International Journal of Multidisciplinary Studies and Innovative Technologies

e-ISSN: 2602-4888
dergipark.org.tr/en/pub/ijmsit
Research Article

2024, 8 (2), 118 - 126
DOI: 10.36287/ijmsit.8.2.11
Received: November 3, 2024; Accepted: December 11, 2024

Machine Learning and Vision Transformer for CT Scanners' Calibration and Quality Assessment

Khanh Quoc Man[1*], Majeed Soufian[2], Amani Mansour Alsaeedi[3], Jon Fulford [4] and Hairil Abdul Razak [5]

[1*] *Department of Computer Science, University of Exeter, Exeter, UK (km827@exeter.ac.uk) (ORCID: 0009-0003-0565-787X)*
[2] *Department of Computer Science, University of Exeter, Exeter, UK (m.soufian@exeter.ac.uk/magid@ieee.org) (ORCID: 0000-0002-8976-9187)*
[3] *Medical School, University of Exeter, Exeter, UK (amda201@exeter.ac.uk) (ORCID: 0000-0000-0000-0000)*
[4] *Medical School, University of Exeter, Exeter, UK (J.Fulford@exeter.ac.uk) (ORCID: 0000-0002-5945-1688)*
[5] *Medical School, University of Exeter, Exeter, UK (H.Abdul-Razak@exeter.ac.uk) (ORCID: 0000-0002-8266-0381)*

Abstract – In this study, we present the process and research for finding the best machine learning methodology and innovative approach to evaluate the image quality in Computed Tomography (CT) scanners by predicting Signal-to-Noise Ratio (SNR) and Contrast-to-Noise Ratio (CNR) from low-resolution CT images of a series of phantoms. Traditional methods of Image Quality Assessment (IQA), reliant on subjective evaluation by radiologists, often suffer from variability and inefficiency. To address these limitations, we explored both interpretable models like the Adaptive Neuro-Fuzzy Inference System (ANFIS) and other advanced deep learning architectures. Initially, ANFIS combined with Gray Level Co-occurrence Matrix (GLCM) features yielded suboptimal results, with an R-squared value of 0.634. Experimenting with various deep learning methodologies for improving the performance, directed us to develop a hybrid model integrating DenseNet, Vision Transformers, and reparameterization techniques, which showed that can achieve superior results with an R-squared value of 0.8892. This research paper focuses on searching for the optimal machine learning model and lays the groundwork for an automated tool that can optimize imaging protocols by providing a comprehensive quality assessment of CT images in CT calibration.
Keywords *– Machine learning, Deep learning, Vision Transformer, CT calibration, IQA.*

Citation: Man, K., Soufian, M., Alsaeedi, A.M., Fulford, J., Razak, H.A. (2024). Machine Learning and Vision Transformer for CT Scanners' Calibration and Quality Assessment. International Journal of Multidisciplinary Studies and Innovative Technologies, 8(2): 118 - 126.

I. INTRODUCTION

Machine learning and artificial intelligence methodologies have been applied increasingly in various medical fields such as medical imaging and pathogen identifications in recent years, started in 3 decades ago [1 and 2], when very few methodologies existed. The computed tomography machines provide pictorial anatomical information about the physiological state of internal organs by using X- rays and gives sensitive discrimination between healthy and diseased tissue. Ensuring the quality of CT images is essential for accurate medical diagnosis. Naturally calibration is a critical step in this process. In CT imaging, calibration is the first and most crucial step to ensure the reliability and accuracy of images used for diagnosis using an object called "Phantom" to simulate the organ. This includes adjusting the CT scanning equipment to correct any errors that might negatively affect image quality. Such calibration should be conducted regularly to maintain accuracy without distorting the images or reducing their value. Any distortion or lack of proper contrast in CT images can lead to diagnostic errors. To support accurate image analysis and the gathering of diagnostic information, producing high-quality CT images is essential.

