and hyperparameters were then retained without further tuning, while the held-out reference group remained excluded from the feature-removal decision.

Reference-test content stability was assessed within the fixed radiomics ROI using translation-only alignment based on gradient correlation. For each of the 192 reference-test pairs, the estimated translation magnitude was expressed in pixels and normalized by the diagonal length of the corresponding ROI. Normalized mutual information (NMI) was computed before and after alignment as $NMI = [H(A) + H(B)]/H(A, B)$ [53], where $H(A)$ and $H(B)$ denote the marginal entropies of the reference and test ROIs and $H(A, B)$ their joint entropy. Entropies were estimated using a fixed 64×64 joint histogram on grayscale values scaled to [0,1]. Valid ROI overlap after translation and gradient correlation before and after alignment were also recorded as descriptive checks. The alignment analysis was used only to quantify residual spatial displacement and content correspondence. Registered images were not used for radiomic feature extraction, model fitting, or performance evaluation.

2.9. Performance evaluation and statistical analysis

Predictive performance was evaluated against the subjective target using the pooled out-of-reference predictions from the four outer folds and separately within each held-out reference group. Let y_i and $\hat{y}_i$ denote the subjective target and corresponding prediction for pair i. Performance was quantified using PLCC, SROCC, MAE, RMSE, and R^2:

$$PLCC = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2 \sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}}, \tag{3}$$

$$SROCC = \frac{\sum_{i=1}^{N}(R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{N}(R_i - \bar{R})^2 \sum_{i=1}^{N}(S_i - \bar{S})^2}}, \tag{4}$$

where R_i and S_i are the ranks of the subjective targets and evaluated scores, respectively. Error-based measures were defined as

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \tag{5}$$

and

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}. \tag{6}$$

For the proposed framework, SROCC was computed from the out-of-reference predictions. For benchmark methods, SROCC was computed from the consistently oriented raw scores, whereas PLCC, MAE, RMSE, and R^2 were computed from the corresponding outer-training-calibrated predictions described in Section 2.6.

Uncertainty was quantified while preserving dependence among pairs sharing the same reference image. Overall 95% confidence intervals (CIs) were obtained by resampling the four reference groups as intact blocks, with all $4^4 = 256$ possible ordered bootstrap draws enumerated. Reference-group-, anatomy-, and probe-scanner-specific summaries were treated as descriptive. For these subset analyses, 95% confidence intervals for PLCC, SROCC, and RMSE were obtained from 10,000 conditional bootstrap replicates within the corresponding subset.

Pairwise comparisons between the proposed framework and each benchmark were performed separately for SROCC and RMSE on the ROI and full-image supports. Metric differences were computed within each of the four reference groups and evaluated using exact two-sided sign-flip tests over all $2^4 = 16$ possible sign patterns. Block-bootstrap 95% confidence intervals were also obtained for the paired effects. Multiplicity was controlled using the Holm procedure [54] within four prespecified families: ROI SROCC, ROI RMSE, full-image SROCC, and full-image RMSE, each comprising ten benchmark comparisons. Statistical significance was assessed at $\alpha = 0.05$, and non-rejection was not interpreted as evidence of equivalence.

2.10. Computational runtime analysis

Computational runtime on the primary ROI support was evaluated on 48 balanced reference-test pairs, with 12 pairs from each reference group. Each method was evaluated over five measured repetitions after one unmeasured warm-up, yielding 240 timing measurements per condition. Images were preloaded and disk input/output operations were excluded. Measurements were performed on a workstation with an Intel Core i5–1135G7 processor at 2.40 GHz, 8 GB RAM, and Windows 11 Home 64-bit. MATLAB R2026a was used for radiomic and classical-metric computations, whereas Python 3.13.9 with scikit-learn 1.9.0 and PyTorch 2.9.1 was used for predictive and learned-model inference.

For the proposed framework, runtime was evaluated by extracting only the fold-specific radiomic features retained by the frozen pipelines described in Section 2.5, followed by predictor inference. Feature extraction was measured both with and without precomputed reference features. Because radiomic extraction and prediction were implemented in separate software environments, their execution times were measured independently and summed component-wise. Extraction of the complete 280-feature set was additionally timed as a non-optimized reference

implementation. Classical benchmark timings comprised metric computation on preloaded grayscale images, whereas LPIPS and DISTS timings included input conversion and model inference. All outputs produced during runtime measurements were numerically verified against the corresponding frozen scores or predictions. Runtime distributions were summarized using the mean and standard deviation and the median with first (Q1) and third (Q3) quartiles.

3. Results

3.1. Subjective image quality scores and observer reliability

Across the 192 reference-test pairs, the subjective quality target ranged from 0.0194 to 0.9780, with a mean of 0.5473, a median of 0.5706, and a standard deviation of 0.2768. Its distribution on the equivalent 0–100 scale spanned nearly the full perceptual quality range considered in the study (Fig. 5). The separate Sturges-based uniformity analysis yielded nine bins, and the chi-square goodness-of-fit test gave $\chi^2(8) = 12.38$ ($p = 0.135$); therefore, the null hypothesis of a uniform distribution was not rejected.

Within-observer repeatability between mirrored presentations yielded CCC values of 0.9535, 0.9472, 0.6045, and 0.9183 for Observers 1–4, respectively. Inter-observer absolute agreement was ICC(A,1) = 0.8577 (95% CI, 0.7766–0.9053) for individual observer ratings and ICC(A,k) = 0.9602 (95% CI, 0.9322–0.9747) for the average of the four observers, showing higher agreement for the aggregated ratings than for individual ratings. The leave-one-observer-out analysis showed consistently high concordance between each three-observer target and the complete four-observer target, with CCC values of 0.9978, 0.9942, 0.9851, and 0.9967 after excluding Observers 1–4, respectively.

3.2. Out-of-reference performance of the proposed framework

The reference-grouped nested procedure generated one out-of-reference prediction for each of the 192 reference-test pairs and selected a distinct pipeline in each outer fold (Supplementary Table S9). For the folds holding out the Esaote kidney and Esaote thigh groups, the training-only selection procedure chose absolute-difference representations with Ridge regression, retaining 11 and 16 features, respectively. The folds holding out the Telemed kidney and Telemed thigh groups selected a normalized-difference representation with RBF-SVR and 3 retained features and a stabilized log-ratio representation with RBF-SVR and 4 retained features, respectively.

