REVIEW ARTICLE

Evaluation Metrics in Digital Image Processing: A Task-Oriented Review and Practical Selection Framework

Muzammil Khan[1], **Muhammad Hamza**[1], **Sikandar Ali**[1], **Ibrar Ali**[1] and **Zaheer Khan**[2,*]

[1] Department of Computer and Software Technology, University of Swat, Swat 19120, Pakistan
[2] Spinghar Technical and Vocational Training Center, Jalalabad 2601, Afghanistan

Abstract

The rapid development of digital image processing and machine learning has increased the need for reliable and task-appropriate evaluation methods. However, the wide range of available metrics, each with different assumptions, strengths, and limitations, makes metric selection challenging across diverse image-processing tasks. This article presents a task-oriented review of commonly used evaluation metrics and proposes a practical framework for selecting appropriate metrics across five major categories: image quality assessment, image classification, image segmentation, image restoration, and object detection. Representative metrics are reviewed based on their mathematical basis, interpretation, applications, advantages, and limitations. Particular attention is given to pixel-wise measures, including MAE, MSE, and RMSE, perceptual and structural measures such as PSNR, SSIM, MS-SSIM, FSIM, VIF, and LPIPS, and classification and detection measures including accuracy, precision, recall, F1-score, ROC-AUC, confusion matrices, and mean average precision. Overlap-based metrics such as Dice and IoU are also examined for segmentation. The review emphasizes that no single metric adequately captures all dimensions of image-processing performance. Therefore, metric selection should consider task characteristics, data properties, reference availability, and perceptual requirements, with complementary metrics used where appropriate. The proposed framework aims to support researchers and practitioners in selecting appropriate and interpretable evaluation metrics for digital image processing and machine learning applications.

Keywords: digital image processing, evaluation metrics, image quality assessment, image classification, image segmentation, image restoration, object detection.

1 Introduction

Digital image processing is a rapidly evolving research field that uses computational algorithms to manipulate digital images and to analyze and enhance their content. As digital imaging technology continues to advance, its applications span various domains, including healthcare, entertainment, security, transportation, and scientific research [1]. This

Submitted: 08 August 2026
Revised: 30 August 2026
Accepted: 15 September 2026
Published: 11 October 2026

Vol. 2, **No.** 4, 2026.
10.62762/JIAP.2026.838036

***Corresponding author:**
✉ Zaheer Khan
xaheerkhancs@gmail.com

Citation
Khan, M., Hamza, M., Ali, S., Ali, I., & Khan, Z. (2026). Evaluation Metrics in Digital Image Processing: A Task-Oriented Review and Practical Selection Framework. *ICCK Journal of Image Analysis and Processing*, 2(4), 231–251.

© 2026 by the Authors. Published by Institute of Central Computation and Knowledge. This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

proliferation of digital images and the associated data has driven researchers and engineers to develop sophisticated methods for processing and interpreting visual information efficiently and accurately [1–3].

The range of applications is broad. In healthcare, image processing supports diagnostics, treatment planning, and surgical guidance [1, 4, 5]. In security and surveillance it underlies facial recognition, license plate recognition, and anomaly detection in video footage [6, 7], and closely related techniques allow autonomous vehicles to detect objects and lanes, map their environment, and combine camera and sensor data for real-time decisions [6, 7]. Remote sensing depends on the analysis of satellite and aerial imagery for climate monitoring, disaster management, and urban planning [8, 9]. Image and video processing are also central to content creation, special effects, video enhancement, and augmented reality in the entertainment and media sector [2, 10], while in astronomy, biology, and materials science they help researchers visualize and analyze complex phenomena [2, 8, 10].

These applications share a set of persistent difficulties. The growing resolution of imaging devices and the sheer volume of image data call for efficient storage, processing, and retrieval [2, 3, 7], and images captured in real-world conditions are frequently degraded by noise, blur, and other distortions that algorithms must tolerate [3, 4]. Autonomous driving and surveillance additionally impose real-time constraints, so that low latency becomes as important as accuracy [1, 7]. The widespread collection of image data raises privacy and security concerns, which motivate secure processing and encrypted storage [1]. Finally, models trained on one dataset often fail to generalize to new environments; improving robustness and adaptability, for example through data augmentation [11], remains an open research problem [1, 3, 7].

Machine learning (ML) has changed digital image processing substantially by enabling automated feature extraction, pattern recognition, and decision-making. Traditional rule-based methods often required manual feature engineering and domain-specific knowledge, whereas ML models can learn these features directly from the data [1, 3, 12]. Convolutional neural networks (CNNs) in particular extract hierarchical features automatically, which has improved performance in classification and segmentation [1, 3, 10]. In image classification, such models assign images to predefined categories with high accuracy, in settings that range from medical diagnostics to everyday photographs [1, 3, 6, 13, 14]. For detecting and delineating objects, architectures such as Faster R-CNN, YOLO (You Only Look Once), and Mask R-CNN are in wide use [1, 8]. Learning-based methods have also been successful in super-resolution, denoising, and inpainting [1, 3, 7], and generative models such as Generative Adversarial Networks (GANs) can synthesize realistic images, increase resolution, and transfer styles between images [1, 3, 7].

Several research directions are currently shaping the field. Because supervised learning requires large annotated datasets, self-supervised learning seeks to learn representations from unlabeled data, and few-shot learning aims to generalize from a handful of examples [1, 15, 16]. Vision Transformers (ViTs) and their variants, which adapt an architecture originally designed for natural language processing to sequences of image patches, now achieve state-of-the-art results in classification and other vision tasks [1, 17, 18]. The demand for real-time processing has made edge computing important, and considerable effort goes into adapting algorithms to resource-constrained devices [1]. In the same decentralized spirit, federated learning trains models across devices without sharing raw data, which is attractive for sensitive images such as those in healthcare [1]. As models grow more complex, explainable AI (XAI) techniques are needed to show how they interpret images and reach their predictions [19, 20]. Biomedical image analysis continues to move towards more accurate and interpretable models, with multi-modal fusion and 3D analysis gaining prominence [4, 5]. There is also growing attention to the environmental cost of these models, which has prompted work on energy-efficient architectures and better hardware utilization [1].

Taken together, machine learning has made image processing more accurate, more efficient, and more automated, while data privacy, real-time operation, and generalization remain unresolved. Whether a new method actually improves on these points can only be established by measuring its performance, and this is where the choice of evaluation metrics becomes decisive. The remainder of this article therefore reviews the metrics in common use, organizes them by task, and discusses how they should be selected and interpreted.

2 Importance of Evaluation Metrics

Evaluating machine learning (ML) models is central to digital image processing because it determines how effective, efficient, and adaptable they are judged to be. Evaluation metrics provide standardized ways to assess how well models perform on specific tasks, identify their limitations, and drive improvements. In the context of image processing, where tasks can range from image classification and segmentation to object detection and image restoration, selecting and applying the right evaluation metrics is essential for the development and deployment of robust models [5, 21, 22]. The assessment of image quality encompasses both subjective evaluation by human observers and objective computational measures, and understanding their relationship is essential for designing metrics that reflect perceived quality [23].

2.1 Assessing Model Performance

One of the primary reasons for evaluating ML models is to objectively assess their performance. Performance metrics such as accuracy, precision, recall, F1-score, Intersection over Union (IoU), and the structural similarity index (SSIM) [24] offer quantitative measures to gauge how well a model meets the desired objectives [21, 22]. Which measure is appropriate depends on the task. In image classification, accuracy is the usual starting point, but it can be insufficient for imbalanced datasets in which a few classes dominate; precision, recall, and the F1-score then give a more nuanced picture [21, 22]. In segmentation, IoU and the Dice coefficient quantify the overlap between predicted and ground truth regions and therefore indicate how accurately objects are delineated [25–28]. Denoising, super-resolution, and restoration are typically assessed with the Peak Signal-to-Noise Ratio (PSNR) and SSIM, which compare the reconstructed image with a reference [5, 21, 29]. Beyond a single test set, evaluating models across datasets and acquisition conditions, for instance with the mean absolute error (MAE) or with confusion matrices, shows how stable their behavior is under varying inputs [21, 22].

2.2 Comparative Analysis of Models

Evaluation metrics facilitate the comparison of different models. In research and development, comparing the performance of new models against state-of-the-art baselines is essential for validating improvements [5, 22, 25]. Public benchmarks such as ImageNet and COCO serve this purpose by fixing standardized evaluation protocols, with Top-1 and Top-5 accuracy as the customary classification measures; they allow researchers to judge the relative strengths of their models and support reproducibility [17, 30–33]. The same metrics are indispensable in ablation studies, where they provide quantitative evidence for the contribution of individual architectural components and design choices [32].

2.3 Model Interpretability and Explainability

Evaluation metrics are not only used to measure accuracy but also to understand how models make decisions. Explainability and interpretability are becoming increasingly important in ML, particularly in high-stakes applications [17, 19]. Techniques such as Grad-CAM, for example, produce saliency maps that highlight the image regions most responsible for a prediction, and dedicated measures are then needed to assess how faithful these explanations are [19, 20, 34]. A related concern is behavior under adversarial attack: adversarial accuracy and other robustness scores quantify the resilience of a model and are essential for building secure systems [17, 35, 36].

2.4 Driving Model Optimization

The feedback provided by evaluation metrics is essential for model optimization. By identifying performance bottlenecks, developers can fine-tune hyperparameters, optimize architectures, and improve training processes [5, 22, 37]. In practice, validation metrics guide the choice of hyperparameters such as the learning rate, batch size, and regularization strength, and they determine when training should be stopped to prevent overfitting [5, 22].

2.5 Ensuring Fairness and Bias Mitigation

Evaluation metrics are also needed to identify and mitigate biases in ML models. In image processing, biases can manifest in the form of differential performance across demographic groups [38, 39]. Fairness measures such as demographic parity and equalized odds make these disparities explicit [39, 40], and evaluating models on diverse datasets helps to verify that they generalize across populations and environments [39, 41].