There are two main methods in IQA, subjective and objective evaluation [3]. While subjective assessment is conducted by experts, such as diagnostic radiologists, objective assessment, is based on using various logical and mathematical algorithms. The subjective evaluation which has been formed by manual qualitative assessment of CT images by radiologists, usually involves identifying and measuring phantom image features. This process is often considered the gold standard but is limited by poor inter-observer agreement and the risk of fatigue and perceptual biases. At the same time, manual assessment by radiologists, is indeed time-consuming and prone to inconsistencies despite it requires significant expertise. These factors can lead to variations in diagnosis and inefficiencies in the workflow. Especially in CT calibration, to evaluate the quality of a phantom image, radiologists are traditionally required to manually indicate the location of the holes in each square in the phantom image [4]. Such challenges underscore the need for automated methods that can consistently and accurately assess CT image quality, particularly during the calibration process. They particularly highlight the need for developing automated methods based on machine learning and artificial intelligence that can reliably evaluate image quality metrics like SNR and CNR, improving efficiency and reducing variability in CT imaging.

This research focuses on the process of finding an innovative optimal machine learning methodology, which can evaluate SNR and CNR in CT images in the most efficient manner and at the same time can be transparent and interpretable. In contrary to 3 decades ago, a vast number of various machine/deep learning methodologies are available,

which makes it difficult to find an optimal model by using each individual one or by combining them together. We tried to cover a wide range of them to solve the problem of automated measurement of SNR and CNR values in the phantoms' hole images. The data consists of 45,500 holes images cropped from phantom CT images, with labels representing the SNR and CNR values of the images, which were manually assigned. We started our research with a simple model called ANFIS, a learning model commonly used in medical imaging due to its ability to integrate fuzzy systems with neural networks and its transparency and interpretability. Evaluating and recognizing the limitations of traditional methods such as ANFIS, we developed many robust solutions and transitioned to more complex wider deep learning models including ResNet, RNN, SE-ResNet, Fast-ViT, SE-ResNet, Unet-NILM and SqueezeNet in order to assess their performance in predicting SNR and CNR values from CT images. These directed us to introduce a hybrid architecture called SynQ-ViT (Synthetic Quality assessment for computed tomography calibration with Vision Transformer), which leverages a hybrid architecture combining DenseNet, Vision Transformers, and reparameterization techniques. This modification enables the model to effectively learn both local and global features. We reported the success of SynQ-ViT for this application with a view from medical imaging separately [5]. Here in the rest of this paper, after portraying related works, searching for an optimal machine learning model, which fit this application best, will be highlighted by presenting above methodologies in some details with their evaluations and experimental results.

II. Related works

Valdes et al. [6] developed a Virtual IMRT QA framework using a machine learning algorithm that accurately predicted gamma passing rates within 3% across different institutions and measurement techniques. In the studies on using ANFIS in medical imaging, Sharma and Mukharjee [7] utilized ANFIS to classify MR images. The integration of ANFIS using GLCM fuzzy rules in medical imaging provided superior classification accuracy when compared to traditional methods like Fuzzy C-Means (FCM) and K-Nearest Neighbor (K-NN). In early detection of COVID-19 through CT image analysis, with the ANFIS-based model achieving superior performance with an accuracy of 98.63% and rapid testing time [8]. In the study of Bahonar et al. [9] ANFIS model significantly outperforms multiple linear regression (MLR) in predicting breast dose during chest CT scans, with a correlation coefficient R of 0.93 and a Root Mean Square Error (RMSE) of 0.172. These findings suggest that ANFIS offers an accurate and efficient approach to medical imaging, especially in CT images.