Across the pooled outer held-out predictions, the proposed framework achieved PLCC = 0.8702 (95% CI, 0.8375–0.9544), SROCC = 0.8930 (95% CI, 0.8681–0.9459), MAE = 0.1199 (95% CI, 0.0942–0.1528), RMSE = 0.1462 (95% CI, 0.1109–0.1861), and R^2 = 0.7196 (95% CI, 0.5363–0.8509) (Table 1; Fig. 6). Performance varied across the four held-out reference groups (Table 1). For three of the four

Table 1
Out-of-reference predictive performance of the proposed radiomics-based framework. Each held-out reference-group row comprises 48 reference-test pairs evaluated after model development on the remaining 144 pairs. The pooled row was computed from the 192 outer held-out predictions obtained across the four reference groups.

Held-out reference group	MAE	RMSE	R^2	PLCC	SROCC
Kidney – Esaote	0.1201	0.1374	0.6845	0.9740	0.9398
Kidney – Telemed	0.0979	0.1120	0.8755	0.9680	0.9684
Thigh – Esaote	0.1711	0.2049	0.4394	0.9540	0.9480
Thigh – Telemed	0.0904	0.1098	0.7964	0.9155	0.8950
Pooled out-of-reference	0.1199	0.1462	0.7196	0.8702	0.8930

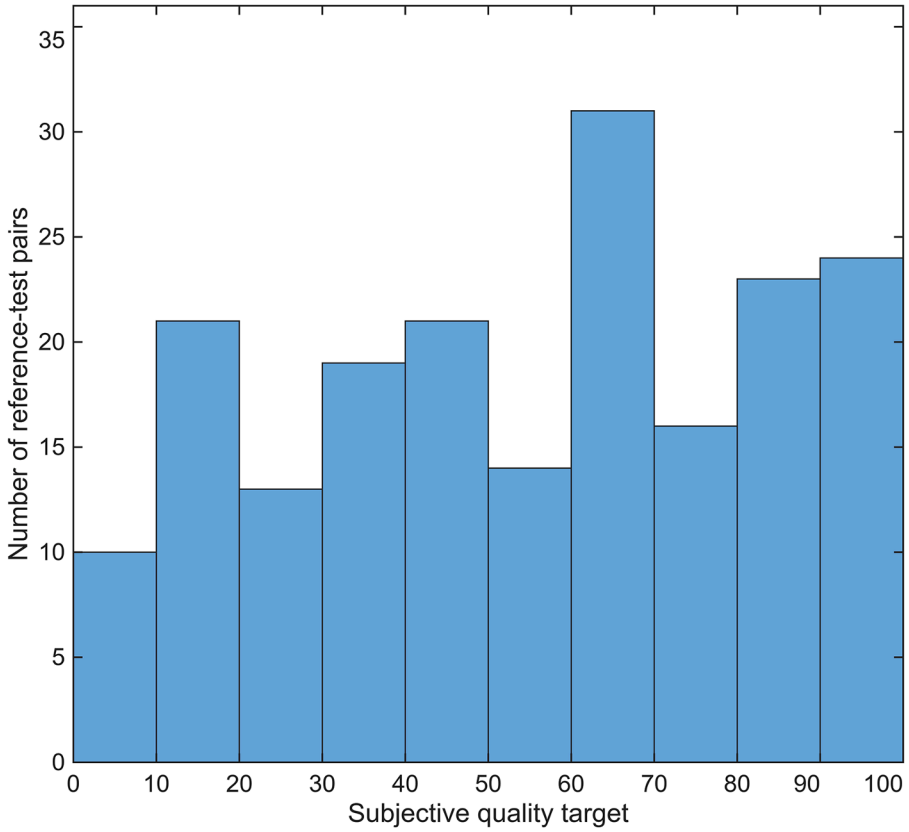


Fig. 5. Distribution of the subjective quality target across the 192 reference-test pairs. Scores are shown on the equivalent 0–100 scale using fixed 10-point intervals, with higher values indicating smaller perceived quality differences between the paired images.

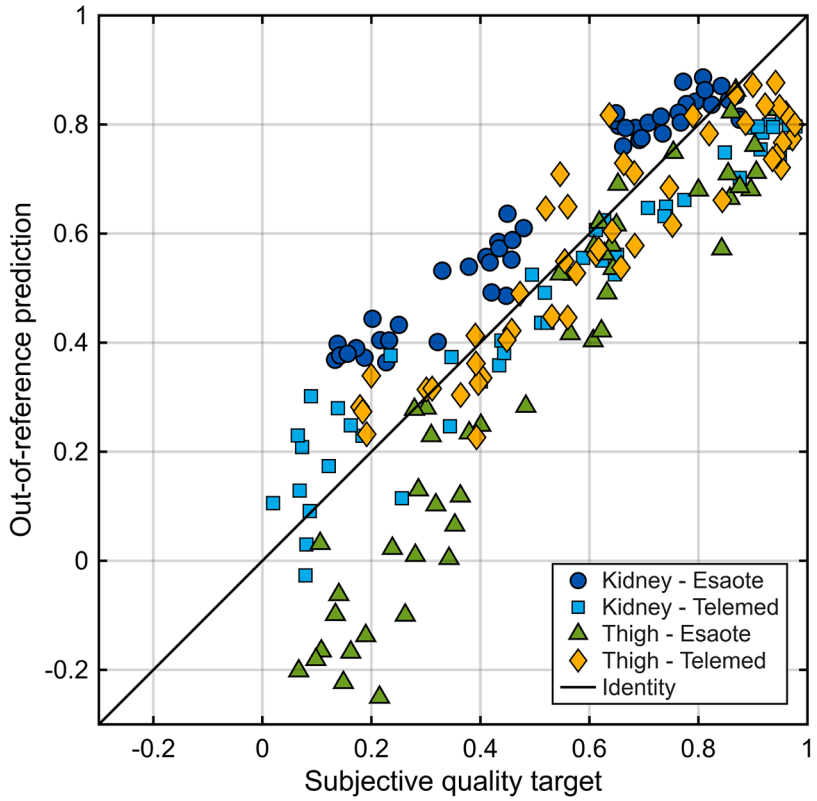


Fig. 6. Out-of-reference predictions of the proposed radiomics-based framework versus the subjective quality target across the 192 reference-test pairs. Each marker represents a prediction obtained while the corresponding reference group was completely excluded from model development. Marker colors and shapes distinguish the four reference groups. The diagonal line represents identity.

groups, R^2 ranged from 0.6845 to 0.8755 and RMSE from 0.1098 to 0.1374. For the Esaote thigh group, PLCC and SROCC remained high at 0.9540 and 0.9480, respectively, whereas R^2 was 0.4394 and RMSE was 0.2049.