2.6 Keeping Up with Emerging Trends

The field of image processing is evolving rapidly, with new challenges and evaluation metrics emerging. Recent advances, such as vision transformers (ViTs) and self-supervised learning models, require novel evaluation approaches [17, 25, 32, 49]. New tasks



Figure 1. Taxonomy of evaluation metrics for digital image processing.

such as multimodal image processing and 3D image analysis have led to task-specific evaluation metrics [58]. Likewise, recent surveys of Vision Transformers suggest that traditional metrics may not fully capture

Table 1. Task-oriented framework for selecting evaluation metrics in digital image processing.

Category	Metrics That Can Apply	Preferred Metrics (Current Research)	Why Preferred?
Image Quality Assessment (IQA)	PSNR, SSIM, MS-SSIM, FSIM, VIF, LPIPS, NIQE, MSE, RMSE, MAE	SSIM, MS-SSIM, VIF, LPIPS	Better correlation with human perception; capture structures, semantics, and deep features; robust to distortions [5, 42–46].
Image Classification	Accuracy, Precision, Recall, F1-Score, AUC-ROC, Confusion Matrix	Accuracy, F1-Score, AUC-ROC	Handle class imbalance; provide global and threshold-independent performance measurement; widely used in ML benchmarks [5, 17, 21, 47, 48].
Image Segmentation	IoU, Dice Coefficient, Pixel Accuracy, Precision, Recall, F1-Score, Hausdorff Distance	IoU, Dice Coefficient	Measure region overlap; directly reflect segmentation quality; standard in medical imaging and deep learning segmentation tasks [5, 49–53].
Image Restoration	PSNR, SSIM, MS-SSIM, RMSE, LPIPS, VIF, Spectral Angle Mapper (SAM)	SSIM, LPIPS, PSNR	SSIM and LPIPS capture perceptual quality; PSNR remains a strong baseline for fidelity comparison [5, 42–46, 54].
Object Detection	mAP, AP@0.5, AP@[0.5:0.95], Precision-Recall, IoU, Average Recall (AR)	mAP, AP@[0.5:0.95]	Provide threshold-sweep evaluation; standard in COCO and Pascal VOC benchmarks; measure both localization and classification accuracy [31, 55–57].

the capabilities of these architectures, which motivates complementary evaluation criteria [17, 49].

In summary, evaluation metrics serve several distinct purposes: they establish how well a model performs, make models comparable, support interpretation, steer optimization, and expose bias. No single metric fulfils all of these purposes, which is why the following sections treat the metrics by task.

3 Categories of Metrics

Evaluation metrics in digital image processing are essential for assessing the quality and performance of image processing algorithms. In this review they are grouped into five categories according to the task they serve; Table 1 summarizes the resulting task-oriented selection framework, and Figure 1 shows the corresponding taxonomy.

Image quality assessment (IQA) metrics evaluate how well an image preserves structural details, contrast, and perceptual quality. This group comprises the Peak Signal-to-Noise Ratio (PSNR), the Structural Similarity Index (SSIM), the Mean Squared Error (MSE), Multi-Scale SSIM (MS-SSIM) [59], and the Feature Similarity Index (FSIM). Image classification metrics are used with machine learning and deep learning models for image recognition and include accuracy, precision, recall, the F1-score, and the Receiver Operating Characteristic (ROC) curve with its Area Under the Curve (AUC). Image segmentation metrics evaluate how well an image is partitioned into meaningful regions; the most common are the Intersection over Union (IoU), the Dice coefficient (the F1-score for segmentation), and the Hausdorff Distance (HD) [60].

Image restoration metrics are used to evaluate denoising, super-resolution, and inpainting methods. In addition to the fidelity measures above, they include the Learned Perceptual Image Patch Similarity (LPIPS), the Natural Image Quality Evaluator (NIQE) [61], and Visual Information Fidelity (VIF) [62]. Finally, object detection algorithms are evaluated with Average Precision (AP), mean Average Precision (mAP), and the IoU between predicted and reference bounding boxes.

Figure 1 gives the reader a visual map of the entire review before the detailed metric sections. The taxonomy follows the five task categories introduced above.

4 Understanding Evaluation Metrics

In this section, we briefly discuss common evaluation metrics and their interpretation. Unless stated otherwise, I denotes the reference (original) image and K the processed image, both of size $M \times N$ pixels. The interpretation ranges given in the tables of this section are indicative rules of thumb for 8-bit images (intensity range 0-255); acceptable values depend on the bit depth, the image content, and the application.

4.1 Mean Squared Error (MSE)

Mean Squared Error (MSE) is a fundamental evaluation metric in digital image processing that quantifies the difference between two images: the original image and a processed (reconstructed or enhanced) image. It calculates the average squared difference between the corresponding pixel intensities of the two images [42, 43, 63].

MSE appears throughout image processing. In image compression (e.g., JPEG) it is a standard measure of fidelity and is usually reported together with PSNR [64, 65]. In restoration tasks such as denoising and inpainting, a lower MSE indicates a better result [66–68], and in enhancement tasks such as super-resolution it measures how closely the output matches the original [69, 70]. It is also one of the most common loss functions for training deep reconstruction models [69, 71].

A lower MSE indicates that the processed image closely resembles the original image, meaning fewer distortions or errors. Conversely, a higher MSE suggests greater deviation and image quality degradation. The mathematical formula for MSE is:

$$\mathrm{MSE} = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} \big(I(i,j) - K(i,j)\big)^2 \qquad (1)$$

where M and N are the dimensions (height and width) of the image, $I(i,j)$ is the pixel intensity at position (i,j) in the original image, $K(i,j)$ is the pixel intensity at the same position in the processed image, and $1/(MN)$ is the normalization factor that yields the average squared difference.

MSE is often reported together with PSNR. Table 2 relates MSE values to perceived image quality.

Despite its simplicity, MSE does not always align with human perception, and it should therefore be complemented by a perceptual measure such as SSIM.

Table 2. Interpretation of Mean Squared Error (MSE) values for image quality assessment.

MSE Value	Image Quality	Perceived Difference
0	Identical images	No difference
< 1	Excellent quality	Almost imperceptible distortion
1-10	Good quality	Minor blurring/noise
10-50	Noticeable degradation	Moderate loss of detail
> 50	Poor quality	Severe distortion

4.2 Peak Signal-to-Noise Ratio (PSNR)

Peak Signal-to-Noise Ratio (PSNR) is one of the most widely used metrics in digital image processing for assessing image quality. It quantifies the ratio between the maximum possible pixel intensity (signal) and the level of noise (error) introduced during image processing, such as compression, filtering, or restoration.

PSNR is derived from the MSE, making it useful for evaluating the extent to which an image has degraded after processing. A higher PSNR value indicates better image quality, meaning the processed image is closer to the original, whereas a lower PSNR value signifies greater distortion and lower quality [43, 45, 63, 72].

PSNR is the customary figure of merit for lossy compression (e.g., JPEG, WebP, H.264/AVC, HEVC), where it indicates how much quality is lost as the file size is reduced [43, 45, 63]. It is used in the same way to evaluate noise removal and restoration methods [73–75] and the accuracy of reconstructed high-resolution images in super-resolution [45, 75]. In video quality assessment, frames of a compressed sequence are compared against the original frames [45, 74], and PSNR is used in streaming and broadcasting to check that quality remains acceptable.

Its popularity rests on practical grounds: PSNR is easy to compute and interpret, it yields a single number on which methods can be compared, and it is accepted across academia and industry [43, 63, 75]. Its main weakness, discussed below, is that it does not always correlate well with human visual perception. PSNR is calculated using the formula:

$$\mathrm{PSNR} = 10 \log_{10} \left(\frac{\mathrm{MAX}_I^2}{\mathrm{MSE}} \right) \qquad (2)$$

where $\mathrm{MAX}_I = 2^B - 1$ is the maximum possible pixel value of an image with bit depth B (255 for an 8-bit grayscale image and 65535 for a 16-bit grayscale

image), and MSE is the mean squared error between the original and the processed image defined in Eq. (1). PSNR is expressed in decibels (dB), a logarithmic scale that emphasizes small differences in noise levels, and it is undefined (infinite) when $\mathrm{MSE} = 0$. As a rough guide, a PSNR above 40 dB indicates a high-quality image with little noise or distortion, values between 30 and 40 dB indicate minor but acceptable distortion, values between 20 and 30 dB indicate noticeable artifacts, and values below 20 dB indicate significant degradation.

A limitation of PSNR is that it does not account for human visual perception: two images may have the same PSNR but appear different to the human eye. It works best for small, localized distortions but struggles with large structural differences. A complementary metric such as the Structural Similarity Index (SSIM) is therefore often used alongside PSNR [76]. The interpretation of PSNR values is presented in Table 3.

Table 3. Interpretation of Peak Signal-to-Noise Ratio (PSNR) values for image quality assessment.

PSNR (dB)	Image Quality	Perceived Distortion
> 40	Excellent	Almost no visible distortion
30-40	Good	Minor distortion, but acceptable
20-30	Moderate	Noticeable artifacts and noise
< 20	Poor	Significant degradation

4.3 Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) is a widely used evaluation metric in digital image processing that measures the difference between the original and processed images. It provides a numerical estimate of the distortion introduced during image processing tasks such as compression, denoising, enhancement, and super-resolution [44, 77, 78].

RMSE is the square root of the MSE and quantifies how much the processed image deviates from the original. A lower RMSE value indicates better image quality, whereas a higher RMSE suggests significant image degradation [77, 78].

Its uses parallel those of MSE. RMSE is reported for the quality loss in compressed images (JPEG, HEVC, etc.) [42, 45], for the error remaining after denoising and sharpening [78, 79], and for the agreement between super-resolved images and the ground truth [45]; it is also used in evaluating deep learning models for image processing, and RMSE (or MSE) often serves as their training loss in image synthesis, enhancement, and denoising [44, 78].

