Iranian Journal of Medical Physics

ijmp.mums.ac.ir

Image Analysis Metrics in Evaluation of Generative Adversarial Network (GAN) Models Performance in Prediction of ^{177}Lu Dose Voxel Kernels

Erick Otieno Kapis[1,2*], M. I. Kaniu[1], Giuseppe Iaccarino[3], H. K. Angeyo[1]

1. Department of Physics, University of Nairobi, P. O. Box 30197, 00100, Nairobi, Kenya.
2. Department of Radiation Oncology, Cancer Treatment Center, The Nairobi Hospital, P.O. Box 30026, 00100 Nairobi, Kenya.
3. Laboratory of Medical Physics and Expert Systems, National Cancer Institute Regina Elena, Rome, 00144, Italy.

ARTICLE INFO

Article type:
Original Paper

Article history:
Received: May 27, 2025
Accepted: Sep 04, 2025

Keywords:
Lutetium-177
Machine Learning
Monte Carlo Method

ABSTRACT

Introduction: In Generative Adversarial Networks (GANs), validation datasets are typically excluded from the adversarial optimization loop, unlike in many conventional machine learning architectures. This study evaluates GAN performance for predicting Dose Voxel Kernels (DVKs) using dedicated image-based metrics and compares these outcomes with training accuracy.

Material and Methods: Density Kernels (DKs) of size 15×15×15 voxels (2.43 mm³ voxel size) were generated from homogeneous materials and CT images using ctcreate/EGSnrc. Each DK incorporated a centrally located isotropic ^{177}Lu source, and corresponding DVKs were simulated using DOSXYZnrc/EGSnrc Monte Carlo methods. Paired DK-DVK datasets were used to train multiple GAN models. Model performance on unseen validation data comprising DKs from water, soft tissue, kidney, and bone was assessed using Structural Similarity Index Method (SSIM), Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and the Relative Global Dimensionless Error of Synthesis (ERGAS).

Results: Twelve GAN models with training accuracies between 96.5% and 99.26% were evaluated. Despite achieving the highest training accuracy, the 99.26% model did not exhibit the best predictive quality. Instead, the 98.4% model achieved superior performance, showing lower MSE (0.020 vs. 0.029 mGy²/MBq·s²), higher PSNR (43.25 vs. 41.35 dB), and a markedly lower ERGAS (10.96 vs. 15.75). SSIM values were consistently high (>0.99) across all models, with no statistically significant differences ($p > 0.05$), indicating comparable structural fidelity.

Conclusion: Training accuracy alone does not reliably reflect GAN performance. Image-based similarity and error metrics provide a more comprehensive and discriminative evaluation of ^{177}Lu DVK prediction quality.

► *Please cite this article as:*
Kapis EO, Kaniu MI, Iaccarino G, Angeyo HK. Image Analysis Metrics in Evaluation of Generative Adversarial Network (GAN) Models Performance in Prediction of ^{177}Lu Dose Voxel Kernels. Iran J Med Phys 2025; 22 (5): 331-340. 10.22038/ijmp.2025.89964.2590.

Introduction

Deep Neural Networks (DNN) and machine learning (ML) techniques have become highly valuable tools across many domains, particularly in medical image generation, manipulation, and analysis [1, 2]. For example, deep learning has been used to manage tasks such as image classification, segmentation and image generations [3]. Well-validated convolutional neural network (CNN) models have been utilized in radiation dose calculations and estimation in conventional radiotherapy and internal radiation dosimetry [3–5].

The accuracy and credibility of the trained machine learning models depend on the training dataset, validation and evaluation techniques applied [6]. Most machine learning architectures, like CNN and DNN follows a supervised learning paradigm, where the data is divided into training, validation, and test subsets. CNN for instance, learns to minimize the loss function between the predicted and the reference dose using paired input-output data [2]. In contrast, Generative Adversarial Networks (GANs) employs a dual-network architecture comprising of generator that creates synthetic outputs and discriminator that distinguishes generated data from real samples [7, 8]. The GANs mostly do not directly use the validation datasets during the adversarial training loop. Post training validation is performed using independent datasets to test the generator ability to produce realistic and quantitatively accurate outputs [7].

In nuclear medicine dosimetry, ML models can be trained to directly calculate whole body dose maps or predict dose voxel kernels (DVKs) that can be used in dose convolution with Positron Emission Tomography (PET) or Single Photon Emission Tomography (SPECT) time integrated activity (TIA) maps [9]. Most studies often use evaluation metrics such as mean squared error (MSE), mean relative absolute error (MARE), mean absolute error (MAE) and mean

**Corresponding Author:* Email: kapotin@ymail.com

absolute percentage error (MAPE) to evaluate the performance of DNN or CNN on the resulting dosimetry in comparison to Monte Carlo standards [1, 2, 8, 9].

Lee *et al.*[2] designed a U-Net CNN model to predict dose voxel values for ^{68}Ge-NOTA-RGD PET/CT imaging. PET and CT images were used as input datasets compared to Monte Carlo simulations as the ground truth. The voxel dose errors were 2.54±2.09% and mean organ dose error was about 1.07%. Akhavanallaf *et al.*[9] worked on whole-body voxel-based internal dosimetry using DNN on the ResNet architecture. Whole body CT density maps and reference voxel-wise S-values generated from the Monte Carlo N-Particle (MCNP) simulator were used as inputs. The dose calculation involved convolving S-value kernels with TIA maps derived from dynamic 18F-FDG PET images. They compared the DNN model results with standard single S-value (SSV) kernels, multiple S-value (MSV) kernels and Monte Carlo simulation using quantitative tools like mean absolute error (MAE), mean relative absolute error (MRAE %) and root mean square error (RMSE). The DNN results were superior to the SSV estimates but comparable to MC calculation. The correlation between predicted kernels and the MC simulation achieved a coefficient of determination, R^2 of 0.98.

The DNN or CNN performances as proposed by Naqa *et al.*, are dependent on the amount of dataset used for training and validation [6]. To achieve agreeable model accuracies, they need a lot of training dataset. In contrast, GANs while may use limited data, their model validation should include both prediction accuracies and realism scores [10]. Technical model training evaluation by Atta *et al.*, suggested the use of model accuracy with addition of other evaluation metrics [11]. Nonetheless, it has been reported that even two ML models with similar training accuracies may not always yield the same prediction accuracies [12] on the same input data, thus calling for better ways of accurate model performance evaluation criteria.