Confidence intervals for reference-group-, anatomy-, and probe-scanner-specific PLCC, SROCC, and RMSE are reported in Supplementary Table S10. For kidney images, PLCC was 0.9085 (95% CI, 0.8867–0.9276) and SROCC was 0.8865 (95% CI, 0.8255–0.9157), whereas thigh images yielded PLCC = 0.9160 (95% CI, 0.8914–0.9378) and SROCC = 0.9297 (95% CI, 0.8947–0.9473). For images acquired with the Esaote system, PLCC was 0.8419 (95% CI, 0.7995–0.8795) and SROCC was 0.8766 (95% CI, 0.8264–0.9039). For Telemed images, the corresponding values were PLCC = 0.9481 (95% CI, 0.9282–0.9648) and SROCC = 0.9382 (95% CI, 0.8964–0.9592).

3.3. Comparison with full-reference IQA benchmarks

On the primary ROI support, the proposed framework showed higher observed aggregate performance than all eight classical FR-IQA methods across the five evaluated performance measures (Table 2). Among the classical methods, PSNR provided the highest SROCC (0.8233) and PLCC (0.7141) and the lowest RMSE (0.1989), compared with SROCC = 0.8930, PLCC = 0.8702, and RMSE = 0.1462 for the proposed framework. The learned/perceptual methods yielded higher observed aggregate performance: LPIPS achieved SROCC = 0.9226 and RMSE = 0.1086, whereas DISTS showed the highest observed performance on the ROI support, with SROCC = 0.9530, PLCC = 0.9519, MAE = 0.0678, RMSE = 0.0847, and R^2 = 0.9060.

On the complete displayed-image support, several classical methods showed higher performance than on the ROI support (Supplementary

Table 2
Performance of the proposed radiomics-based framework and full-reference image quality assessment benchmarks on the primary ROI support. SROCC was computed from the proposed framework predictions and from consistently oriented raw benchmark scores. For benchmark methods, PLCC, MAE, RMSE, and R^2 were computed from outer-training-calibrated predictions; no held-out target was used for calibration. Values in parentheses indicate reference-block bootstrap 95% confidence intervals for SROCC and RMSE.

Method	SROCC (95% CI)	PLCC	MAE	RMSE (95% CI)	R^2
Proposed framework	0.8930 (0.8681–0.9459)	0.8702	0.1199	0.1462 (0.1109–0.1861)	0.7196
PSNR	0.8233 (0.7930–0.9148)	0.7141	0.1639	0.1989 (0.1532–0.2359)	0.4808
SSIM	0.7696 (0.7026–0.8797)	0.6193	0.1808	0.2310 (0.1496–0.3132)	0.2999
MS-SSIM	0.3630 (0.0553–0.7526)	0.2957	0.2536	0.3351 (0.2105–0.4851)	−0.4734
IW-SSIM	0.2884 (−0.0873–0.6919)	0.1869	0.2764	0.3633 (0.2227–0.5296)	−0.7320
FSIM	0.6907 (0.6318–0.8002)	0.6467	0.1708	0.2144 (0.1928–0.2424)	0.3965
VSNR	0.6800 (0.5394–0.8836)	0.5418	0.1921	0.2418 (0.1502–0.3196)	0.2325
IFC	0.1836 (0.1372–0.4297)	−0.0075	0.2900	0.3488 (0.2537–0.4485)	−0.5961
VIF	0.4485 (0.2877–0.7695)	0.1307	0.2687	0.3196 (0.2608–0.3692)	−0.3406
LPIPS	0.9226 (0.8783–0.9513)	0.9197	0.0857	0.1086 (0.0965–0.1221)	0.8452
DISTS	0.9530 (0.9278–0.9717)	0.9519	0.0678	0.0847 (0.0724–0.1006)	0.9060

Table S11). PSNR achieved the highest SROCC among the classical full-image methods (0.8814), whereas SSIM yielded the lowest classical RMSE (0.1606); both remained below the corresponding aggregate performance of the proposed framework. LPIPS and DISTS again showed higher observed performance, with SROCC values of 0.9330 and 0.9532 and RMSE values of 0.1101 and 0.0937, respectively. At the individual reference-group level, selected classical methods nevertheless matched or exceeded the proposed framework for specific endpoints. On the ROI support, PSNR matched the proposed framework for the Telemed kidney group (SROCC = 0.9684 for both) and exceeded it for the Telemed thigh group (0.9533 vs. 0.8950), while SSIM yielded a lower RMSE for the Esaote thigh group (0.1332 vs. 0.2049).

Reference-aware paired comparisons are reported in Supplementary Table S12. On the ROI support, the 95% confidence intervals for the SROCC difference were entirely above zero, indicating higher observed SROCC for the proposed framework, for six of the eight classical methods, while the intervals for PSNR and VSNR included zero. For RMSE, the corresponding interval included zero only for SSIM. On the full-image support, the SROCC difference interval included zero for PSNR, whereas the RMSE difference intervals included zero for SSIM and FSIM. For DISTS, the paired-effect confidence intervals were entirely below zero for both SROCC and RMSE on both spatial supports, corresponding to higher observed performance for DISTS. For LPIPS, the intervals included zero for both ROI endpoints and for full-image SROCC, whereas the full-image RMSE interval was entirely below zero, corresponding to lower observed RMSE for LPIPS. None of the 40 reference-aware pairwise comparisons reached statistical significance after Holm correction (Supplementary Table S12). Given the four reference groups, the minimum attainable two-sided exact *p*-value was 0.125.

3.4. Robustness to ROI, discretization, and noise perturbations

Small perturbations of ROI geometry produced limited changes in the extracted radiomic features. Across the eight 5-pixel translations and two scaling conditions, median feature-level CCC ranged from 0.9711 to 0.9953 and SROCC from 0.9628 to 0.9951 (Supplementary Table S13, panel A). All six full-reference representations also showed high stability, with median condition-level CCC values ranging from 0.9394 to 0.9775 and SROCC values from 0.9090 to 0.9814 (Supplementary Table S14, panel A). Frozen-model predictions remained highly stable, with CCC = 0.9826–0.9961 and SROCC = 0.9670–0.9948 (Supplementary Figure S4; Supplementary Table S13, panel B).

Gray-level discretization substantially affected the numerical values of the 205 discretization-dependent test-image features while largely preserving their rank ordering. Median CCC was 0.4143 for FBN16 and 0.5014 for FBN64, whereas median SROCC remained 0.9875 and 0.9931, respectively; the remaining 75 non-discretized intensity features were invariant (Supplementary Table S13, panel A). At the representation level, high rank correlations were observed across all six full-reference representations, while absolute concordance varied across formulations. Among the difference-based representations, normalized difference and stabilized log-ratio showed the highest CCC values, reaching 0.9843 and 0.9761 for FBN16 and 0.9928 and 0.9905 for FBN64, respectively (Supplementary Table S14, panel B). At prediction level, FBN16 yielded CCC = 0.9766 and SROCC = 0.9847, whereas FBN64 yielded CCC = 0.8650 and SROCC = 0.9570 (Supplementary Figure S4; Supplementary Table S13, panel B).