RMSE is more interpretable than MSE because it is expressed in the same unit as the original pixel intensity values. Unlike MAE, RMSE gives more weight to large errors, making it useful for applications where small image artifacts are acceptable but large distortions must be minimized [77]. RMSE is derived from the MSE in Eq. (1) and is given by:

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} = \sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\big(I(i,j) - K(i,j)\big)^2} \tag{3}$$

RMSE is always non-negative, because it is the square root of an average of squared differences. A value of zero means that the processed image is identical to the original, and increasingly large values indicate increasingly strong deviation from it.

A limitation of RMSE is that it does not reflect perceptual quality: two images with the same RMSE might look different to the human eye. Its sensitivity to large errors is also not always the best indicator of perceived image quality, so SSIM is often reported alongside it. Table 4 presents how to interpret RMSE values.

Table 4. Interpretation of Root Mean Squared Error (RMSE) values for image quality assessment.

RMSE Value	Image Quality	Perceived Error
< 1	Excellent	Almost no visible error
1-5	Good	Small distortion, acceptable quality
5-10	Moderate	Noticeable artifacts present
> 10	Poor	Significant distortion

4.4 Signal-to-Noise Ratio (SNR)

Signal-to-Noise Ratio (SNR) is a fundamental metric in digital image processing used to quantify the level of the desired signal relative to the background noise in an image. A higher SNR indicates a cleaner image with less noise, while a lower SNR suggests a degraded or noisy image [80–82]. It is used in image acquisition, medical imaging, remote sensing, and the evaluation of denoising techniques [83, 84].

In medical imaging (MRI, CT, ultrasound, X-ray), SNR is routinely reported as a baseline indicator of image quality and system performance, since diagnostic accuracy depends on it [80, 85]. In satellite and airborne remote sensing it characterizes the quality of images acquired under varying atmospheric conditions [86]. It is equally relevant to low-light and otherwise noisy acquisitions, where it is used both to guide enhancement and to quantify the detectability of defects and features [87–89]. Finally, SNR provides a quantitative basis for comparing denoising methods, from Gaussian and wavelet filtering to self-supervised deep denoisers [84, 90]. SNR is defined as the ratio of signal power to noise power:

$$\mathrm{SNR} = \frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}} \tag{4}$$

Since images are represented as matrices of pixel intensities, SNR is often calculated in decibels as:

$$\mathrm{SNR} = 10 \log_{10} \left(\frac{\sum_{i=1}^{M} \sum_{j=1}^{N} I(i,j)^2}{\sum_{i=1}^{M} \sum_{j=1}^{N} \big(I(i,j) - K(i,j)\big)^2} \right) \tag{5}$$

where the numerator represents the total power of the signal (the original image I), the denominator represents the total power of the noise (the difference between the original image and the noisy or distorted image K, i.e., $MN \cdot \mathrm{MSE}$), and the base-10 logarithm converts the ratio into decibels (dB) for easier interpretation. A high SNR (>40 dB) means that the signal is strong and the noise is minimal, whereas a low SNR (<20 dB) means that the noise is significant, leading to poor image quality.

Beyond the evaluation of processed images, SNR guides the design of camera sensors and of medical and remote-sensing acquisition systems, and it is decisive in low-light imaging, where night-vision cameras and astronomical telescopes depend on a high SNR.

SNR has two main limitations: (i) it does not account for structural distortions, so an image might have high SNR but still look visually incorrect; and (ii) perceptual quality is not always correlated with SNR, which is why metrics such as SSIM are often used alongside it. Table 5 presents how to interpret SNR values.

Table 5. Interpretation of Signal-to-Noise Ratio (SNR) values for image quality assessment.

SNR (dB)	Image Quality	Noise Level
> 40	Excellent	Minimal noise
30-40	Good	Slight noise
20-30	Moderate	Noticeable noise
10-20	Poor	High noise
< 10	Very poor	Very high noise and poor clarity

4.5 Feature Similarity Index Measure (FSIM)

The Feature Similarity Index Measure (FSIM) is an advanced image quality assessment metric used in digital image processing. Unlike traditional metrics such as MSE and PSNR, FSIM evaluates image quality based on low-level features extracted from the image [43, 44, 91].

FSIM is motivated by human visual perception and relies on two features: (i) phase congruency (PC), which measures the most perceptually significant features in an image (edges, textures) and is largely invariant to contrast and illumination changes [42, 91]; and (ii) gradient magnitude (GM), which captures contrast and texture details and is used as a complementary feature to PC [43, 91].

This design reflects the way the human visual system relies on low-level features rather than raw pixel errors [43, 92]. It is effective for complex images such as natural scenes and textures, for which it correlates strongly with subjective quality scores [93], and on benchmark IQA databases it agrees with human judgments better than PSNR and MSE do [91, 93].

FSIM is applied wherever perceptual fidelity matters. In image compression (JPEG and other codecs) it is used to characterize quality, including the visually lossless range in which FSIM remains very close to 1 [43, 93]. Denoising filters, whether Gaussian, wavelet, non-local means, or CNN-based and hybrid, are commonly evaluated with FSIM alongside PSNR and SSIM [42, 94], and enhancement methods such as histogram equalization, super-resolution, and contrast enhancement are compared in terms of how well they preserve image features [42]. In medical imaging (e.g., MRI, mammography), FSIM is used to measure restoration quality and has been compared with radiologists' assessments of diagnostic quality [44, 94]. The FSIM metric is calculated as follows:

$$\mathrm{FSIM} = \frac{\sum_{x\in\Omega} S_L(x)\, PC_m(x)}{\sum_{x\in\Omega} PC_m(x)} \tag{6}$$

where Ω is the entire image domain, x is a pixel location, $PC_m(x) = \max\big(PC_1(x), PC_2(x)\big)$ is the larger of the phase congruency values of the reference image (PC_1) and the processed image (PC_2) at x, which weights perceptually significant locations, and $S_L(x)$ is the local similarity map, computed from the phase congruency similarity $S_{PC}(x)$ and the gradient magnitude similarity $S_G(x)$ as:

$$S_L(x) = S_{PC}(x)\, S_G(x) \tag{7}$$

with

$$\begin{aligned} S_{PC}(x) &= \frac{2\, PC_1(x)\, PC_2(x) + T_1}{PC_1^2(x) + PC_2^2(x) + T_1}, \\ S_G(x) &= \frac{2\, G_1(x)\, G_2(x) + T_2}{G_1^2(x) + G_2^2(x) + T_2} \end{aligned} \tag{8}$$

where G_1 and G_2 are the gradient magnitudes of the two images and T_1, T_2 are small positive constants that ensure numerical stability [91]. The numerator of Eq. (6) weights the local similarity by perceptual significance, and the denominator normalizes the measure. FSIM values lie in the interval $(0, 1]$: $\mathrm{FSIM} = 1$ means that the images are identical, and lower values indicate more distortion.

FSIM is sensitive to structural distortions, which makes it suitable for denoising and image restoration, and it is effective for texture-rich images such as medical images, satellite images, and artistic photographs. Its key limitations are that (i) it is computationally more expensive than pixel-wise metrics, and (ii) it may not work well on extremely low-resolution images, where feature extraction is challenging. Table 6 presents the interpretation of FSIM values.

Table 6. Interpretation of Feature Similarity Index Measure (FSIM) values for perceptual image quality assessment.

FSIM Score	Perceived Image Quality	Example Scenario
0.95-1.00	Excellent (almost identical)	Visually lossless or near-lossless compression
0.85-0.95	Good	High-quality lossy compression
0.70-0.85	Moderate	Slightly blurred or compressed
0.50-0.70	Poor	Noticeable noise and artifacts
< 0.50	Very poor	Strong degradation, artifacts, and high noise

4.6 Mean Absolute Error (MAE)

The Mean Absolute Error (MAE) is a widely used evaluation metric that measures the average absolute difference between the corresponding pixel intensities of two images [77]. It is used to assess the accuracy of image reconstruction, compression, enhancement, and restoration algorithms by quantifying the mean magnitude of the pixel-wise error in the same units as the image data [95]. Unlike MSE, which squares the errors, MAE directly considers the absolute differences, making it easier to interpret as the "average error" and preserving the original scale [79]. This also makes it more robust to outliers than MSE, which overemphasizes large errors due to squaring [95, 96].

MAE is often used as a full-reference discrepancy measure between an original and a distorted image, with lower values indicating better quality and more accurate reconstruction [79]. Because it is less sensitive to extreme errors than MSE or RMSE, it is preferred as a measure of average error in many contexts [95]. MAE is used in image compression, denoising, super-resolution, medical image analysis, and other image processing tasks. It is given by:

$$\mathrm{MAE} = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\big|I(i,j) - K(i,j)\big| \tag{9}$$

where $\big|I(i,j) - K(i,j)\big|$ is the absolute difference between corresponding pixel intensities. $\mathrm{MAE} = 0$ means that the two images are identical, and a higher MAE indicates greater distortion or image degradation.

The limitations of MAE are that (i) it does not consider perceptual image quality (unlike SSIM or FSIM), and (ii) it weights all errors linearly and therefore does not emphasize large errors. Table 7 shows how to interpret MAE values.

4.6.1 *Comparison of MAE, MSE, and RMSE in Digital Image Processing*

In digital image processing, selecting an appropriate error metric is crucial for evaluating image quality. MAE, MSE, and RMSE are three commonly used

Table 7. Interpretation of Mean Absolute Error (MAE) values for image quality assessment.