The current study is a new improvement of the 2D pix2pix GAN to a 3D GAN network for training GAN models using a 3D input dataset. This is an attempt to use the 3D GAN architecture to train models for predicting ^{177}Lu DVKs. ^{177}Lu is a therapeutic radionuclide used in targeted radionuclide therapy (TRT) in the form of Lu-177-DOTATAE and Lu-177-PSMA-617 for neuroendocrine tumors and metastatic castration resistant prostate cancer respectively [13–15]. The aim of this study was to apply a combination of image analysis metrics other than training accuracies to evaluate the performance of the selected trained GAN models when predicting ^{177}Lu DVKs. It aimed to check whether the models with high training accuracies would hold the best overall performances. GANs were deemed best suited for this study because they can be trained with resource constrained dataset and can reproduce high frequency three-dimensional spatial features unlike other ML architectures that requires a lot of training dataset.

Image analysis metrics such as the Structural Similarity Index Method (SSIM), Mean Squared Error (MSE), Peak Signal to Noise Ratio (PSNR), and Relative Global Dimensionless Error of Synthesis (ERGAS) were used to evaluate the predicted ^{177}Lu DVK images against the corresponding reference Monte Carlo-calculated DVK images. We hope that these metrics can be integrated in future GAN core designs to improve learning speed and accuracy. In addition, other parameters such as the overall kernel accuracy based on the per-voxel absorbed dose numerical accuracy were also investigated.

Materials and Methods

Monte Carlo Reference DVK Simulation

Homogeneous Density Kernels (DKs) of matrix size 15×15×15, with voxel dimensions of 2.4 mm×2.4 mm×2.4 mm, were generated using the EGSnrc/DOSXYZnrc Monte Carlo code. Each homogeneous kernel was assigned a uniform density corresponding to specific media, including water, lung, kidney, soft tissue, liver, and bone. In addition, heterogeneous CT-based density kernels representing selected human organs (kidney, liver, lung, bone, and soft tissue) were created using the ctcreate module in EGSnrc, allowing realistic spatial variations in tissue composition and density. The conversion of CT numbers to material densities was performed according to the calibration described by Schneider *et al.*[16].

For each DK, full beta decay spectrum of ^{177}Lu was simulated, characterized by an effective mean beta energy of 133 keV (100% intensity) and the two most prominent gamma emissions at 113 keV (6.23%) and 208 keV (10.41%). These emissions were isotropically emitted from the central voxel, defined as an isotropic point source. Beta particles and photons were tracked down to 10 keV kinetic energy, corresponding to energy cutoffs of ECUT = 0.521 MeV for electrons and PCUT = 0.01 MeV for photons. The PRESTA-II electron transport algorithm and XCOM photon cross sections were employed. A total of 5.0×10^8 particle histories as illustrated in Figure 1 were simulated in EGSnrc/DOSXYZnrc code to compute the corresponding DVKs.

The voxel-wise statistical uncertainties from the simulations were evaluated directly from the generated .3ddose DVK files using Equation 1. Across the full 15 × 15 × 15 voxel kernel, the median relative voxel uncertainty was 1.1%, with 95% of voxels below 2.1%. In high-dose regions near the central source voxel (≥ 1% of maximum dose), the median and 95th-percentile uncertainties were 0.47% and 0.87%, respectively. The DVK volume-averaged statistical uncertainty was 0.93%, substantially lower than typical systematic model uncertainties (~5%). These results demonstrate that the Monte Carlo statistical noise was negligible compared to model-related uncertainties, and does not affect the accuracy of the simulated dose voxel kernels.

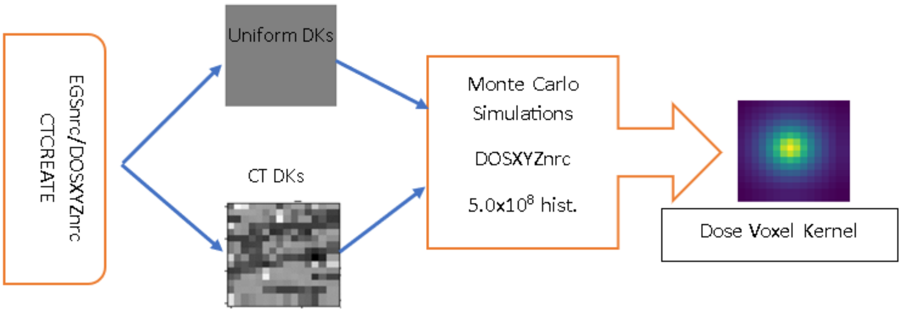


Figure 1. Generation of uniform density and CT based DKs using EGSnrc/ctcreate code and Monte Carlo simulation into Dose Voxel Kernels in DOSXYZnrc/EGSnrc codes using 5.0×108 primary particles

EGSnrc/DOSXYZnrc code was also preferred in this study because it is a widely validated, open-source Monte Carlo radiation transport code developed and maintained by the national Research Council of Canada. It has been extensively benchmarked against experimental measurements and other well-established Monte Carlo codes like MCNP and GEANT4 [17].

$$u_{kernel}(\%) = 100 \times \frac{\sqrt{\sum_{i=1}^{N}\sigma_i^2}}{\sum_{i=1}^{N} D_i} \quad (1)$$

u_{kernel} is the relative statistical uncertainty of the dose in kernel, N is the number of voxels in the kernel, D_i is dose in voxel i, $\sigma_i$is the statistical uncertainty of the dose in voxel i.

Generative Adversarial Networks (GAN) Models

A new three-dimensional GAN network was designed in python code following the basic ideas in pix2pix two-dimensional GAN [7, 12, 18]. It was implemented to train models that can predict ^{177}Lu DVKs from input density kernels (DKs) (16×16×16 voxels). The architecture comprises a U-Net-based generator and a 3D PatchGAN discriminator as illustrated in Figure 2. The DK and DVK kernels from Monte Carlo simulation were edge padded from (15×15×15) to (16×16×16) and used as input data.

The generator was designed as a 3D U-Net encoder-decoder structure. The encoder consists of four 3D convolutional layers (kernel = 4×4×4, stride = 2, padding = 1) with feature depths of 64, 128, 256, and 512, respectively. Each layer uses LeakyReLU (α = 0.2) activation and batch normalization (momentum = 0.8). The bottleneck includes one convolutional layer (kernel = 4×4×4, stride = 1) with 512 filters, batch normalization, and LeakyReLU activations, capturing global dose features.