Under controlled additive Gaussian-noise perturbations, stability decreased progressively with increasing noise magnitude. For test-image radiomic features, median CCC decreased from 0.8808 at $\sigma = 0.01$ to 0.1743 at $\sigma = 0.05$, while median SROCC decreased from 0.9659 to 0.7832 (Supplementary Table S13, panel A). Representation-level stability also decreased with increasing noise, with different representations showing different CCC and SROCC profiles across noise levels (Supplementary Table S14, panel C). Frozen-prediction CCC decreased from 0.8453±0.0056 to 0.6491±0.0094, while SROCC decreased from 0.8524±0.0173 to 0.5014±0.0144 and prediction-drift RMSE increased from 0.1524±0.0023 to 0.2630±0.0018 (Supplementary Figure S4; Supplementary Table S13, panel B). These synthetic post-acquisition perturbations were used only to quantify sensitivity and do not constitute repeated-acquisition test-retest measurements.

3.5. Model and dataset diagnostic analyses

Selected feature sets varied across the four outer folds. The 34 fold-specific feature selections corresponded to 28 unique base radiomic features. Five features were selected in at least two folds, one in three folds, and none in all four. Pairwise Jaccard indices between the selected base-feature sets ranged from 0 to 0.1739 (Supplementary Table S15, panel A).

In the grouped fixed-configuration data-size diagnostic, median RMSE was 0.1572 [Q1–Q3, 0.1250–0.1946] with one training reference group (48 pairs), 0.0866 [0.0814–0.1011] with two groups (96 pairs), and 0.1247 [0.1115–0.1543] with three groups (144 pairs) (Supplementary Figure S5; Supplementary Table S15, panel B). The lowest median RMSE was observed with two training reference groups, and no monotonic decrease in RMSE was observed across the three diagnostic conditions. The three-reference condition corresponded to the frozen outer held-reference evaluation, whereas the one- and two-reference conditions were grouped development-validation diagnostics.

Among the four acquisition factors, gain showed the largest median η^2 across evaluable group-feature combinations (0.5082), followed by power (0.1159), dynamic range (0.0417), and frequency (0.0332). Gain had the largest η^2 in 78.5% of these combinations (Supplementary Table S16, panel A). Gain also had the highest η^2 for both the subjective target and the frozen predictions in all four reference groups. The median maximum absolute difference between their corresponding four-factor η^2 profiles was 0.0541 (Supplementary Figure S6; Supplementary Table S16, panel B). Six of the 34 fold-specific feature selections showed a maximum $\eta^2 \geq 0.80$ in the per-reference diagnostic (Supplementary Table S16, panel B). In the separate outer-training-only removal analysis, 9 of the 34 fold-specific feature selections met the gain/dynamic-range removal criterion. All nine were gain-triggered and none were dynamic-range-triggered. After removal, aggregate RMSE was 0.1750 and R^2 was 0.5981, compared with 0.1462 and 0.7196 for the frozen framework (Supplementary Table S16, panel C).

Across the 192 reference-test pairs, the median translation-only displacement was 3.94 pixels, corresponding to 0.93% of the ROI diagonal. The 95th percentile and maximum normalized displacements were 3.09% and 3.27%, respectively, and no pair exceeded 5%. Median valid ROI overlap was 0.9813. Median NMI increased from 1.0726 before alignment to 1.1082 after alignment, with increases in all 192 pairs (Supplementary Table S17, panel A). The Telemed thigh group showed the largest residual translation, with a median normalized displacement of 2.32%. Thirty-eight of 48 pairs exceeded 2%, whereas none exceeded 5% (Supplementary Table S17, panel B).

3.6. Computational runtime and complexity

Using the fold-specific selected features and precomputed reference features, the proposed framework had a component-wise mean runtime estimate of 0.2147±0.2064 s per image pair, with a median of 0.1774 s [Q1–Q3, 0.1163–0.3002]. Without precomputed reference features, the corresponding mean runtime was 0.4244±0.3123 s per pair. Predictor inference alone required 0.000165±0.000066 s per pair, with radiomic feature extraction accounting for nearly all of the total computation. The selected-feature cached-reference implementation was 3.68 times faster than the complete 280-feature cached-reference implementation, which required 0.7897±0.5023 s per pair.

Across the classical FR-IQA methods, mean ROI runtimes ranged from 0.0011 s for PSNR to 0.2285 s for FSIM (Table 3; Fig. 7).

Table 3
Computational runtime on the primary ROI support. Runtime values are reported as mean±standard deviation and median [Q1–Q3] across 240 measured evaluations per method. For the proposed framework, runtime corresponds to the component-wise sum of selected-feature radiomic extraction with precomputed reference features and predictor inference. Disk I/O was excluded from all measurements.

Method	Runtime, mean ±SD (s/pair)	Runtime, median [Q1–Q3] (s/pair)
Proposed radiomics + ML	0.2147±0.2064	0.1774 [0.1163–0.3002]
PSNR	0.0011±0.0047	0.0006 [0.0006–0.0007]
SSIM	0.0034±0.0026	0.0033 [0.0029–0.0034]
MS-SSIM	0.0091±0.0039	0.0088 [0.0083–0.0091]
IW-SSIM	0.0575±0.0092	0.0569 [0.0544–0.0609]
FSIM	0.2285±0.0262	0.2372 [0.2254–0.2432]
VSNR	0.0208±0.0166	0.0190 [0.0174–0.0221]
IFC	0.2097±0.0240	0.2157 [0.2045–0.2283]
VIF	0.2156±0.0251	0.2206 [0.2093–0.2343]
LPIPS	0.0373±0.0069	0.0369 [0.0328–0.0410]
DISTS	0.6352±0.0854	0.6406 [0.5784–0.6931]

IFC, VIF, and FSIM had mean runtimes similar to that of the proposed framework (0.2097, 0.2156, and 0.2285 s, respectively), while showing lower predictive performance. The remaining classical methods required less computation but also showed lower performance than the proposed framework. Among the learned/perceptual methods, LPIPS was faster (0.0373 s per pair) and showed higher performance, whereas DISTS showed higher performance but required 0.6352 s per pair, approximately 2.96 times the runtime of the proposed implementation.

4. Discussion

This study developed and evaluated a radiomics-based machine-learning framework for full-reference assessment of ultrasound image quality under controlled changes in acquisition settings. Its use of radiomics builds on prior image-quality applications in PET and MRI [26–28], while addressing an ultrasound setting in which acquisition-dependent variations affect intensity, speckle texture, contrast, and spatial detail [4–6]. Rather than characterizing each reference-test pair with a single predefined similarity metric, the framework describes it using multiple quantitative radiomic descriptors and learns how their combined changes relate to expert-perceived quality. Four sonographers independently evaluated the paired acquisitions using the SSRQS protocol [19]. Observer ratings showed high inter-observer agreement, and the resulting aggregate target spanned almost the full normalized quality range. Leave-one-observer-out analysis further indicated that the target was not materially driven by any single observer. This provided a stable perceptual reference for model evaluation.