MAE Score	Perceived Image Quality	Example Scenario
0-5	Excellent (almost identical)	Near-lossless compression
5-15	Good	High-quality lossy compression
15-30	Moderate	Slight blur or compression artifacts
30-50	Poor	Visible noise and artifacts
> 50	Very poor	Strong degradation and significant loss of detail

pixel-wise error metrics that measure the difference between an original and a processed image. This section compares them in terms of their properties, interpretation, advantages, and limitations.

As a simplified illustration, consider an 8×8 grayscale image and its blurred version for which $\mathrm{MAE} = 10$, $\mathrm{MSE} = 150$, and $\mathrm{RMSE} = \sqrt{150} \approx 12.25$. The MAE states that the average pixel intensity difference is 10; the MSE is much larger because squaring increases the emphasis on large errors; and the RMSE returns to the intensity scale but remains higher than the MAE because of the squared averaging. The advantages and limitations of each metric are presented in Table 8.

Table 8. Comparative analysis of MAE, MSE, and RMSE: advantages and limitations.

Metric	Advantages	Limitations
MAE	Easy to interpret; less sensitive to outliers; linear error growth	Does not penalize large errors strongly; less useful in applications where large distortions matter
MSE	Penalizes large errors more strongly (useful for critical applications); differentiable and convenient as a machine learning loss function	Harder to interpret (squared units); sensitive to outliers
RMSE	Provides the error in the same units as the image; balances sensitivity and interpretability	As sensitive to outliers as MSE, although expressed on the original intensity scale

Table 9 briefly summarizes the scenarios in which each error metric is typically used.

Table 10 shows which error metric should be considered in different scenarios.

In short, MAE is the natural choice when outliers should not dominate the evaluation; MSE is preferable when large errors must be penalized heavily; and RMSE offers a compromise, since it restores the original intensity scale while still penalizing large errors.

Table 9. Recommended error metrics for different image processing use cases.

Use Case	Suggested Metric	Reason
General image quality assessment	MAE	Easier to interpret and not dominated by extreme values.
High-priority error sensitivity (e.g., medical imaging)	MSE	Strongly penalizes large errors, helping to ensure that critical details are preserved.
Machine learning (loss function for image restoration)	MSE or RMSE	Smooth optimization behavior, useful for deep learning models.
Natural image processing (error reported in intensity units while retaining sensitivity to large errors)	RMSE	Combines the sensitivity of MSE with the interpretability of the original scale.

Table 10. Selection of MAE, MSE, and RMSE based on evaluation requirements.

Metric	Best For	Avoid When
MAE	Image compression and applications with low sensitivity to large errors	Large errors must be emphasized
MSE	High-risk applications (medical imaging, critical reconstructions)	Interpretability is key
RMSE	Balancing sensitivity and interpretability (e.g., image denoising)	Emphasis on large errors should be reduced

4.7 Normalized Absolute Error (NAE)

Normalized Absolute Error (NAE) is an evaluation metric used in digital image processing to measure the difference between an original image and a processed image (e.g., compressed, restored, or denoised images) [97]. Unlike MAE, which gives an absolute error value, NAE normalizes the error by dividing it by the total pixel intensity of the original image, making it a normalized, scale-adjusted variant of the absolute error [97]. This normalization allows a fairer comparison across images of different intensities and resolutions and across varying image types and noise conditions [97, 98]. NAE is calculated as:

$$\mathrm{NAE} = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} \left| I(i,j) - K(i,j) \right|}{\sum_{i=1}^{M} \sum_{j=1}^{N} \left| I(i,j) \right|} \tag{10}$$

where the numerator is the total absolute error between the two images and the denominator is the total pixel intensity of the original image (the normalization factor).

NAE is non-negative: $\mathrm{NAE} = 0$ means that the

images are identical, and a higher NAE indicates a greater difference between the original and processed images. In practice, values usually fall between 0 and 1, although NAE can exceed 1 when the total absolute error is larger than the total intensity of the original image. The metric is particularly useful when comparing images of varying brightness or contrast levels, as it accounts for variations in intensity values.

NAE measures the overall pixel-wise deviation between two images relative to the total intensity of the reference image; it does not capture structural information. Unlike MSE, which emphasizes larger errors, it weights all errors linearly and therefore gives a balanced account of overall differences.

Table 11 presents the interpretation of NAE values.

Table 11. Interpretation of NAE values and corresponding image quality.

NAE Value	Interpretation	Quality of Processed Image
0.00-0.05	Very small error	Excellent quality (almost identical images)
0.05-0.15	Small error	Good quality (slight differences visible)
0.15-0.30	Moderate error	Noticeable degradation
0.30-0.50	High error	Significant degradation
> 0.50	Very high error	Poor quality, major distortion

4.8 Dice Coefficient

The Dice coefficient, or Dice Similarity Coefficient (DSC), is a widely used evaluation metric in image segmentation that measures the similarity between two binary images (such as a ground truth segmentation mask and a predicted mask), defined as twice the overlap divided by the sum of the two regions [99, 100]. It is a measure of spatial overlap between segmentation masks and is particularly prominent in medical imaging tasks where agreement between automated and manual segmentations must be quantified [101]. DSC values range from 0 (no overlap) to 1 (perfect overlap, i.e., identical segmentations), making it an intuitive indicator of segmentation quality [100].

Unlike pixel-wise metrics such as MSE, the Dice coefficient focuses on spatial consistency rather than pixel-wise intensity differences. It is given by:

$$\mathrm{DSC} = \frac{2\,|A \cap B|}{|A| + |B|} \tag{11}$$

where A is the ground truth segmentation mask (the actual region of interest), B is the predicted segmentation mask (the model output), $|A \cap B|$ is the number of overlapping pixels (true positives), and $|A|$ and $|B|$ are the total numbers of pixels in the ground truth mask and the predicted mask, respectively. The DSC is symmetric, i.e., $\mathrm{DSC}(A, B) = \mathrm{DSC}(B, A)$, and it ignores true-negative (background) pixels, which makes it far less affected by a dominant background class than pixel accuracy. The closely related Intersection over Union (IoU), or Jaccard index, is defined as:

$$\mathrm{IoU} = \frac{|A \cap B|}{|A \cup B|} = \frac{\mathrm{DSC}}{2 - \mathrm{DSC}} \tag{12}$$

so that the two measures are monotonically related and rank segmentations identically, with $\mathrm{IoU} \leq \mathrm{DSC}$ [101].

The DSC measures the overlap between predicted and actual segmentation regions and is preferred over simple pixel-wise accuracy because it better reflects size and localization agreement, especially under strong class imbalance where background pixels dominate (e.g., small tumors within large scans) [101, 102]. In such cases, pixel accuracy can be misleadingly high even with poor lesion segmentation, whereas the DSC remains sensitive to foreground performance [99]. The DSC is therefore widely used in medical image segmentation benchmarks and as a loss function or target metric in deep learning models where both precision and recall are important [103].

The DSC is used well beyond its origins. In medical imaging (MRI, CT, X-ray) it is the standard measure for evaluating tumor detection, organ and liver/tumor segmentation, and the delineation of white-matter lesions or brain tumors [99, 100, 102]. In autonomous driving, the segmentation of roads, lanes, and objects in urban scenes is evaluated with Jaccard- or Dice-type overlap scores [104], and in remote sensing the same measures are applied to building and land-cover segmentation from high-resolution satellite or aerial images [105]. More generally, Dice-type similarity measures quantify the overlap between detected and reference regions whenever segmentation algorithms are compared [28, 50].

Table 12 presents the interpretation of DSC values.

Table 13 presents the comparison of the DSC with MSE, PSNR, and SSIM.

If the focus is on segmentation accuracy rather than pixel intensity differences, the Dice coefficient is the

Table 12. Interpretation of Dice Similarity Coefficient (DSC) values for image segmentation.

DSC Value	Interpretation	Quality of Segmentation
0.90-1.00	Excellent overlap	Highly accurate segmentation
0.75-0.90	Good overlap	Reliable segmentation with minor errors
0.50-0.75	Moderate overlap	Some segmentation errors present
0.25-0.50	Poor overlap	Significant segmentation errors
0.00-0.25	No or minimal overlap	Very poor segmentation

Table 13. Comparison of pixel-based, signal-based, structural, and region-overlap evaluation metrics.

Metric	Focus Area	Sensitivity to Imbalance?	Best Used For
MSE	Pixel-wise difference	Yes	Image denoising, restoration
PSNR	Signal-to-noise ratio	Yes	Compression quality
SSIM	Structural similarity	No	Perceptual image quality
Dice coefficient	Overlapping regions	No	Image segmentation

more appropriate metric: it provides an overlap-based similarity measure that reflects how well a segmented region matches the ground truth, and it should be one of the primary evaluation metrics for segmentation tasks. The DSC is mathematically equivalent to the F1-score computed at the pixel level:

$$\mathrm{DSC} = F_1 = \frac{2\,TP}{2\,TP + FP + FN} \tag{13}$$

where TP (true positives) is the number of pixels correctly classified as part of the segmented object, FP (false positives) is the number of pixels incorrectly included in the segmented object, and FN (false negatives) is the number of pixels of the object that were missed. The two names differ only in usage: the term Dice coefficient is primarily used in image segmentation (pixel-based evaluation), whereas the term F1-score is used in classification problems (sample-based evaluation). The designation "Dice coefficient (F1-score for segmentation)" is therefore valid and helps readers with a machine learning background.

4.9 Precision and Recall

Precision and recall are fundamental evaluation metrics used in image classification, object detection, and image segmentation tasks. They help to assess the performance of models by measuring how well relevant pixels or objects are identified [106].

Precision measures how many of the detected objects or segmented pixels are actually relevant, while recall measures how many of the actual relevant objects or pixels were detected by the model [107, 108].

In object detection (e.g., face recognition, license plate recognition), performance is commonly summarized by precision-recall (PR) curves and Average Precision, which capture the trade-off between missing objects and generating false alarms [57]. In medical image segmentation, such as tumor or lesion detection, the two measures separate missed lesions (false negatives) from spurious detections (false positives) [51]. Edge detectors are compared in the same way: precision expresses how many detected edges are true boundaries, and recall how many true boundaries are found [109].