The decoder mirrors the encoder with four 3D transposed convolutional layers (kernel = 4×4×4, stride = 2, padding = 1) and channel depths of 512, 256, 128, and 64. Each layer employs ReLU activation, batch normalization, and skip connections to corresponding encoder layers. A final 3D convolutional layer (kernel = 1×1×1, stride = 1, padding = 0) with sigmoid activation outputs normalized DVK in the [0,1] range.

The discriminator adopts a 3D PatchGAN design, taking concatenated pairs of DKs and DVKs as input. It includes three 3D convolutional layers (kernel = 4×4×4) with filter counts of 64, 128, and 1. The first two layers use stride = 2, and the final layer stride = 1, producing a local realism map corresponding to 22×22×22 receptive fields. Each layer applies instance normalization and LeakyReLU (α = 0.2) activation. The final output is a 3D probability map indicating the physical plausibility of each local patch. The GAN design is illustrated in Figure 2.

The generator loss, L_G combines an adversarial term and voxel-wise L1 loss as given in Equation 2.

$$L_G = L_{adv} + \lambda L_{L1}, \quad (2)$$

where $L_{adv}$is the adversarial loss encouraging the generator to produce dose distributions that the discriminator classifies as realistic, and $L_{L1}$is the voxel-wise L1 reconstruction loss, enforcing numerical similarity between the generated $(G(x))$ and true (y) dose kernels. The weighting factor λ (set between 50–100) balances perceptual realism and voxel-level accuracy.

The discriminator loss, L_D was optimized using a binary cross-entropy (BCE) loss, Equation 3.

$$L_D = -E[\log D(y,x)] - E[\log(1 - D(G(x),x))], \quad (3)$$

where $D(y,x)$ represents the discriminator's probability that the real pair (true DK and DVK) is authentic, and $D(G(x),x)$ represents the probability that the generated DVK pair is real. This loss penalizes incorrect classifications, enabling the discriminator to guide the generator toward physically consistent and realistic dose predictions.

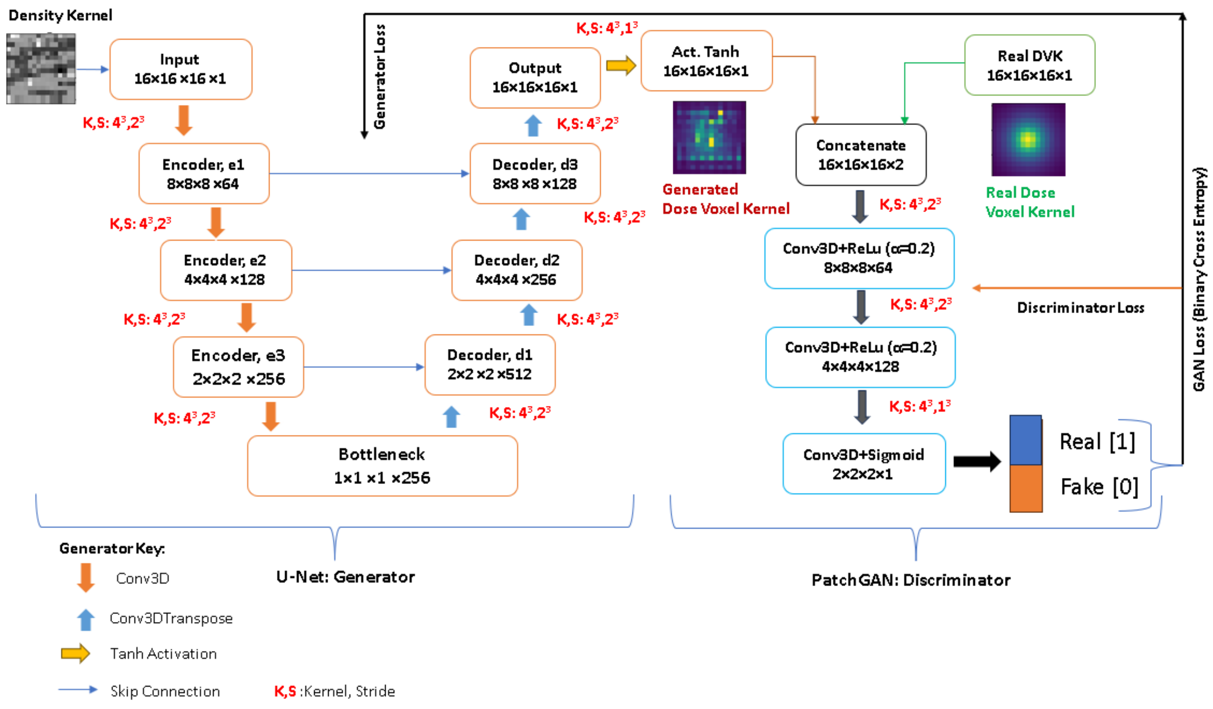


Figure 2. Three-dimensional GAN with U-Net Generator and PatchGAN Discriminator. The discriminator and generator losses are fed back respectively to improve the adversarial training processes.

Physical correspondence preserving data augmentations like rotation by 90 and 180 degrees were applied on the Monte Carlo simulated DK-DVK pairs to increase training dataset. The augmentation was useful in enhancing model stability and generalization. Uniform density DK-DVK pairs and heterogenous density DK-DVK pairs were trained separately. A total of 3,020 and 6,040 input data was used to train Uniform and Mixed models respectively. The batch size equal to 1 recommended by Isola *et al*.[10] was found to be realistic.

The models generated were saved automatically at the end of each epoch. The model's prediction accuracy was calculated by comparing the voxel values of the generated DVK with the reference DVK using Equation 4. It calculated the fraction of voxels whose predicted dose values were numerically within absolute tolerance of 10^{-15} Gy/particle of the reference DVK. Given the voxel dose values range from 1×10^{-9} to 1×10^{-17} Gy/particle, the tolerance value was reasonable as it allows numerical noise in low dose voxels while being strict enough in high dose regions.

Percent Accuracy (%) = $\frac{100}{N}\sum_{i=1}^{N}[|x_i - y_i| \le 10^{-15}]$ (4)

x_i and y_i are the voxel values from the reference DVK and generated DVK respectively. N is the total number of voxels in the kernel while the value 10^{-15} is the absolute tolerance for the compared i^{th} voxel values. The models' performances were also analyzed based on the type of training dataset used (either from uniform density DKs or heterogenous DKs). Twelve models with training accuracies between 96.5% and 99.26% were randomly selected for performance analysis.