In recent advancements within CT imaging quality assurance, a deep learning approach using convolutional neural networks (CNN) has been explored to predict whether CT scans meet the minimal diagnostic image quality threshold. Lee et al. [10] introduce a pre-trained VGG19 network was fine-tuned to analyze a dataset consisting of 74 high-resolution axial CT scans, with image quality rated by a radiologist. The network achieved an accuracy of 0.76 and an AUC of 0.78, highlighting the potential of deep learning methods in assessing and ensuring diagnostic quality in CT imaging, despite challenges posed by the relatively small number of cases. Study of a novel Blind Image Quality Assessment (BIQA) method for low-dose CT images [11], utilized a Denoising Diffusion Probabilistic Model (DDPM) and a transformer-based evaluator. The DDPM is employed to generate high-quality primary content from distorted images, mimicking the human visual system's active inference process, while a transformer-based evaluator predicts image quality by integrating this content with a dissimilarity map. Jensen et al. [12] evaluated the performance of a Deep Learning Image Reconstruction (DLIR) algorithm in contrast-enhanced oncologic abdominal CT, comparing it to the standard 30% Adaptive Statistical Iterative Reconstruction V (ASIR-V). The results demonstrated that DLIR significantly improved image quality, particularly at higher strengths, with a notable 4% reduction in noise and a 92-94% increase in contrast-to-noise ratio compared to 30% ASIR-V.

These studies emphasize the role of machine learning models, particularly deep learning models, in medical imaging and CT quality assurance. However, these studies focus on image evaluation during diagnosis rather than during the CT calibration process. In this paper, we address that gap by using a dataset collected during CT calibration. We aim to develop an objective method based on an optimal machine learning methodology to evaluate the quality of CT images during the preparation phase of a CT machine before it is put into use.

III. AI and Machine Learning methodologies

A. *In search of the optimal model*

The optimal model, which will be used as the most effective method for automated CT image quality assessment and calibration, not only should produce optimal accuracy among all other candidate models but also must have optimal number of parameters, easy and quick to train and implement in real medical working environment. The primary means to evaluate the accuracy of each model, is the Mean Squared Error (MSE) for the error of SNR and CNR, and the coefficient of determination, R^2 (R-squared) as an additional performance metric, which are defined in the next section. The optimal model must also be transparent and interpretable while capturing both local and global image features effectively. In search of such model, first ANFIS was considered.

B. *Basic model*

The main advantage of ANFIS [13] is in its transparent and interpretable architecture (Fig. 1) in the form of fuzzy "if-then" rules, which made it our first choice for this application.

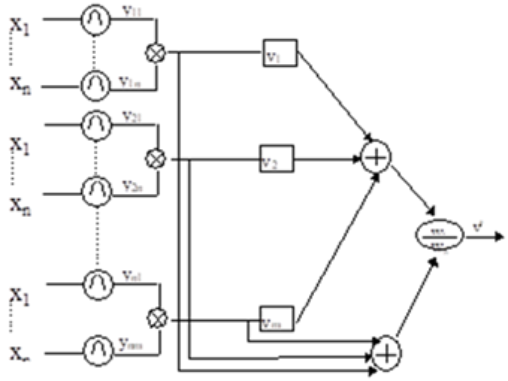

Fig 1. A typical structure of Adaptive Neuro-Fuzzy Inference System with n inputs (X_1 ... X_n) and one output (v') as bock diagram.

ANFIS is a hybrid intelligent system that integrates the benefits of both artificial neural networks from machine learning and fuzzy logic from artificial intelligence, allowing it to model complex, nonlinear relationships effectively. In this study, ANFIS was employed to predict SNR and CNR from phantom hole images, following steps below:

1) Feature Extraction: Before applying the ANFIS model, we first extract features from the CT images using GLCM. GLCM is a statistical method that analyzes the spatial relationships of pixels in an image. It is particularly useful for capturing texture information and is widely used in medical imaging, especially in CT [14]. The GLCM computes how frequently pairs of pixels with specific values and in a specified spatial relationship occur in an image, generating a matrix from which various texture features can be derived. The formula for GCLM feature extraction [15] included:

- Contrast: Measures the intensity contrast between a pixel and its neighbor over the whole image.

$$Contrast = \sum_{i,j} |i-j|^2 \, P(i,j) \tag{1}$$

- Dissimilarity: Similar to contrast but provides a more direct measure of the difference between pairs of pixels.

$$Dissimilarity = \sum_{i,j} |i-j| \, P(i,j) \tag{2}$$

- Homogeneity: Measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal.

$$Contrast = \sum_{i,j} \frac{P(i,j)}{1+|i-j|} \tag{3}$$

- Energy: Provides the sum of squared elements in the GLCM, reflecting image uniformity.