Precision and recall are calculated as:

$$\mathrm{Precision} = \frac{TP}{TP + FP} \tag{14}$$

$$\mathrm{Recall} = \frac{TP}{TP + FN} \tag{15}$$

where TP (true positives) is the number of pixels or objects correctly identified as belonging to the target class, FP (false positives) is the number of pixels or objects incorrectly identified as belonging to the target class, and FN (false negatives) is the number of pixels or objects that belong to the target class but were missed by the model. A high precision means that when the model predicts a positive, it is likely correct (few false positives); a high recall means that few actual positives are missed (few false negatives). The F1-score combines precision and recall as their harmonic mean:

$$F_1 = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{16}$$

The F1-score is useful when false positives and false negatives are both important, providing a balanced measure of precision and recall.

The relative weight given to the two measures depends on the cost of each type of error. Precision is decisive when false positives are costly, for example in security and surveillance systems, where benign objects must

Table 14. Interpretation of precision-recall scenarios in image processing.

Scenario	Precision	Recall	Interpretation
Perfect model	100%	100%	No false positives or false negatives
High precision, low recall	90%	50%	The model makes very few false detections but misses many actual objects (strict model)
Low precision, high recall	50%	90%	The model detects most objects but makes many incorrect detections (lenient model)
Low precision, low recall	50%	50%	Poor detection, with both many false positives and many false negatives

not be labeled as threats [108]. Recall matters most when false negatives are unacceptable, as in medical imaging, where a missed tumor or lesion can be critical, or in remote-sensing segmentation, where important land-cover objects should not be overlooked [107]. For this reason, many evaluation frameworks analyze the precision-recall trade-off explicitly, through PR curves, AP, or derived measures such as Localization Recall Precision, in order to choose an operating point that matches the tolerance of the application for either kind of error [107, 108].

These metrics are particularly useful in segmentation and detection, where identifying the relevant pixels or objects is more important than overall accuracy. A model with high precision but low recall is selective: it makes few false detections but may miss many relevant objects. A model with high recall but low precision captures most relevant objects at the price of many incorrect predictions. The ideal model achieves both, which corresponds to the top-right corner of a precision-recall plot; if precision drops sharply as recall increases, the model is producing too many false positives. Table 14 describes different possible scenarios.

When precision and recall are equally important and a balanced metric is needed, the F1-score in Eq. (16) should be used. In all cases, the two measures should be analyzed together, and a precision-recall curve shows how a model behaves at different decision thresholds.

4.10 Confusion Matrix

The confusion matrix is a fundamental tool for evaluating classification models in image processing tasks such as image classification, object detection, and image segmentation, because it tabulates true positives, false positives, true negatives, and false negatives in a single table [110]. In binary and multi-class problems, rows typically represent actual classes and columns represent predicted classes, allowing clear visualization of which classes are being confused with which others [111, 112].

In image classification (e.g., cat vs. dog), the confusion matrix summarizes per-class performance and supports the computation of accuracy, precision, recall, and F1-score [110]. In detection and recognition problems, such as glaucoma detection from fundus images, it helps to analyze the false positives and false negatives of a detector [113], and in segmentation and feature-detection tasks, such as corner detection, it counts correctly and incorrectly detected structures, from which sensitivity, specificity, and related measures are derived [112].

The value of the confusion matrix lies in the fact that it shows how many samples fall into each actual-predicted combination and thus reveals which classes are most often confused [110]. Its four entries (TP, FP, TN, FN) can be read off directly and distinguish the types of error, which is crucial in medical diagnosis or quality inspection, where false positives and false negatives have very different costs [110, 112, 114]. Moreover, the standard performance measures (accuracy, precision, recall or sensitivity, F1-score, specificity, and the Matthews correlation coefficient, MCC) are all defined as functions of these entries, so the matrix is the primary source from which they are derived [110, 111]. The binary form of the confusion matrix is shown in Table 15.

Table 15. Confusion matrix for binary image classification.

Actual \ Predicted	Positive (1)	Negative (0)
Positive (1)	True positive (TP)	False negative (FN)
Negative (0)	False positive (FP)	True negative (TN)

Here, a true positive (TP) is a positive sample correctly classified as positive; a false positive (FP) is a negative

sample incorrectly classified as positive (Type I error); a false negative (FN) is a positive sample incorrectly classified as negative (Type II error); and a true negative (TN) is a negative sample correctly classified as negative.

The confusion matrix serves as the foundation for calculating multiple evaluation metrics. The specificity, or true negative rate (TNR), measures how well the model identifies negatives and is useful in medical diagnosis (e.g., screening for diseases):

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{17}$$

The false positive rate (FPR) and the false negative rate (FNR) are defined as:

$$\begin{aligned} \text{FPR} &= \frac{FP}{FP + TN} = 1 - \text{Specificity}, \\ \text{FNR} &= \frac{FN}{TP + FN} = 1 - \text{Recall} \end{aligned} \tag{18}$$

The FPR is the probability of incorrectly classifying a negative as positive, and the FNR is the probability of missing an actual positive. The FPR, together with the true positive rate, is used in ROC curves for threshold-based analysis.

The confusion matrix is the basis for evaluating models in image processing, and visualizing it as a heatmap helps in interpretation. Because accuracy alone is not sufficient, precision, recall, and the F1-score should be used for a complete evaluation.

4.11 Accuracy and F1-Score

Accuracy is a fundamental evaluation metric in digital image processing for measuring the correctness of a classification model. It represents the proportion of correctly classified pixels, objects, or images relative to the total number of samples [115]. While accuracy is a simple and intuitive metric, it may be misleading in cases of imbalanced datasets, where one class dominates the dataset [51].

The F1-score is the harmonic mean of precision and recall, providing a more balanced measure of a model's performance, especially when dealing with imbalanced classes [116, 117]. It helps in scenarios such as medical image segmentation, object detection, and facial recognition, where both false positives and false negatives are critical [51]. Accuracy is computed from the entries of the confusion matrix in Table 15 as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{19}$$

The F1-score, precision, and recall are defined in Eqs. (16), (14), and (15), respectively.

Accuracy remains popular because it is simple and intuitive [115], because it is informative when the class distribution of an image dataset is balanced [116], and because it is the conventional measure in image recognition and classification [22, 115].

The F1-score, in contrast, is more informative for imbalanced datasets, for instance when rare diseases are to be detected in medical images [116]. It is also valuable in segmentation, where missing a small but important part of an object can change the result substantially [51], and in applications such as biometric security, where false positives and false negatives must be balanced [22, 51]. Table 16 shows when to use accuracy and the F1-score.

Taken together, accuracy is appropriate for balanced datasets but can remain high even when a model is biased towards the majority class, whereas the F1-score exposes the limitations of a model when false negatives and false positives both matter. In medical imaging, biometric security, and segmentation, the F1-score is therefore generally preferred.

Accuracy is nevertheless frequently reported as the sole metric in the literature. Plausible reasons include its simplicity and interpretability, the use of balanced benchmark datasets in many studies, the convenience of benchmarking against earlier accuracy-based results, and its computational simplicity. In image processing research, both accuracy and the F1-score should be reported to provide a more complete evaluation.

4.12 Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC)

The Receiver Operating Characteristic (ROC) curve is a graphical representation of a classification model's ability to distinguish between classes. It is particularly useful in digital image processing applications such as image segmentation, object detection, and medical imaging classification [118]. The ROC curve plots the true positive rate (TPR, also called sensitivity or recall) against the false positive rate (FPR) at various classification thresholds [118–120].

The Area Under the Curve (AUC) quantifies the overall ability of the model to discriminate between classes [118–120]: $\text{AUC} = 1$ corresponds to perfect

Table 16. Comparison of accuracy and F1-score under different evaluation scenarios.

Scenario	Use Accuracy	Use F1-Score
Balanced dataset (equal positives and negatives)	Yes	Optional
Imbalanced dataset (rare class detection)	Not alone	Yes
Image classification with multiple classes	Yes	Yes
Medical image analysis (e.g., cancer detection)	Not alone	Yes

classification, $\mathrm{AUC} = 0.5$ means that the model performs no better than random guessing, and $\mathrm{AUC} < 0.5$ means that the model performs worse than random guessing. AUC-ROC is widely used in medical image processing (e.g., tumor detection), biometric recognition, and other classification-based tasks.

The two quantities needed to plot the ROC curve are the TPR,

$$\mathrm{TPR} = \frac{TP}{TP + FN} \tag{20}$$

which measures how well the model detects positives and is identical to the recall in Eq. (15), and the FPR defined in Eq. (18), which measures how often the model incorrectly classifies negatives as positives. The AUC is computed as the integral of the ROC curve:

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR}\, \mathrm{d}(\mathrm{FPR}) \tag{21}$$

In practice, the ROC curve is available only at a finite number of thresholds, so the integral is evaluated numerically, for example with the trapezoidal rule.

ROC analysis has three main attractions. First, it is threshold-independent: whereas accuracy and precision depend on a chosen classification threshold, the ROC curve describes performance across all thresholds [119]. Second, the ROC curve is insensitive to class prevalence, unlike accuracy, which can be misleading for imbalanced datasets [119, 121]; for highly skewed data, however, precision-recall curves are often considered more informative [122]. Third, in tasks such as cancer detection or face recognition, ROC-AUC makes the balance between false alarms and missed detections explicit [118].

AUC-ROC applies most directly to binary problems, such as distinguishing tumor from non-tumor images [118]. It extends to multiclass classification by computing one-vs-all ROC curves for each class [119, 121], and it is also used for target and anomaly detection in imagery, for example in hyperspectral target detection [120].