Data Analysis

In addition to model training accuracy, image analysis metrics were quantitatively used to evaluate the performances of each of the selected models using their predicted ^{177}Lu DVKs.

Mean Squared Error (MSE)

MSE quantifies the average squared difference between reference and predicted DVKs as given in Equation 5.

$$MSE=\frac{1}{MN}\sum_{n=0}^{M}\sum_{m=1}^{N}[g(n,m)-h(n,m)]^2 \quad (5)$$

where g and h represent the pixel values the reference and generated DVKs respectively. M and N represent the image dimensions.

Peak Signal to Noise Ratio (PSNR)

PSNR, in decibels (dB), measures image similarity relative to noise as illustrated in Equation 6.

$$PSNR=\frac{10log_{10}(Peakval^2)}{MSE} \quad (6)$$

Higher PSNR indicates better quality (typically 30–50 dB for 8-bit images). MSE = 0 implies infinite PSNR.

Structural Similarity Index Method (SSIM)

SSIM compares luminance, contrast, and structure between images. Values range from –1 to +1. SSIM is calculated using Equation 7.

$$SSIM(x,y)=\frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x{}^2+\mu_y{}^2+C_1)(\sigma_x{}^2+\sigma_y{}^2+C_2)} \quad (7)$$

where μ, σ, and σ_{xy} are mean, standard deviation, and covariance over local windows. Window size of 11×11 and other default parameter settings were used.

Relative Global Dimensionless Error of Synthesis (ERGAS)

ERGAS is a relative global error metric that measures the overall root mean square error (RMSE) between the reference DVK image and the GAN-predicted DVK as given in Equation 8.

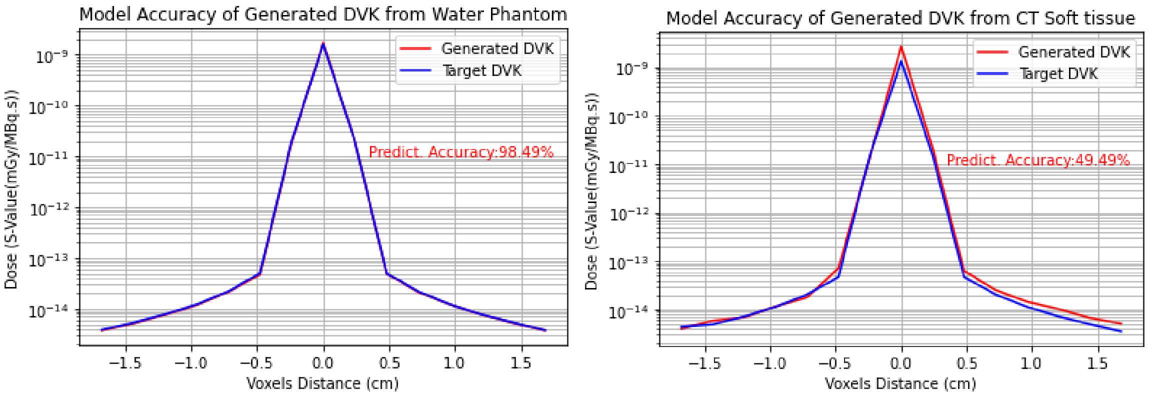
$$\text{ERGAS}=100\ \times\frac{1}{r}\sqrt{\frac{1}{N_b}\sum_{b=1}^{N_b}\left(\frac{RMSE_b}{\overline{M}_{ref,b}}\right)^2} \quad (8)$$

N_b is the number of image channels, r is the ratio of high resolution to low resolution pixels (set equal to 4), $RMSE_b$ is the root mean square of generated and reference DVKs for channel b while $\overline{M}_{ref,b}$ is the mean value of the reference DVK image for channel b.

Results

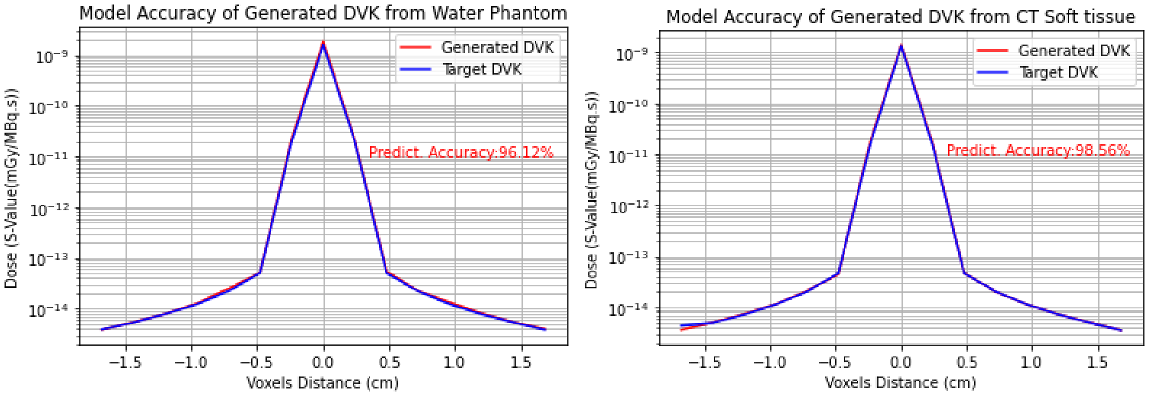
Two sets of trained models, Uniform and Mixed models were compared in terms of their DVK kernel prediction accuracies. Figure 3 (a) and (b), show profile of DVKs predicted from water and CT soft tissue DKs using Uniform model respectively. The Uniform model has high DVK prediction accuracy (98.49%) in water DKs and significantly low prediction accuracy (49.49%) in CT soft tissue. This observation is an indication of a model collapse [12] owing to the lack of training diversity in the Uniform model. In contrast, DVK prediction by the Mixed model for the same dataset is shown in Figure 4 (a) and (b), respectively. It shows a good prediction performance for both water and CT soft-tissue media.

The DVK prediction accuracies for the two sets of models are plotted in Figure 5. They were used to predict the DVKs of four DK materials: water, CT kidney, CT bone, and CT soft tissue.



(a) (b)

Figure 3. Uniform model showing DVK prediction accuracies in (a) water DK and (b) CT soft-tissue DK medium.