With this perceptual target established, the central question was whether the learned mapping from radiomic representations to perceived quality could generalize to reference acquisitions excluded from model development. The reference-grouped nested design addressed this by excluding one complete reference group from representation fitting, feature selection, model selection, and hyperparameter optimization. The resulting predictions therefore reflect transfer to an unseen reference acquisition rather than interpolation among images derived from the same reference (Table 1; Fig. 6). Performance varied among groups, but the pattern was informative. In the Esaote-thigh fold, rank and linear correlations remained high while R^2 and absolute error deteriorated, consistent with a group-specific calibration mismatch rather than loss of quality ordering. Correlation measures are comparatively insensitive to systematic offset or scaling, whereas R^2 penalizes absolute disagreement. Different feature subsets and model configurations across folds suggest that perceptually relevant information may be distributed across several radiomic descriptors rather than confined to a single universal signature. The limited cross-fold feature recurrence, however, also indicates some dependence on the available reference groups. The non-monotonic data-size diagnostic should not be interpreted as showing that fewer training images are preferable; instead, it indicates that pair count alone does not fully describe training information in clustered data, where reference-group diversity also matters.

The benchmark analysis places these findings in a broader IQA context. Previous studies have shown that agreement between objective

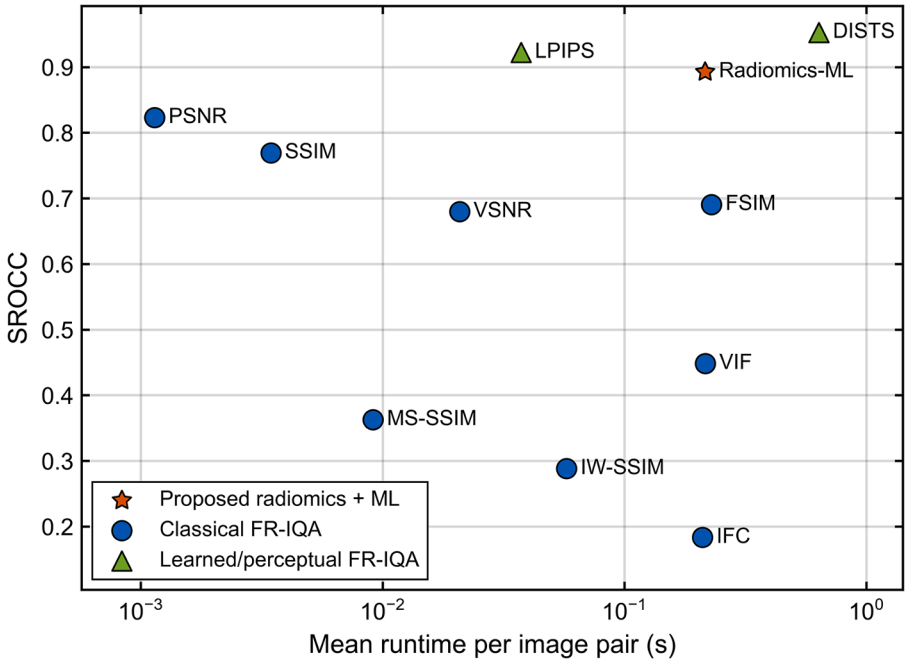


Fig. 7. Accuracy-runtime trade-off on the primary ROI support. SROCC is plotted against mean runtime per image pair on a logarithmic scale for the proposed radiomics-based framework, eight classical FR-IQA methods, LPIPS, and DISTS. The proposed-framework runtime corresponds to the component-wise sum of selected-feature radiomic extraction and predictor inference with precomputed reference features.

IQA measures and expert judgment varies across medical modalities and assessment tasks [10,11,14–16]. Consistent with this, the radiomics-based framework showed stronger aggregate performance than the evaluated classical metrics on the primary ROI support, whereas several classical methods became more competitive on the complete displayed image (Table 2; Supplementary Table S11). Classical metrics should therefore not be regarded as intrinsically inadequate for ultrasound: PSNR and SSIM remain simple, fast, and informative baselines. LPIPS and DISTS achieved higher observed aggregate performance on the primary ROI support, illustrating the effectiveness of learned perceptual representations in this dataset [51,52]. The main value of the radiomic framework in this setting is therefore not peak predictive accuracy, but the interpretability of its intermediate descriptors and the acquisition-setting analyses they enable. With only four reference groups, the exact sign-flip tests had limited inferential resolution and very low statistical power, particularly after Holm correction. Consequently, the nonsignificant comparisons should not be interpreted as evidence of equivalence between methods. The benchmark results are therefore best interpreted as comparative evidence rather than a universal ranking. Comparisons with recent ultrasound-specific learned approaches [22,23] should likewise remain qualitative because their targets, supervision strategies, and clinical tasks differ.

Robustness analyses provide a second layer of evidence by showing where sensitivity enters the pipeline. Small ROI translations and scaling produced little change at feature, representation, and prediction levels, indicating that the framework was not critically dependent on exact pixel-level ROI placement (Supplementary Figure S4). These controlled perturbations characterize local geometric sensitivity of the fixed reference-defined ROIs rather than variability associated with independently defined ROI boundaries. Gray-level discretization produced a different pattern: numerical values of discretization-dependent features changed substantially, while their rank ordering was largely preserved. The full-reference transformation partly moderated this sensitivity, with normalized-difference and stabilized log-ratio representations showing particularly high concordance under FBN changes (Supplementary Table S14). This supports evaluating several representations within training rather than imposing one formulation a priori. Separating feature-, representation-, and prediction-level robustness also shows whether a perturbation affects the descriptors, their reference-test combination, or the final output. Stronger synthetic noise eventually degraded all three levels, defining a clear boundary to robustness and reinforcing that these experiments complement, but do not replace, true repeated-acquisition assessment [29,30].

A further interpretive advantage emerges from the acquisition-setting analysis. Gain had the largest effect on many radiomic descriptors and was also the factor most strongly associated with subjective quality and frozen predictions across all four reference groups (Supplementary Figure S6; Supplementary Table S16). Model sensitivity to gain therefore mirrors a perceptual dependency present in the experiment rather than, by itself, demonstrating shortcut learning. Removing strongly gain- or dynamic-range-associated selected features reduced performance without eliminating predictive capability, indicating that setting-related information contributes meaningfully to prediction but does not fully explain it. This is an important consequence of the radiomic formulation: instead of observing only whether a final score correlates with perception, one can trace how acquisition factors affect image descriptors and how this information propagates into prediction. Such analyses make model behavior more inspectable and may help distinguish perceptually meaningful sensitivity from excessive dependence on a single setting.