An AUC above 0.9 indicates a strong model with a high capability to differentiate the two classes (e.g., tumors vs. non-tumors). In summary, the ROC curve evaluates model performance across all classification thresholds, and the AUC provides a single score to compare classifiers. Both should be used alongside precision, recall, and the F1-score for a complete evaluation.

5 Conclusions and Future Work

Evaluation is a fundamental component of digital image processing and machine learning because the validity of a model or image-processing method depends not only on its development but also on how its performance is measured. This review examined commonly used evaluation metrics across five major task categories: image quality assessment, image classification, image segmentation, image restoration, and object detection. The findings show that metric selection is inherently task-dependent, and no single metric can adequately capture all dimensions of image-processing performance. For image quality and restoration, pixel-wise measures such as MSE, RMSE, MAE, and PSNR provide useful information about numerical fidelity and reconstruction error but may not fully reflect structural or perceptual similarity. Metrics such as SSIM, MS-SSIM, FSIM, VIF, and LPIPS therefore provide complementary information. Similarly, classification should not rely on accuracy alone, particularly for imbalanced datasets; precision, recall, F1-score, confusion matrices, and ROC-AUC provide additional insight into classification performance. For segmentation, Dice and IoU measure agreement between predicted and reference regions, while object detection commonly relies on AP and mAP to evaluate localization and classification performance. Thus, metric selection should be guided by the characteristics and objectives of the specific task. An important implication of this review is that evaluation metrics should be selected according to the evaluation objective rather than their popularity or convenience. Pixel-level fidelity, structural similarity, perceptual quality, classification discrimination, localization accuracy, robustness, and class imbalance

represent different performance dimensions and may require complementary measures. The limitations of conventional metrics such as MSE, RMSE, and PSNR further support combining numerical and perceptual measures when appropriate.

Future research should focus on standardized, reproducible, and task-aware evaluation protocols, with stronger links between objective metrics and human perceptual judgments and greater consistency across datasets and implementations. Such frameworks will be increasingly important for emerging areas including multimodal image processing, foundation models, generative image restoration, three-dimensional imaging, and vision-language systems. Therefore, metric selection should be treated as an integral part of experimental design rather than simply a reporting step. A well-designed evaluation strategy should use interpretable, task-appropriate metrics capable of capturing the relevant dimensions of performance.

Data Availability Statement

Not applicable.

Funding

This work was supported without any funding.

Conflicts of Interest

Muzammil Khan served as an Editor-in-Chief of the *ICCK Journal of Image Analysis and Processing* at the time of manuscript submission. To ensure the integrity of the peer-review process, Muzammil Khan was not involved in the editorial handling, peer review, or decision-making process for this manuscript, which was handled independently by another editor. Zaheer Khan is affiliated with Spinghar Technical and Vocational Training Center, Jalalabad 2601, Afghanistan. The authors declare that this affiliation had no influence on the study design, data collection, analysis, interpretation, or the decision to publish. The remaining authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Trigka, M., & Dritsas, E. (2025). A comprehensive survey of deep learning approaches in image processing. *Sensors, 25*(2), 531. [CrossRef]

[2] Valente, J., António, J., Mora, C., & Jardim, S. (2023). Developments in image processing using deep learning and reinforcement learning. *Journal of imaging, 9*(10), 207. [CrossRef]

[3] Archana, R., & Jeevaraj, P. S. E. (2024). Deep learning models for digital image processing: A review. *Artificial Intelligence Review, 57*(1), 11. [CrossRef]

[4] Gomathi, S., & Roopa Chandrika, R. (2025). Advancing medical image processing with deep learning: Innovations and impact. *ICTACT Journal on Image and Video Processing, 15*(3), 3489-3494. [CrossRef]

[5] Alnaggar, O. A. M. F., Jagadale, B. N., Saif, M. A. N., Ghaleb, O. A., Ahmed, A. A., Aqlan, H. A. A., & Al-Ariki, H. D. E. (2024). Efficient artificial intelligence approaches for medical image processing in healthcare: comprehensive review, taxonomy, and analysis. *Artificial Intelligence Review, 57*(8), 221. [CrossRef]

[6] Gorai, S. K., Sarangi, A., & Pradhan, S. (2025). Deep learning for image classification: Methods, challenges, and future directions. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 11*, 484-496. [CrossRef]

[7] Saravanan, V., Adaikkammai, A., Sasikala, G., Iyswariya, A., Babisha, A., & Latha, M. (2024). Advancements in artificial intelligence for better image analysis, recognition, and interpretation in the field of image processing. In *2024 Global Conference on Communications and Information Technologies (GCCIT)* (pp. 1-6). IEEE. [CrossRef]

[8] Jiao, L., & Zhao, J. (2019). A survey on the new generation of deep learning in image processing. *IEEE Access, 7*, 172231-172263. [CrossRef]

[9] Zhu, X. X., Tuia, D., Mou, L., Xia, G. S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine, 5*(4), 8-36. [CrossRef]

[10] Hsieh, W., Bi, Z., Liu, J., Peng, B., Zhang, S., Pan, X., ... & Liu, M. (2024). Deep learning, machine learning - Digital signal and image processing: From theory to application. *arXiv preprint arXiv:2410.20304*. [CrossRef]

[11] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of big data, 6*(1), 60. [CrossRef]

[12] Varga, D. (2025). AI-driven digital image processing: Advancements, challenges, and future directions. *Electronics, 14*(9), 1754. [CrossRef]

[13] Chen, J. (2025). Research on image recognition model

based on deep learning. *Applied and Computational Engineering, 177*, 159-165. [CrossRef]

[14] Wanjari, A., & Verma, S. (2025). A review of machine learning and deep learning-based approaches for image recognition and its applications. In *2025 4th International Conference on Sentiment Analysis and Deep Learning (ICSADL)* (pp. 459-464). IEEE. [CrossRef]

[15] Gui, X., Wang, W., & Tao, D. (2024). A survey of self-supervised learning from multiple perspectives: Algorithms, theory, applications and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 47*(1), 248-267. [CrossRef]

[16] Eapen, B. T., & George, S. T. (2024). Exploring self-supervised learning architectures for image processing. *International Journal on Recent and Innovation Trends in Computing and Communication, 12*(2), 358-372. [CrossRef]

[17] Wang, Y., Lv, F., Li, S., Liu, X., & Peng, H. (2025). A comparative survey of vision transformers for image classification. *Neurocomputing, 629*, 129752. [CrossRef]

[18] Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. *ACM Computing Surveys, 54*(10s), 1-41. [CrossRef]

[19] Saeed, W., & Omlin, C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems, 263*, 110273. [CrossRef]

[20] Akhtar, N. (2023). Revolutionizing digital image processing: Advancements and applications of deep visual models explainability. *arXiv preprint arXiv:2301.13445*. [CrossRef]

[21] Salmanpour, M. R., Shamsaei, M., Hajianfar, G., Askari, A., Maghsudi, M., Bloise, I., ... & Rahmim, A. (2025). Machine learning metrics: Assessing performance differences across modalities and tasks in medical imaging. *Physica Medica, 130*, 104890. [CrossRef]

[22] Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific reports, 14*(1), 6086. [CrossRef]

[23] Zhai, G., & Min, X. (2020). Perceptual image quality assessment: a survey. *Science China Information Sciences, 63*(11), 211301. [CrossRef]

[24] Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing, 13*(4), 600-612. [CrossRef]

[25] Urrea, C., & Vélez, J. M. (2025). Advances in semantic segmentation using deep learning techniques: A systematic review. *IEEE Access, 13*, 4501-4523. [CrossRef]

[26] Cho, J. (2024). Weighted Intersection over Union (wIoU): A metric for evaluating image segmentation. *IEEE Access, 12*, 116030-116045. [CrossRef]

[27] Taciuc, I. A., Dandu, A. R., Badiger, R., Kalyana Sundaram, M. N., & Marinas, M. C. (2025). Enhancing AI-driven detection of lymph nodes in ultrasound imaging: Comparative analysis of DICE and IoU metrics. *Sensors, 25*(9), 2727. [CrossRef]

[28] Merkulova, N., & Jayakumar, A. (2025). Evaluation framework for image segmentation. *arXiv preprint arXiv:2504.04435*. [CrossRef]

[29] Chin, S. P., Yap, W. S., Tee, C. A. T. H., & Kim, Y. J. (2025). Challenges and evaluation metrics in medical image enhancement. *arXiv preprint arXiv:2510.13638*. [CrossRef]

[30] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision, 115*, 211-252. [CrossRef]

[31] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13* (pp. 740-755). Springer International Publishing. [CrossRef]

[32] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9650-9660). IEEE. [CrossRef]

[33] Shao, R., Shi, Z., Du, J., Tang, X., & Qiao, Y. (2022). On the open-world classification challenge in image classification. *arXiv preprint arXiv:2202.00179*. [CrossRef]

[34] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618-626). IEEE. [CrossRef]

[35] Dong, Y., Fu, Q. A., Yang, X., Pang, T., Su, H., Xiao, Z., & Zhu, J. (2020). Benchmarking adversarial robustness on image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 321-331). IEEE. [CrossRef]

[36] Wang, J., Ai, J., Lu, M., Jiang, J., Yang, T., Bai, Z., ... & Wang, D. (2026). A survey of neural network robustness assessment in image recognition. *ACM Computing Surveys, 58*(12), 1-40. [CrossRef]

[37] Pokkuluri, K. S., Latha, Y. M., Kumar, P. S., & Narayana, B. L. (2024). Enhancing the accuracy of image segmentation using deep learning techniques. *International Journal of Intelligent Systems and Applications in Engineering, 12*(4), 474-482. https://ijisae.org/index.php/IJISAE/article/view/6128

[38] Dehdashtian, M., Rafiei, M. H., & Adeli, H. (2024). Fairness and bias mitigation in computer vision: A

survey. *arXiv preprint arXiv:2408.02464*. [CrossRef]

[39] Wang, T., Zhao, J., Yatskar, M., Chang, K. W., & Ordonez, V. (2020). Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 5310-5319). IEEE. [CrossRef]

[40] Hänel, A., Tschernezki, V., Wiegand, T., & Samek, W. (2022). Enhancing the fairness of visual attribute predictors. In *Proceedings of the Asian Conference on Computer Vision* (pp. 2707-2724).