(a) (b)

Figure 4. Mixed model showing DVK prediction accuracies in (a) water DK and (b) CT soft-tissue DK medium.

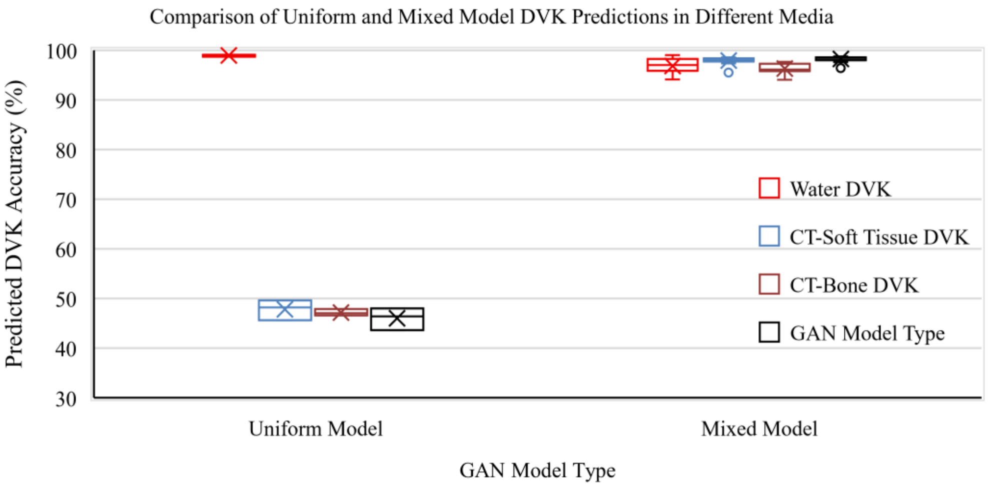


Figure 5.Uniform and mixed model DVK prediction accuracies on both homogenous (water medium) and heterogenous materials (CT soft tissue, CT bone and CT Kidney).

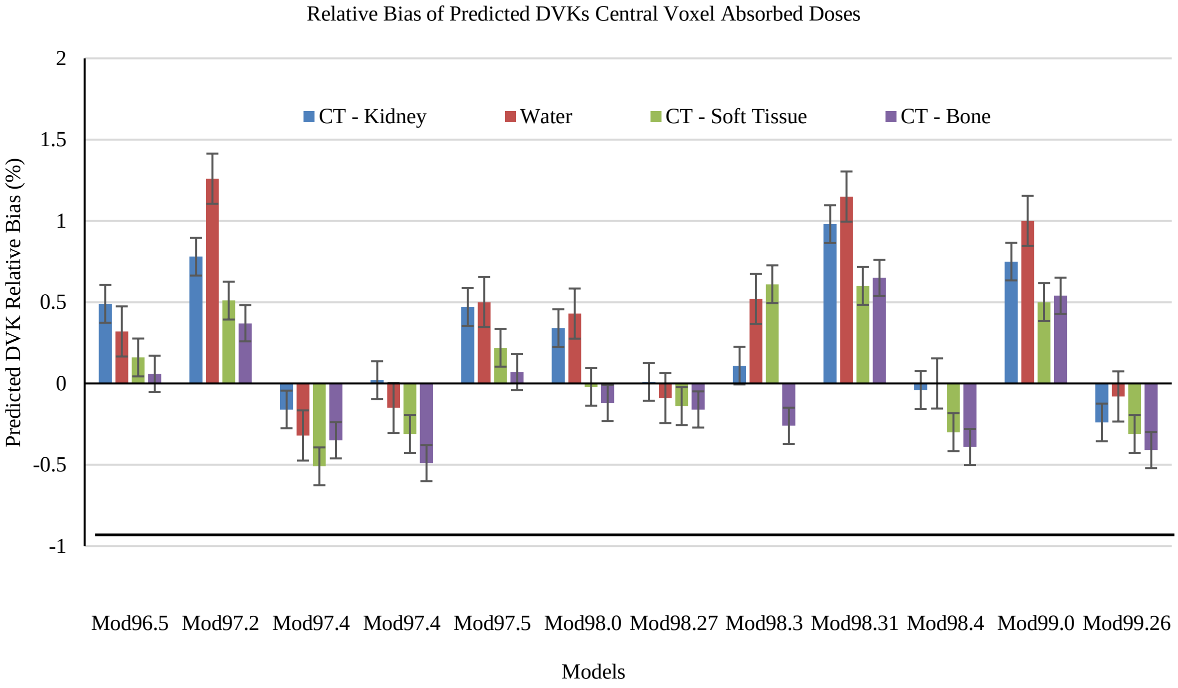


Figure 6.Relative differences in the central voxel absorbed doses of the model- predicted DVKs compared to reference DVKs.

The Uniform models performed well with an average prediction accuracy of 98.91% on water-based DKs, but failed in heterogeneous media such as CT soft tissue, CT bone, and CT kidney, with average accuracies of 47.81%, 47.17%, and 46.01%, respectively. In contrast, the mixed models demonstrate strong generalization across homogeneous and heterogeneous tissues leading to high DVK prediction accuracies ranging between 96% and 99.27%. These observations are consistent with the model training requirements that for the model to be robust and to avoid overfitting, the training must be exposed to a variety of datasets to improve generalization [4].

Image Quality Tests Analysis

Various image quality analysis algorithms were applied to the generated DVK images against the corresponding Monte Carlo-simulated reference DVKs. To avoid directional bias, a 2D image plane was randomly extracted from both the GAN model generated and reference DVK images for comparative analysis. The image analyses metrics such as MSE, PSNR, SSIM, and ERGAS were applied to DVK images generated by the mixed models only. Twelve mixed models with training accuracies varying between 96.50% (mod96.50%) and 99.26% (mod99.26%) were used to predict the ^{177}Lu DVKs from four different DK media. While all the predicted DVKs had

SSIM values greater than 0.99, the kernel prediction accuracy, PSNR, and ERGAS were the most indicative metrics of the models' performances.

Importantly, dose absorption at the central source voxel was considered a significant characteristic of DVKs. In all cases, the DVKs showed over 90% dose absorption at the central voxel compared to all other voxels. The percentage differences in the predicted central voxel doses were compared with the reference DVKs. Figure 6 presents the differences in the central voxel percentage contribution compared with the reference DVKs. Models with differences within ±0.5% were considered as best performing.