These characteristics also clarify the intended application. Full-reference IQA requires a meaningful matched reference and is therefore most naturally suited to controlled system evaluation, acquisition-protocol development, parameter optimization, and quality-control experiments rather than unconstrained bedside assessment. In these settings, the framework could complement conventional metrics by providing both a continuous quality estimate and an analyzable description of how image characteristics change relative to the reference. Restricting extraction to retained features substantially reduced computational burden, although classical metrics remained markedly faster and LPIPS offered a favorable accuracy-runtime combination in the present CPU benchmark. DISTS achieved higher observed performance at greater computational cost (Fig. 7). Method choice should therefore reflect the evaluation objective, including the relative importance of computational simplicity, predictive performance, and access to acquisition- and feature-level information.

The controlled design should nevertheless be considered when interpreting the findings. Four reference groups derived from a single rabbit cadaver, spanning two anatomies and two probe-scanner configurations, provided a focused setting for isolating acquisition-setting effects, but broader biological, in vivo, and inter-system variability remains to be established. Reference images were selected through expert-supervised visual inspection rather than a predefined quantitative criterion or multi-rater consensus, making the full-reference comparisons dependent on the selected references. Subjective ratings were based on complete displayed images, whereas radiomic analysis used manually defined fixed ROIs, introducing a difference in spatial support between perception and prediction. Subjective assessment involved four sonographers, with lower within-observer repeatability for Observer 3 indicating some heterogeneity in rating consistency across observers. Nevertheless, leave-one-observer-out concordance of the aggregate target was consistently high, indicating that the aggregate target was robust to the exclusion of any single observer. Image format was coupled to scanner configuration, preventing separation of device-related and storage-format contributions. True acquisition test-retest reliability could not be quantified using ICC or CCC because repeated identical acquisitions were not available. Future studies should therefore evaluate larger multi-system in vivo cohorts with repeated acquisitions and independently or consensually defined references.

5. Conclusions

This study presents a radiomics-based machine-learning framework for full-reference ultrasound image quality assessment under controlled acquisition-setting variations. By combining quantitative intensity and texture descriptors with reference-grouped model development, the framework generated out-of-reference predictions that remained closely aligned with expert subjective assessments across the four reference groups. On the primary ROI support, it showed stronger observed aggregate performance than the evaluated classical FR-IQA methods, whereas learned perceptual methods showed higher observed aggregate performance. Robustness analyses indicated stable predictions under modest ROI and discretization changes, while acquisition-setting analyses enabled tracing how setting-related image changes propagated from radiomic features to predictions. These findings support radiomic full-reference representations as an interpretable framework for controlled imaging-system evaluation and protocol optimization. Larger multi-system, in vivo studies with repeated acquisitions and independently defined references are needed to establish broader generalizability and practical applicability.

Ethics statement

No experiments were performed on live animals. Ultrasound images were acquired from rabbit cadaver tissue only. According to institutional guidelines, no specific ethical approval was required for the use of cadaveric animal material in this study.

Funding

This study was supported by the Italian Ministry of University and Research (MUR) through the Fondo Italiano per le Scienze Applicate

(FISA) under the project *SURF-UP – Surgery by Using a Robotic Focused Ultrasound Platform* (grant code: FISA-2022-00770), by SOUNDSAFE CARE SRL (project "*Soundsafe Vet Care*" – CUP: I57H24001520008, COR: 22717933 – funded by Fondo Next Generation EU – PNRR – "M4C2I2.3"), and by the Italian Ministry of University and Research (MUR) through the PRIN 2022 program (Progetti di Rilevante Interesse Nazionale) under the project *FUtuRo – Focused Ultrasound Surgery enabled by Robotics and Simulation* (project code: 2022342MEF).

CRediT authorship contribution statement

Alessandro Piccolo: Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis, Conceptualization. **Vito Papapicco:** Writing – review & editing, Software, Resources, Methodology, Investigation, Data curation. **Andrea Mariani:** Writing – review & editing, Resources, Methodology, Funding acquisition. **Arianna Menciassi:** Writing – review & editing, Supervision, Funding acquisition. **Andrea Cafarelli:** Writing – review & editing, Validation, Supervision, Methodology, Funding acquisition.

Declaration of competing interest

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

Acknowledgements

The authors would like to thank Dr. Paolo Spinnato and a collaborating specialist (Diagnostic and Interventional Radiology, IRCCS Istituto Ortopedico Rizzoli, Bologna, Italy), Dr. Alessandro Locatelli (AniCura, Clinica Veterinaria Tibaldi, Milan, Italy), and Dr. Caterina Puccinelli (Department of Veterinary Sciences, University of Pisa, San Piero a Grado, Pisa, Italy) for their valuable contribution to the subjective image quality assessment and for providing expert clinical insight.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2026.109672.

Data availability

The code, derived data, metadata, subjective targets (MOS-derived data), ROI coordinates and binary ROI masks, reference-grouped split assignments, random seeds, model configurations, and radiomic feature tables supporting the computational analyses of this study are publicly available at https://github.com/alepiccolo99/ultrasound-iqa-radiomics. These derived data and metadata enable reproduction of the downstream modeling and statistical analyses reported in this study. The original ultrasound images are not publicly distributed because they are subject to third-party data-sharing restrictions.

References

[1] M. Maar, J. Lee, A. Tardi, Y.Y. Zheng, C. Wong, J. Gao, Inter-transducer variability of ultrasound image quality in obese adults: qualitative and quantitative comparisons, Clin. Imaging 92 (2022) 63–71, https://doi.org/10.1016/j.clinimag.2022.09.010.

[2] M.P. Caserta, S.L. Bonnett, M.C. La Valley, S. De Meo, A.W. Bowman, Ultrasound practice redesign to improve image quality: implementation of a quality control sonographer, J. Am. Coll. Radiol. 17 (12) (2020) 1644–1652, https://doi.org/10.1016/j.jacr.2020.07.015.

[3] M.K. Welleweerd, A.G. de Groot, S.O. De Looijer, F.J. Siepel, S. Stramigioli, Automated robotic breast ultrasound acquisition using ultrasound feedback, in: 2020 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2020, pp. 9946–9952, https://doi.org/10.1109/ICRA40945.2020.9196736.

[4] J.A. Noble, Reflections on ultrasound image analysis, Med. Image Anal. 33 (2016) 33–37, https://doi.org/10.1016/j.media.2016.06.015.