[41] Halder, A., Gharami, S., Sadhu, P., Singh, P. K., Woźniak, M., & Ijaz, M. F. (2024). Implementing vision transformer for classifying 2D biomedical images. *Scientific Reports, 14*(1), 12567. [CrossRef]

[42] Sara, U., Akter, M., & Uddin, M. S. (2019). Image quality assessment through FSIM, SSIM, MSE and PSNR—a comparative study. *Journal of Computer and Communications, 7*(3), 8-18. [CrossRef]

[43] Al Najjar, Y. (2024). Comparative analysis of image quality assessment metrics: PSNR, SSIM, FSIM, and NAE. *International Journal of Advanced Natural Sciences and Engineering Researches, 8*(4), 128-133. [CrossRef]

[44] Mason, A., Rioux, J., Clarke, S. E., Costa, A., Schmidt, M., Keough, V., ... & Beyea, S. (2020). Comparison of objective image quality metrics to expert radiologists' scoring of diagnostic quality of MR images. *IEEE Transactions on Medical Imaging, 39*(4), 1064-1072. [CrossRef]

[45] Arabboev, M., Muksimov, A., Mukhtorov, D., Muminov, B., Abdullaev, S., & Djuraev, O. (2024). A comprehensive survey on image super-resolution: Evaluation metrics from signal processing to deep learning perspectives. *Journal of Imaging, 10*(10), 252. [CrossRef]

[46] Sharma, A. (2024). Image quality assessment using distortion metrics for hyperspectral image compression. *Remote Sensing Letters, 15*(9), 899-910. [CrossRef]

[47] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427-437. [CrossRef]

[48] Khan, A., Rauf, Z., Sohail, A., Khan, A. R., Asif, H., Asif, A., & Farooq, U. (2024). Early disease identification of rice and wheat crops using a transfer learning approach on smart handheld devices. *Frontiers in Plant Science, 15*, 1383748. [CrossRef]

[49] Kim, H., Seo, H., & Kim, E. (2025). A systematic review of hybrid vision transformers for radiological image analysis: Architectures, clinical applications, and future directions. *Diagnostics, 15*(2), 221. [CrossRef]

[50] Vlăsceanu, G., Botezan, A., Dinu, A., & Acatrinei, C. (2024). Selecting metrics to evaluate image segmentation performance. *Applied Sciences, 14*(22), 10605. [CrossRef]

[51] Müller, D., Soto-Rey, I., & Kramer, F. (2022). Towards a guideline for evaluation metrics in medical image segmentation. *BMC Research Notes, 15*(1), 210. [CrossRef]

[52] Taha, A. A., & Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging, 15*(1), 29. [CrossRef]

[53] Khan, A., Sohail, A., Ali, A., & Khan, A. (2024). Hybrid deep learning model for efficient classification and segmentation of chest X-ray images to detect tuberculosis. *Signal, Image and Video Processing, 18*(1), 713-723. [CrossRef]

[54] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 586-595). IEEE. [CrossRef]

[55] Padilla, R., Passos, W. L., Dias, T. L., Netto, S. L., & da Silva, E. A. (2021). A comparative analysis of object detection metrics with a companion open-source toolkit. *Electronics, 10*(3), 279. [CrossRef]

[56] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision, 88*, 303-338. [CrossRef]

[57] Oksuz, K., Cam, B. C., Akbas, E., & Kalkan, S. (2022). One metric to measure them all: Localisation recall precision (LRP) for evaluating visual detection tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(9), 4942-4959. [CrossRef]

[58] Rodríguez-Lira, A., Mata, C., Solano-Rojas, B., Rodríguez-Bermúdez, G., Villanueva-Polanco, R., Martínez-Arroyo, M., & Quesada-López, C. (2025). Recent advances in deep learning-based 3D reconstruction: A survey. *Electronics, 14*(15), 3032. [CrossRef]

[59] Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003, November). Multiscale structural similarity for image quality assessment. In *The thirty-seventh Asilomar Conference on Signals, Systems & Computers, 2003* (Vol. 2, pp. 1398-1402). IEEE. [CrossRef]

[60] Celaya, A., Actor, J. A., Muthusivarajan, R., Gates, E., Chung, C., Schellingerhout, D., ... & Fuentes, D. (2023). A generalized surface loss for reducing the Hausdorff distance in medical imaging segmentation. *arXiv preprint arXiv:2302.03868*. [CrossRef]

[61] Mittal, A., Soundararajan, R., & Bovik, A. C. (2012). Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters, 20*(3), 209-212. [CrossRef]

[62] Sheikh, H. R., & Bovik, A. C. (2006). Image information and visual quality. *IEEE Transactions on Image Processing, 15*(2), 430-444. [CrossRef]

[63] Al-Khafaji, S. L., & Ramaha, G. (2025). Hybrid deep learning model for image compression. *Baghdad Science Journal, 22*(3), 1030-1040. [CrossRef]

[64] Chen, T., Liu, H., Ma, Z., Shen, Q., Cao, X., & Wang, Y. (2020). End-to-end learnt image compression via non-local attention optimization and improved context quantization. *arXiv preprint arXiv:2007.02711*. [CrossRef]

[65] Gu, S., Zhang, L., Zuo, W., Feng, X., & Sang, N. (2016). A no-reference perceptual sharpness assessment for ultra-high-definition 4k displays. In *2016 IEEE International Conference on Image Processing (ICIP)* (pp. 2369-2373). IEEE. [CrossRef]

[66] Zhao, H., Gallo, O., Frosio, I., & Kautz, J. (2017). Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging, 3*(1), 47-57. [CrossRef]

[67] Babu, V. R., Chelladurai, D., & Jeyabharathi, S. (2024). Image compression and reconstruction via enhanced quality using deep-learning. *Journal of Intelligent & Fuzzy Systems, 46*(4), 9591-9606. [CrossRef]

[68] Ohayon, G., Michaeli, T., & Elad, M. (2025, May). Posterior-mean rectified flow: Towards minimum MSE photo-realistic image restoration. In *International Conference on Learning Representations* (Vol. 2025, pp. 63101-63140).

[69] Arumugam, P., & Banothu, B. (2025). Integrating multi-scale attention network for image restoration task with noise estimation. *Image and Vision Computing, 156*, 105419. [CrossRef]

[70] Seif, G., & Androutsos, D. (2018). Edge-based loss function for single image super-resolution. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1468-1472). IEEE. [CrossRef]

[71] Seitzer, M., Yang, G., Schlemper, J., Oktay, O., Würfl, T., Christlein, V., ... & Maier, A. (2018). Adversarial and perceptual refinement for compressed sensing MRI reconstruction. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I* (pp. 232-240). Springer International Publishing. [CrossRef]

[72] Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. (2023). Fast imaging enhancement network: Iterative refinement for noise removal in coronary angiography images. *Biomedical Signal Processing and Control, 84*, 104827. [CrossRef]

[73] Dobre-Baron, O., Nicola, C., Nicola, M., & Istratoaie, O. (2025). Evaluating image quality metrics as loss functions for deep learning-based image dehazing. *Electronics, 14*(5), 1040. [CrossRef]

[74] Keleş, O., Tarhan, C., & Bartan, O. (2021). On the computation of PSNR for a set of images or video. In *2021 Picture Coding Symposium (PCS)* (pp. 1-5). IEEE. [CrossRef]

[75] Dong, C., Loy, C. C., He, K., & Tang, X. (2015). Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38*(2), 295-307. [CrossRef]

[76] Hore, A., & Ziou, D. (2010). Image quality metrics: PSNR vs. SSIM. In *2010 20th International Conference on Pattern Recognition* (pp. 2366-2369). IEEE. [CrossRef]

[77] Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development, 15*(14), 5481-5487. [CrossRef]

[78] Radhabai, B. R., Subramani, B., & Karuppasamy, P. (2024). An intelligent deep learning approach for diagnosis of brain MRI images. *Computers and Electrical Engineering, 115*, 109123. [CrossRef]

[79] Maliński, Ł. (2023). Decomposed dissimilarity measures for image denoising assessment. *Signal, Image and Video Processing, 17*(4), 945-952. [CrossRef]

[80] Wu, J. (2025). Optimizing MRI signal-to-noise ratio using the Fourier algorithm. *Journal of Imaging, 11*(3), 88. [CrossRef]

[81] Gnanasambandam, A., & Chan, S. H. (2022). Exposure-referred signal to noise ratio for digital image sensors. *IEEE Transactions on Image Processing, 31*, 2653-2665. [CrossRef]

[82] Ayde, R., Engelbert, T., Nikulin, P., Schär, M., Pruessmann, K., & Weiss, K. (2025). MRI SNR: A software solution review. *Magnetic Resonance in Medicine, 93*(3), 933-947. [CrossRef]

[83] Tabuchi, A., Nakamura, Y., Seki, S., Ueda, T., & Ikeda, D. (2022). Estimation of signal-to-noise ratio for a clinical X-ray CT image. *Diagnostics, 12*(12), 3133. [CrossRef]

[84] Zhang, Q., Jiang, X., & Li, L. (2025). Self-supervised denoising methods for medical images: A survey. *Computers in Biology and Medicine, 186*, 109624. [CrossRef]

[85] Kojima, S., Shinohara, H., Ueda, T., Ohira, S., Akino, Y., Shimamoto, H., ... & Konishi, K. (2025). Practical MRI signal-to-noise ratio mapping: A comparative study of methods and clinical applications. *Radiological Physics and Technology, 18*(2), 393-404. [CrossRef]