The relative biases indicate that the models: Mode96.50, Mod97.40, Mod97.50, Mod98.00, Mod98.27, Mod98.4, and Mod99.26, had differences within ±0.50% on all media. The other five models exhibit larger variations especially for water and kidney.

SSIM

The SSIM trends showed that the DVK predictions of the CT kidney had the highest values in all models with an average of 0.9995±0.0003. All models gave an SSIM of more than 0.99 on all media showing no significant differences ($p>0.05$). On average, Mod98.4 and Mod98.27 gave the highest SSIM values at 0.9992±0.0004 and 0.9992±0.0003, respectively. Based on the SSIM metric, these are likely to be the best-performing models. SSIM values closer to unity indicate a high degree of similarity between the predicted DVK and the reference DVK, and hence, a good model prediction performance indicator.

MSE

The mean squared error between the predicted and reference DVK values were generally less than 0.05 mGy^2/MBq.s^2. However, an outlier MSE value of 0.066 was recorded for mod99.26 in the prediction of the CT bone as shown in Figure 7. Models with an MSE closer to zero are considered to be best performing because they indicate that the voxel values of the predicted DVK image are statistically similar to the target DVKs.

Mod98.40 was the best performing with an average MSE of 0.020±0.006 mGy^2/MBq.s^2. The worst-performing model was Mod98, with an average of 0.030±0.012 mGy^2/MBq.s^2. Both Mod99 and Mode99.26 despite being models with high training accuracy, had average MSE of 0.025±0.002 mGy^2/MBq.s^2 and 0.029±0.021 mGy^2/MBq.s^2, respectively.

PSNR Analysis

Models generating DVKs with PSNR above 40 dB are considered high-performing [19]. Mod98.40 achieved the highest average PSNR (43.25±2.92 dB), followed by Mod98.27 (42.69±1.50 dB), while Mod99 and Mod99.26 trailed slightly at 40.86±0.80 dB and 41.35±5.09 dB respectively. Mod98 performed the worst (39.99±3.71 dB), but interestingly, Mod96.5 despite its low prediction accuracy, had a PSNR above 40 dB. These inconsistencies highlight that model accuracy alone may be misleading when deciding the best performing GAN model. PSNR better reflects image quality by comparing peak signal strength to noise, offering a more robust measure of similarity between generated and reference DVKs [12]. Figure 8 illustrate the PSNR for each model on four materials tested.

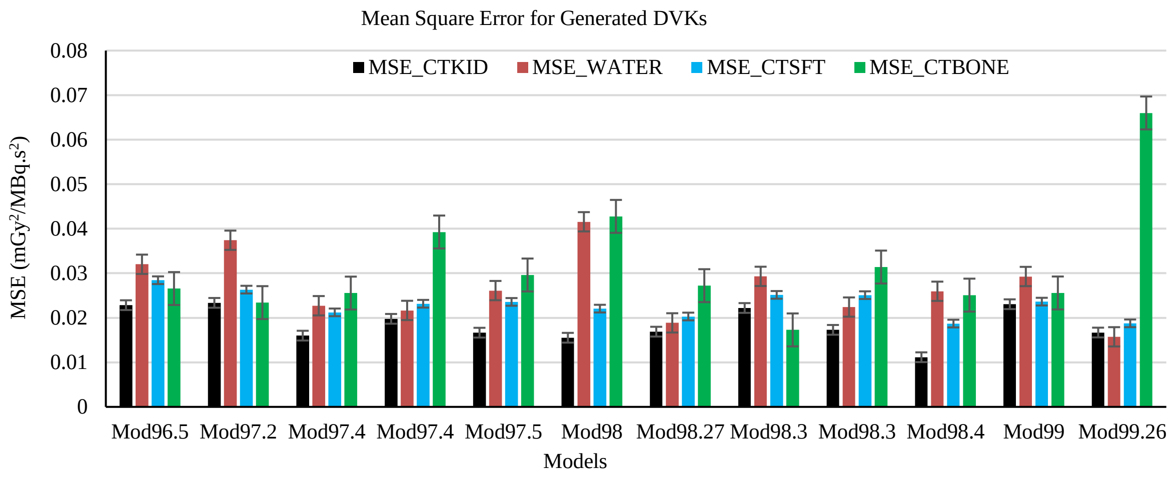


Figure 7. Mean Squared Error (MSE) of the Predicted DVKs by each model. Mod99.26 gave the highest MSE for bone than other models.

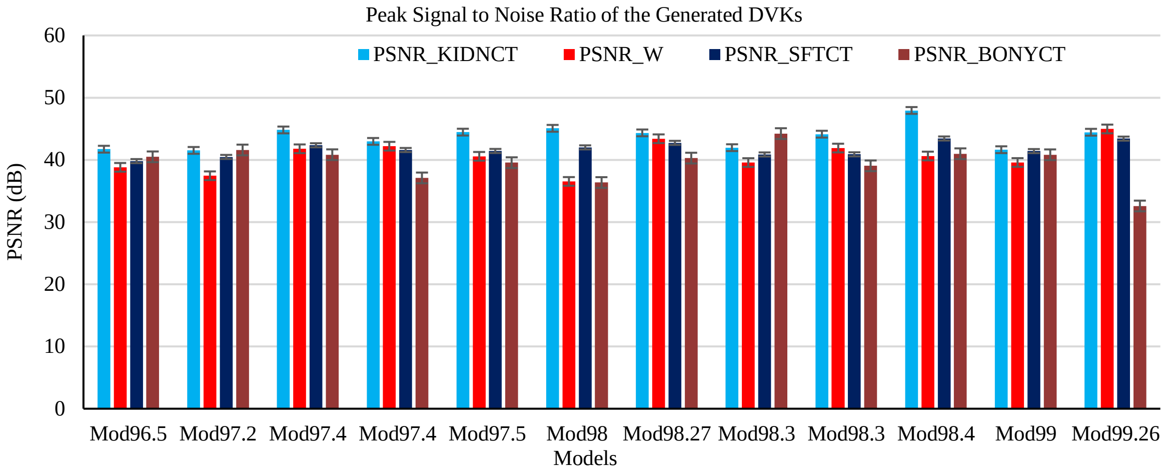


Figure 8. Peak Signal to Noise Ratio (PSNR) of predicted DVKs by all tested models.