[5] N.M. Duggan, A.J. Goldsmith, A.A. Al Saud, I.W. Ma, H. Shokoohi, A.S. Liteplo, Optimizing lung ultrasound: the effect of depth, gain and focal position on sonographic B-lines, Ultrasound Med. Biol. 48 (8) (2022) 1509–1517, https://doi.org/10.1016/j.ultrasmedbio.2022.03.015.

[6] O.M. Rindal, A. Austeng, A. Fatemi, A. Rodriguez-Molares, The effect of dynamic range alterations in the estimation of contrast, IEEE Trans. Ultrason. Ferroelectr. Freq. Control 66 (7) (2019) 1198–1208, https://doi.org/10.1109/TUFFC.2019.2911267.

[7] D. Raina, D. Ntentia, S.H. Chandrashekhara, R. Voyles, S.K. Saha, Expert-agnostic ultrasound image quality assessment using deep variational clustering, in: 2023 IEEE International Conference on Robotics and Automation (ICRA), London, United Kingdom, 2023, pp. 2717–2723, https://doi.org/10.1109/ICRA48891.2023.10160435.

[8] L.S. Chow, R. Paramesran, Review of medical image quality assessment, Biomed. Signal Process. Control 27 (2016) 145–154, https://doi.org/10.1016/j.bspc.2016.02.006.

[9] S. Athar, Z. Wang, A comprehensive performance evaluation of image quality assessment algorithms, IEEE Access 7 (2019) 140030–140070, https://doi.org/10.1109/ACCESS.2019.2943319.

[10] A. Breger, A. Biguri, M.S. Landman, I. Selby, N. Amberg, E. Brunner, J. Gröhl, S. Hatamikia, C. Karner, L. Ning, S. Dittmer, A study of why we need to reassess full reference image quality assessment with medical images, J. Imaging Inform. Med. 38 (6) (2025) 3444–3469, https://doi.org/10.1007/s10278-025-01462-1.

[11] A. Breger, C. Karner, I. Selby, J. Gröhl, S. Dittmer, E. Lilley, J. Babar, J. Beckford, T. R. Else, T.J. Sadler, S. Shahipasand, A study on the adequacy of common iqa measures for medical images, in: International Conference on Medical Imaging and Computer-Aided Diagnosis, Springer Nature Singapore, Singapore, 2024, pp. 451–462, https://doi.org/10.1007/978-981-96-3863-5_41.

[12] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612, https://doi.org/10.1109/TIP.2003.819861.

[13] H.R. Sheikh, M.F. Sabir, A.C. Bovik, A statistical evaluation of recent full reference image quality assessment algorithms, IEEE Trans. Image Process. 15 (11) (2006) 3440–3451, https://doi.org/10.1109/TIP.2006.881959.

[14] A. Mason, J. Rioux, S.E. Clarke, A. Costa, M. Schmidt, V. Keough, T. Huynh, S. Beyea, Comparison of objective image quality metrics to expert radiologists' scoring of diagnostic quality of MR images, IEEE Trans. Med. Imaging 39 (4) (2019) 1064–1072, https://doi.org/10.1109/TMI.2019.2930338.

[15] W. Lee, F. Wagner, A. Galdran, Y. Shi, W. Xia, G. Wang, X. Mou, M.A. Ahamed, A. A. Imran, J.E. Oh, K. Kim, Low-dose computed tomography perceptual image quality assessment, Med. Image Anal. 99 (2025) 103343, https://doi.org/10.1016/j.media.2024.103343.

[16] A. Breger, J. Gröhl, C. Karner, T.R. Else, I. Selby, T. Rix, L.S. Witt, M. Duchêne, J. Weir-McCall, C.B. Schönlieb, PhotIQA: a photoacoustic image data set with image quality ratings, Sci. Data (2026), https://doi.org/10.1038/s41597-026-07610-8.

[17] S. Lang, M. Zhou, X. Wei, J. Yan, Y. Feng, W. Jia, Image quality assessment: exploring the similarity of deep features via covariance-constrained spectra, IEEE Trans. Broadcast. 72 (1) (2026) 303–316, https://doi.org/10.1109/TBC.2025.3609055.

[18] W. Shen, M. Zhou, Y. Chen, X. Wei, Y. Feng, H. Pu, W. Jia, Image quality assessment: investigating causal perceptual effects with abductive counterfactual inference, in: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2025, pp. 17990–17999, https://doi.org/10.1109/CVPR52734.2025.01676.

[19] M.C. Hemmsen, M.M. Petersen, S.I. Nikolov, M.B. Nielsen, J.A. Jensen, Ultrasound image quality assessment: a framework for evaluation of clinical image quality, in: Medical Imaging 2010: Ultrasonic Imaging, Tomography, and Therapy 7629, SPIE, 2010, pp. 105–116, https://doi.org/10.1117/12.840664.

[20] S. Zhang, Y. Wang, J. Jiang, J. Dong, W. Yi, W. Hou, CNN-based medical ultrasound image quality assessment, Complexity 2021 (1) (2021) 9938367, https://doi.org/10.1155/2021/9938367.

[21] K. Nesovic, R.G. Koh, A.A. Sereshki, F.S. Zadeh, M.R. Popovic, D. Kumbhare, Ultrasound image quality evaluation using a structural similarity based autoencoder, in: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE, 2021, pp. 4002–4005, https://doi.org/10.1109/EMBC46164.2021.9630261.

[22] T. Liu, X. Zhang, X. Ma, S. Chen, X. Wang, P. Zhou, Y. Chen, G. Zhou, F. Lv, DAM: degradation-aware model for ultrasound image quality assessment, IEEE J. Biomed. Health Inf. 29 (11) (2025) 8233–8245, https://doi.org/10.1109/JBHI.2025.3572459.

[23] L. Liu, T. Wang, W. Zhu, H. Zhang, H. Tian, Y. Li, W. Cai, P. Yang, Intelligent quality assessment of ultrasound images for fetal nuchal translucency measurement during the first trimester of pregnancy based on deep learning models, BMC Pregnancy Childbirth 25 (1) (2025) 741, https://doi.org/10.1186/s12884-025-07863-y.

[24] E. Fleury, K. Marcomini, Performance of machine learning software to classify breast lesions using BI-RADS radiomic features on ultrasound images, Eur. Radiol. Exp. 3 (1) (2019) 34, https://doi.org/10.1186/s41747-019-0112-7.

[25] S. Latha, P. Muthu, S. Dhanalakshmi, R. Kumar, K.W. Lai, X. Wu, Emerging feature extraction techniques for machine learning-based classification of carotid artery ultrasound images, Comput. Intell. Neurosci. 2022 (1) (2022) 1–14, https://doi.org/10.1155/2022/1847981.