[86] Fiete, R. D., & Tantalo, T. (2001). Comparison of SNR image quality metrics for remote sensing systems. *Optical Engineering, 40*(4), 574-585. [CrossRef]

[87] Xu, X., Wang, R., Fu, C. W., & Jia, J. (2022). SNR-aware low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 17714-17724). IEEE. [CrossRef]

[88] Sundarrajan, B., Ramakrishnan, S., Narayana, S., Sathishkumar, V. E., & Prabhu, J. (2024). Enhancing low-light medical images using a deep learning-based noise reduction approach. *Scientific*

Reports, 14(1), 30590. [CrossRef]

[89] Usamentiaga, R., García, D. F., & Molleda, J. (2018). More than fifty shades of grey: quantitative characterization of defects and anomalies in infrared images for nondestructive testing. *Sensors, 18*(10), 3396. [CrossRef]

[90] Liu, Y., Zhang, L., Lu, Y., Liao, Y., Zhao, Q., Yan, Q., ... & Ma, J. (2025). Joint super-resolution and denoising of MRI images via a multi-branch feature-extraction network. *IEEE Transactions on Instrumentation and Measurement, 74*, 5014014. [CrossRef]

[91] Zhang, L., Zhang, L., Mou, X., & Zhang, D. (2011). FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing, 20*(8), 2378-2386. [CrossRef]

[92] Wandile, K., Dhore, N., Suryawanshi, A., Wanjari, G., Deshmukh, P., & Wasankar, S. (2025). Filtering methods for image denoising and performance evaluation criteria. *International Journal for Research in Applied Science and Engineering Technology, 13*(3), 1505-1520. [CrossRef]

[93] Afnan, Ullah, F., Yaseen, Lee, J., Jamil, S., & Kwon, O. J. (2023). Subjective assessment of objective image quality metrics range guaranteeing visually lossless compression. *Sensors, 23*(3), 1297. [CrossRef]

[94] Singh, G., Sehgal, P., Kaur, A., Bhatt, R., Panchal, B., & Khan, I. R. (2025). An innovative deep learning method based on transformer for denoising of digital mammographic images. *Scientific Reports, 15*(1), 3416. [CrossRef]

[95] Robeson, S. M., & Willmott, C. J. (2023). Decomposition of the mean absolute error (MAE) into systematic and unsystematic components. *PLOS ONE, 18*(2), e0279774. [CrossRef]

[96] Karunasingha, D. S. K. (2022). Root mean square error or mean absolute error? Use their ratio as well. *Information Sciences, 585*, 609-629. [CrossRef]

[97] Eskicioglu, A. M., & Fisher, P. S. (1995). Image quality measures and their performance. *IEEE Transactions on Communications, 43*(12), 2959-2965. [CrossRef]

[98] Asamoah, D., Effah-Baah, J., Owusu-Banahene, W., & Danso, O. (2018). Measuring image contrast in digital images. *Journal of Engineering and Applied Sciences, 13*, 5347-5355. [CrossRef]

[99] Rainio, O., & Klén, R. (2026). Modified dice coefficients for evaluation of tumor segmentation from PET images. *Journal of Imaging Informatics in Medicine, 39*(1), 785-793. [CrossRef]

[100] Zou, K. H., Warfield, S. K., Bharatha, A., Tempany, C. M., Kaus, M. R., Haker, S. J., ... & Kikinis, R. (2004). Statistical validation of image segmentation quality based on a spatial overlap index: scientific reports. *Academic radiology, 11*(2), 178-189. [CrossRef]

[101] Eelbode, T., Bertels, J., Berman, M., Vandermeulen, D., Maes, F., Bisschops, R., & Blaschko, M. B. (2020). Optimization for medical image segmentation: Theory and practice when evaluating with dice coefficient and dice loss. *IEEE Transactions on Medical Imaging, 39*(11), 3679-3690. [CrossRef]

[102] Bertels, J., Eelbode, T., Berman, M., Vandermeulen, D., Maes, F., Bisschops, R., & Blaschko, M. B. (2019). Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22* (pp. 92-100). Springer International Publishing. [CrossRef]

[103] Kato, H., & Hotta, K. (2022). Adaptive t-vMF dice loss: An effective expansion of dice loss for medical image segmentation. *arXiv preprint arXiv:2207.07842*. [CrossRef]

[104] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., ... & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3213-3223). IEEE. [CrossRef]

[105] Ataş, M. (2023). Performance evaluation of the Jaccard-Dice coefficient in building segmentation from high-resolution satellite images. *Sensors, 23*(9), 4505. [CrossRef]

[106] Pont-Tuset, J., & Marques, F. (2016). Supervised evaluation of image segmentation and object proposal techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38*(7), 1465-1478. [CrossRef]

[107] Maxwell, A. E., Warner, T. A., & Fang, F. (2021). Implementation of machine-learning classification in remote sensing: An applied review. *International Journal of Remote Sensing, 39*(9), 2784-2817. [CrossRef]

[108] Oksuz, K., Cam, B. C., Akbas, E., & Kalkan, S. (2018). Localization recall precision (LRP): A new performance metric for object detection. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 504-519). Springer. [CrossRef]

[109] Martin, D. R., Fowlkes, C. C., & Malik, J. (2004). Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 26*(5), 530-549. [CrossRef]

[110] Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics, 17*(1), 168-192. [CrossRef]

[111] Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21*(1), 6. [CrossRef]

[112] Ramli, R., Arof, H., Ibrahim, F., Syamsul Khair Mokhtar, H., & Ayob, M. A. (2022). Confusion

matrix-based performance measure of detection using a corner detector. *Pertanika Journal of Science & Technology, 30*(1). [CrossRef]

[113] de Zarzà, I., de Curtò, J., & Calafate, C. T. (2022). Detection and assessment of glaucoma using EfficientNet deep learning model. *Informatics in Medicine Unlocked, 31*, 100977. [CrossRef]

[114] Chicco, D., & Jurman, G. (2022). An invitation to greater use of Matthews correlation coefficient in robotics and artificial intelligence. *Frontiers in Robotics and AI, 9*, 876814. [CrossRef]

[115] Farhadpour, M., Farajzadeh, K., & Hashemzadeh, M. (2024). Selecting loss functions and evaluation metrics for multi-class imbalanced problems. *Applied Soft Computing, 167*, 112406. [CrossRef]

[116] Hemmatian, B., Kargar, M., & Ghaderpour, E. (2025). Addressing class imbalance via cluster-based noise reduction and SMOTE for classification. *Scientific Reports, 15*(1), 6988. [CrossRef]

[117] Obi, J. C. (2023). A comparative study of classification metrics. *International Journal of Computational and Applied Mathematics & Computer Science, 3*, 19-25. [CrossRef]

[118] Nahm, F. S. (2022). Receiver operating characteristic curve: overview and practical use for clinicians. *Korean Journal of Anesthesiology, 75*(1), 25-36. [CrossRef]

[119] Richardson, B., Bhatt, D., Bhatt, P., Bhatt, D., Bhatt, S., & Tiwari, P. (2024). The ROC curve: An accurate assessment of imbalanced datasets in machine learning. *Plos one, 19*(12), e0315331. [CrossRef]

[120] Chang, C. I. (2020). An analysis of 3D hyperspectral target detection using receiver operating characteristic curves with three-dimensional ROC (3DROC) analysis. *Remote Sensing, 12*(16), 2604. [CrossRef]

[121] Aguilar-Ruiz, J. S., & Michalak, K. (2024). Classification performance evaluation with imbalanced multi-class data. *Neurocomputing, 594*, 127856. [CrossRef]

[122] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE, 10*(3), e0118432. [CrossRef]



Muzammil Khan has over 14 years of experience in teaching, research, and project supervision in Computer Science. He is currently serving as an Assistant Professor at the University of Swat. His expertise spans artificial intelligence, natural language processing (NLP), educational data mining, and digital image processing. He has authored over 70 peer-reviewed publications and supervised multiple PhD and MS research projects. His core research contributions include the development of a multilingual digital news preservation framework using NLP and cross-lingual linking techniques, as well as the design of machine learning-based diagnostic models for diseases such as tuberculosis, cancer, and coronary artery disease under the AuSC framework. He is actively involved in interdisciplinary, applied research aimed at solving real-world problems, ranging from healthcare diagnostics to language preservation for low-resource languages. With strong technical proficiency in Python, Java, and C++, and hands-on experience with AI tools and data mining frameworks, he bridges academic research with scalable, impactful applications.



Muhammad Hamza is a Computer Science graduate and is currently pursuing a Master's degree in Computer Science at the University of Swat, Swat, 19120, Pakistan. His research interests include medical image analysis using machine learning and deep learning, as well as hybrid approaches such as ensemble learning. (Email: mhamzaedu1@gmail.com)



Sikandar Ali is a Computer Science graduate and is currently pursuing a Master's degree in Computer Science at the University of Swat, Pakistan. He is serving as a Visiting Lecturer in the Department of Computer and Software Technology. His research interests include machine learning, deep learning, computer vision, and natural language processing. (Email: sikandarali51006@gmail.com)



Ibrar Ali is a Computer Science graduate and is currently pursuing a Master's degree in Computer Science at the University of Swat, Pakistan. He is serving as a Visiting Lecturer in the Department of Computer and Software Technology. His research interests include machine learning, deep learning, computer vision, and natural language processing. (Email: ibraralin@gmail.com)



Zaheer Khan is an AI/ML Engineer and Independent Researcher, currently serving as a Data Administrator at Spinghar Technical and Vocational Training Center. His research interests focus on artificial intelligence applications, machine learning, natural language processing (NLP), and digital image processing. He is passionate about developing practical AI-driven solutions and exploring emerging technologies through independent research and applied projects. (Email: xaheerkhancs@gmail.com)