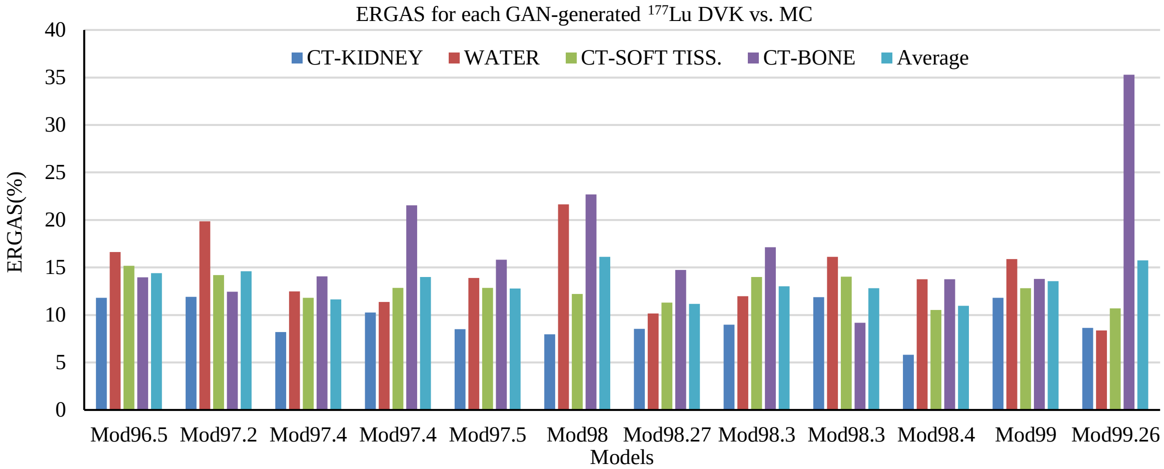


Figure 9. ERGAS values of the models on generated DVKs compared to reference DVKs

ERGAS

It provides a valuable and standardized method for assessing the quality of the generated images. ERGAS gives an overall measure of error across DVK datasets, thus enabling an understanding of the model's performance. Models with the lowest ERGAS values are considered the best performing. In the selected models, mod98 performed the worst with ERGAS 16.127±6.237 while mod98.4 is the best with an average ERGAS of 10.957±3.260 indicating voxel-wise good similarity. Figure 9 shows the ERGAS for the different selected models. Here, good generative models were considered to have an average ERGAS below 12.0.

Discussion

Many machine learning architectures such as GAN networks, are designed to produce trained models at different intervals during training. By default, the model training accuracy is the key indicator for the success of training as well as to inform when to stop further training [4]. However, this study emphasized that there are many model evaluation metrics that one can consider to obtain a high performing model that can be universally applied to predict ^{177}Lu DVKs and other radionuclides from density kernels of diverse heterogeneities. The image quality can be technically described to indicate the deviation from the ideal target image as it relates to the subjective perception and prediction of an image [20]. This implies that a combination of perceptive tests and qualitative analysis on the models are necessary for evaluation.

Table 1. Image Analysis Metrics for the top four models with training accuracies 96.5%, 98.4%, 99.0% and 99.26% respectively, used in predicting ^{177}Lu DVKs

Image Analysis Metrics	Selected Models Average Performance			
	mod96.5	mod98.4	mod99	mod99.26
MSE (mGy^2/MBq·s²)	0.027±0.003	0.020±0.006	0.025±0.002	0.029±0.021
PSNR (dB)	40.203±1.061	43.248±2.919	40.862±0.798	41.354±5.090
ERGAS (%)	14.387±1.759	10.957±3.260	13.572±1.501	15.747±11.326
SSIM	0.9989±0.0003	0.9992±0.0004	0.9990±0.0004	0.9981±0.0024
Central Voxel contribution diff. (%)	0.257±0.153	-0.183±0.166	0.698±0.199	-0.260±0.120

Training accuracy while robust and easily understandable, it can also be misleading since high accuracies may be due to model overfitting or due to class imbalance bias [21, 22]. Training accuracy may also ignore the validation or test performances since the model may only optimize the training dataset but fails to adapt to the variations in real-world data [19].

Incorporating other model evaluation metrics such as MSE, PSNR, SSIM, and ERGAS in addition to the basic model training accuracy can significantly improve training evaluation. Table 1 presents the results of four selected models used to predict ^{177}Lu DVKs from the four different DKs. It was realized that the best-performing models were actually not the models with the highest training accuracies. Mod98.4 performed better than all other models with high training accuracies like mod99 and mod99.26 in all evaluation metrics. It had the lowest MSE, the highest PSNR, lowest ERGAS, highest average SSIM and the lowest relative difference on central voxels absorbed dose. In this study, PSNR and ERGAS were considered the most important metrics for evaluating the overall performance of the models because of high sensitivity and PSNR is also partly derived from MSE.

While the SSIM measure is also quite important, it may not be a good indicator in cases where all models are capable of producing DVKs with SSIM of more than 0.99 in multi-class binary classification problem. It also has a major weakness to vary significantly depending on the implementation of parameters such as window size and constants [22]. The ERGAS allows for a better comparison across different scales or variables as it normalizes the error based on the size of the observations [20, 23].

This study illustrated the importance to carry out extensive full-image metrics analyses to determine the performance of GAN generated models. The application of generative Artificial Intelligence (AI) in nuclear medicine has been recognized in various studies as a key component for faster and more accurate personalized internal radiation dosimetry[5, 8]. Although, many machine learning architectures are being developed, it is necessary to develop model performance evaluation criteria.

Conclusion

This study demonstrates the critical role of image-based evaluation metrics in assessing the performance of Generative Adversarial Networks (GANs) and other machine learning (ML) models, particularly in the context of medical imaging and voxel-based dosimetry. Metrics such as Mean Squared Error (MSE), Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), and Relative Global Dimensionless Error of Synthesis (ERGAS) offer valuable quantitative measures of image fidelity and reconstruction accuracy. Their application is especially important in high-stakes fields like radiopharmaceutical therapy, where even small deviations in voxel intensity can affect clinical outcomes and treatment planning.

The study further underscores the necessity of using diverse and representative datasets to train ML models effectively. Insufficient data variability can lead to overfitting or mode collapse, limiting the generalizability of the generated outputs. By incorporating heterogeneity into the training set, models become more robust and capable of producing accurate dose voxel kernels across a range of anatomical and dosimetric conditions. These findings highlight not only the importance of carefully chosen evaluation metrics but also of comprehensive dataset design in developing reliable, clinically applicable machine learning tools.