[26] G. Reynés-Llompart, A. Sabaté-Llobera, E. Llinares-Tello, J.M. Martí-Climent, C. Gámez-Cenzano, Image quality evaluation in a modern PET system: impact of new reconstructions methods and a radiomics approach, Sci. Rep. 9 (1) (2019) 10640, https://doi.org/10.1038/s41598-019-46937-8.
[27] Y. Hao, J. Zheng, W. Li, W. Zhao, J. Zheng, H. Wang, J. Ren, G. Zhang, J. Zhang, Ultra-high b-value DWI in rectal cancer: image quality assessment and regional lymph node prediction based on radiomics, Eur. Radiol. 35 (1) (2025) 49–60, https://doi.org/10.1007/s00330-024-10958-3.
[28] L.C. Thijssen, M. de Rooij, J.O. Barentsz, H.J. Huisman, Radiomics based automated quality assessment for T2W prostate MR images, Eur. J. Radiol. 165 (2023) 110928, https://doi.org/10.1016/j.ejrad.2023.110928.
[29] V. Kumar, Y. Gu, S. Basu, A. Berglund, S.A. Eschrich, M.B. Schabath, K. Forster, H. J. Aerts, A. Dekker, D. Fenstermacher, D.B. Goldgof, Radiomics: the process and the challenges, Magn. Reson. Imaging 30 (9) (2012) 1234–1248, https://doi.org/10.1016/j.mri.2012.06.010.
[30] A. Zwanenburg, M. Vallières, M.A. Abdalah, H.J. Aerts, V. Andrearczyk, A. Apte, S. Ashrafinia, S. Bakas, R.J. Beukinga, R. Boellaard, M. Bogowicz, The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping, Radiology 295 (2) (2020) 328–338, https://doi.org/10.1148/radiol.2020191145.
[31] Recommendation ITU-R BT. 500-14: methodologies for the subjective assessment of the quality of television images. 2019.
[32] D.W. Scott, Sturges' Rule, 1, Wiley Interdisciplinary Reviews: Computational Statistics, 2009, pp. 303–306, https://doi.org/10.1002/wics.35.
[33] L.I. Lin, A concordance correlation coefficient to evaluate reproducibility, Biometrics 45 (1989) 255, https://doi.org/10.2307/2532051.
[34] K.O. McGraw, S.P. Wong, Forming inferences about some intraclass correlation coefficients, Psychol. Methods 1 (1) (1996) 30–46. https://psycnet.apa.org/doi/10.1037/1082-989X.1.1.30.
[35] T.K. Koo, M.Y. Li, A guideline of selecting and reporting intraclass correlation coefficients for reliability research, J. Chiropr. Med. 15 (2) (2016) 155–163, https://doi.org/10.1016/j.jcm.2016.02.012.
[36] X. Kai, Y. Jie, Z.Y. Min, L.X. Liang, HVS-based medical image compression, Eur. J. Radiol. 55 (1) (2005) 139–145, https://doi.org/10.1016/j.ejrad.2004.09.007.
[37] A.P. Bradley, A wavelet visible difference predictor, IEEE Trans. Image Process. 8 (5) (1999) 717–730, https://doi.org/10.1109/83.760338.
[38] B. Ergen, M. Baykara, Texture based feature extraction methods for content based medical image retrieval systems, Biomed. Mater. Eng. 24 (6) (2014) 3055–3062, https://doi.org/10.3233/BME-141127.
[39] R.M. Haralick, K. Shanmugam, I.H. Dinstein, Textural features for image classification, IEEE Trans. Syst. Man Cybern. SMC-3 (6) (1973) 610–621, https://doi.org/10.1109/TSMC.1973.4309314.
[40] J.A. Oliver, M. Budzevich, G.G. Zhang, T.J. Dilling, K. Latifi, E.G. Moros, Variability of image features computed from conventional and respiratory-gated PET/CT images of lung cancer, Transl. Oncol. 8 (6) (2015) 524–534, https://doi.org/10.1016/j.tranon.2015.11.013.
[41] A. Chu, C.M. Sehgal, J.F. Greenleaf, Use of gray value distribution of run lengths for texture analysis, Pattern Recognit. Lett. 11 (6) (1990) 415–419, https://doi.org/10.1016/0167-8655(90)90112-F.
[42] M.M. Galloway, Texture analysis using gray level run lengths, Comput. Graph. Image Process. 4 (2) (1975) 172–179, https://doi.org/10.1016/S0146-664X(75)80008-6.
[43] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. B: Stat. Methodol. 58 (1) (1996) 267–288, https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
[44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[45] Z. Wang, E.P. Simoncelli, A.C. Bovik, Multiscale structural similarity for image quality assessment, in: The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003 2, IEEE, 2003, pp. 1398–1402, https://doi.org/10.1109/ACSSC.2003.1292216.
[46] Z. Wang, Q. Li, Information content weighting for perceptual image quality assessment, IEEE Trans. Image Process. 20 (5) (2010) 1185–1198, https://doi.org/10.1109/TIP.2010.2092435.
[47] L. Zhang, L. Zhang, X. Mou, D. Zhang, FSIM: a feature similarity index for image quality assessment, IEEE Trans. Image Process. 20 (8) (2011) 2378–2386, https://doi.org/10.1109/TIP.2011.2109730.
[48] D.M. Chandler, S.S. Hemami, VSNR: a wavelet-based visual signal-to-noise ratio for natural images, IEEE Trans. Image Process. 16 (9) (2007) 2284–2298, https://doi.org/10.1109/TIP.2007.901820.
[49] H.R. Sheikh, A.C. Bovik, G. De Veciana, An information fidelity criterion for image quality assessment using natural scene statistics, IEEE Trans. Image Process. 14 (12) (2005) 2117–2128, https://doi.org/10.1109/TIP.2005.859389.
[50] H.R. Sheikh, A.C. Bovik, Image information and visual quality, IEEE Trans. Image Process. 15 (2) (2006) 430–444, https://doi.org/10.1109/TIP.2005.859378.
[51] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2018, pp. 586–595, https://doi.org/10.1109/CVPR.2018.00068.
[52] K. Ding, K. Ma, S. Wang, E.P. Simoncelli, Image quality assessment: unifying structure and texture similarity, IEEE Trans. Pattern Anal. Mach. Intell. 44 (5) (2020) 2567–2581, https://doi.org/10.1109/TPAMI.2020.3045810.
[53] C. Studholme, D.L. Hill, D.J. Hawkes, An overlap invariant entropy measure of 3D medical image alignment, Pattern Recognit. 32 (1) (1999) 71–86, https://doi.org/10.1016/S0031-3203(98)00091-0.
[54] S. Holm, A simple sequentially rejective multiple test procedure, Scand. J. Stat. 65–70 (1979).