Acknowledgment

This research received support from Department of Physics, University of Nairobi in providing high performing computers for Monte Carlo simulations.

References

1. Scarinci I, Valente M, Pérez P. A Machine Learning-Based Model for a Dose Point Kernel Calculation. EJNMMI Phys. 2023 June;10(1):41. doi:10.1186/s40658-023-00560-9.
2. Lee M S, Hwang D, Kim J H, Lee J S. Deep-Dose: A Voxel Dose Estimation Method Using Deep Convolutional Neural Network for Personalized Internal Dosimetry. Sci. Rep. 2019 July;9(1):10308. doi:10.1038/s41598-019-46620-y.
3. Götz T I, Lang E W, Schmidkonz C, Kuwert T, Ludwig B. Dose voxel kernel prediction with neural networks for radiation dose estimation. Z. Für Med. Phys. 2021 Feb.;31(1):23–36. doi:10.1016/j.zemedi.2020.09.005.
4. Currie G, Hawk K E, Rohren E, Vial A, Klein R. Machine Learning and Deep Learning in Medical Imaging: Intelligent Imaging. J. Med. Imaging Radiat. Sci. 2019 Dec.;50(4):477–87. doi:10.1016/j.jmir.2019.09.005.

5. Carter L M, Ocampo Ramos J C, Kesner A L. Personalized Dosimetry of 177 Lu-DOTATATE: A Comparison of Organ- and Voxel-Level Approaches Using Open-Access Images. Biomed. Phys. Eng. Express. 2021 Sept.;7(5):057002. doi:10.1088/2057-1976/ac1550.
6. El Naqa I, Ruan D, Valdes G, Dekker A, McNutt T, Ge Y, et al. Machine Learning and Modeling: Data, Validation, Communication Challenges. Med. Phys. 2018;45(10):e834–e40. doi:10.1002/mp.12811.
7. Kearney V, Chan J W, Wang T, Perry A, Descovich M, Morin O, et al. DoseGAN: A Generative Adversarial Network for Synthetic Dose Prediction Using Attention-Gated Discrimination and Generation. Sci. Rep. 2020 July;10(1):11073. doi:10.1038/s41598-020-68062-7.
8. Kim K M, Lee M S, Suh M S, Cheon G J, Lee J S. Voxel-Based Internal Dosimetry for 177Lu-Labeled Radiopharmaceutical Therapy Using Deep Residual Learning. Nucl. Med. Mol. Imaging. 2023 Apr.;57(2):94–102. doi:10.1007/s13139-022-00769-z.
9. Akhavanallaf A, Shiri I, Arabi H, Zaidi H. Whole-Body Voxel-Based Internal Dosimetry Using Deep Learning. Eur. J. Nucl. Med. Mol. Imaging. 2020 Sept. doi:10.1007/s00259-020-05013-4.
10. Isola P, Zhu J-Y, Zhou T, Efros A A. Image-to-Image Translation with Conditional Adversarial Networks. 2017 IEEE Conf. Comput. Vis. Pattern Recognit. CVPR. 2017 July:5967–76. doi:10.1109/CVPR.2017.632.
11. Agyeman A Y, Tetteh S G. Technical Evaluation of Machine Learning Models: An Empirical Study. 2024.
12. Karimipourfard M, Sina S, Mahani H, Karimkhani S, Sadeghi M, Alavi M, et al. A Taguchi-Optimized Pix2pix Generative Adversarial Network for Internal Dosimetry in 18F-FDG PET/CT. Radiat. Phys. Chem. 2024 May;218:111532. doi:10.1016/j.radphyschem.2024.111532.
13. Chi K N, Yip S M, Bauman G, Probst S, Emmenegger U, Kollmannsberger C K, et al. 177Lu-PSMA-617 in Metastatic Castration-Resistant Prostate Cancer: A Review of the Evidence and Implications for Canadian Clinical Practice. Curr. Oncol. 2024 Mar.;31(3):1400–15. doi:10.3390/curroncol31030106.
14. Sartor O, De Bono J, Chi K N, Fizazi K, Herrmann K, Rahbar K, et al. Lutetium-177–PSMA-617 for Metastatic Castration-Resistant Prostate Cancer. N. Engl. J. Med. 2021 Sept.;385(12):1091–103. doi:10.1056/NEJMoa2107322.
15. Research C for D E and. FDA Approves Lutetium Lu 177 Dotatate for Treatment of GEP-NETS. FDA. 2019 Feb.
16. Schneider W, Bortfeld T, Schlegel W. Correlation between CT Numbers and Tissue Parameters Needed for Monte Carlo Simulations of Clinical Dose Distributions. Phys. Med. Biol. 2000 Feb.;45(2):459. doi:10.1088/0031-9155/45/2/314.
17. Kawrakow I. Accurate Condensed History Monte Carlo Simulation of Electron Transport. I. EGSnrc, the New EGS4 Version. Med. Phys. 2000;27(3):485–98. doi:https://doi.org/10.1118/1.598917.
18. Hölscher D, Reich C, Gut F, Knahl M, Clarke N. Pix2Pix Hyperparameter Optimisation Prediction. Procedia Comput Sci. 2024 Mar.;225(C):1009–18. doi:10.1016/j.procs.2023.10.088.
19. Breger A, Biguri A, Landman M S, Selby I, Amberg N, Brunner E, et al. A Study of Why We Need to Reassess Full Reference Image Quality Assessment with Medical Images. J. Imaging Inform. Med. 2025 Mar. doi:10.1007/s10278-025-01462-1.
20. Sara U, Akter M, Uddin M S. Image Quality Assessment through FSIM, SSIM, MSE and PSNR—A Comparative Study. J. Comput. Commun. 2019;07(03):8–18. doi:10.4236/jcc.2019.73002.
21. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative Adversarial Networks. Commun. ACM. 2020 Oct.;63(11):139–44. doi:10.1145/3422622.
22. Behzadpour M, Ghanbari M. Improving Precision of Objective Image/Video Quality Meters. Multimed. Tools Appl. 2023 Jan.;82(3):4465–78. doi:10.1007/s11042-022-13416-8.
23. Rainio O, Teuho J, Klén R. Evaluation Metrics and Statistical Tests for Machine Learning. Sci. Rep. 2024 Mar.;14(1):6086. doi:10.1038/s41598-024-56706-x.