$$Contrast = \sum_{i,j} P(i,j)^2 \tag{4}$$

- Correlation: Measures how correlated a pixel is to its neighbor over the whole image.

$$Contrast = \sum_{i,j} \frac{(i-\mu_i)(j-\mu_j)P(i,j)}{\sigma_i \sigma_j} \tag{5}$$

In above formulations, $P(i,j)$ represents the probability of the co-occurrence of pixel pairs separated by a specific distance and angle, μ_i, μ_j and σ_i, σ_j are the means and standard deviations of the marginal distributions of i and j respectively.

2) Rule Generation and Fuzzy Inference: Once the features are extracted, the ANFIS model processes these features as input data. Each input variable is associated with a fuzzy membership function. ANFIS generates a set of fuzzy if-then rules based on all possible combinations of membership functions across the input variables. For instance, a typical rule might state that "if the contrast is high and the homogeneity is low, then the SNR will be high". The firing strength of each rule is calculated as the product of the membership values for the input variables involved in the rule:

$$w_j = \prod_{i=1}^{n} \mu A_{i,j} \, (x_i \tag{6}$$

where $\mu A_{i,j}$ is the membership value of input x_i in a fuzzy set $A_{i,j}$.

3) Normalization and output: The final output of the ANFIS model is a weighted sum of the normalized firing strengths and the corresponding linear functions of the inputs:

$$y = \sum_{j=1}^{M} \overline{w}_j f_j \, (x) \tag{7}$$

4) Evaluation: Let $\hat{y}_i = (\hat{y}_{i1}, \hat{y}_{i2})$ denote the predicted values for SNR and CNR, and $y_i = (y_{i1}, y_{i2})$ denote the true values. The overall MSE loss function measures the average squared difference between the predicted and true values:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} [(y_{i1} - \hat{y}_{i1})^2 + (y_{i2} - \hat{y}_{i2})^2 \tag{8}$$

where n is the number of samples. The overall R^2 metric assesses the proportion of variance in the dependent variables that is predictable from the independent variables, i.e.:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} [(y_{i1} - \hat{y}_{i1})^2 + (y_{i2} - \hat{y}_{i2})^2)]}{\sum_{i=1}^{n} [(y_{i1} - \bar{y}_1)^2 + (y_{i2} - \bar{y}_2)^2)]} \tag{9}$$

where $\bar{y}_1$ and $\bar{y}_2$ are the mean values of the true SNR and CNR, respectively.

5) Result: After training, ANFIS achieved an MSE of 39.9, indicating a large deviation between the predicted and actual values of the image quality metrics. Additionally, obtained R^2 of 0.634, suggests that the model could only explain 63.4% of the variance in the overall SNR and CNR values, leaving a considerable portion of the variability unaccounted for. While ANFIS provides a valuable framework for clear understanding and transparent modeling relationships in data, its application to the prediction of SNR and CNR in this study has not yielded satisfactory results.

C. Other Advanced Machine and Deep Learning Models

It's possible to improve the ANFIS performance by some clustering methods [16] however, it is anticipated that trying more complex and sophisticated models from a wide range of machine and deep learning approaches would increase the likelihood of achieving a better predictive performance and accuracy in terms of MSE and R^2 metrics, guiding us toward obtaining an optimal model as necessary condition. Although other metrics such as having minimum number of parameters, the ease and speed of training and implementation in real

medical working environments are important and will also be considered for finding the optimal model, the main drawbacks of machine and deep learning models are their lack of transparency and interpretability. To address these issues, a method is employed to visualize and interpret a model's decision-making process without changing its parameters and will be discussed later. A model with optimal MSE, R^2 and other metrics, which fails to highlight important areas such as the Region Of Interest (ROI) in the CT images, will not be considered as the optimal model.

Apart from ANFIS, for the same dataset, seven other models from a wide range of machine and deep learning methodologies were developed, which are explained in next subsections. For successful training and to optimize the performance of our models, we conducted an extensive hyper-parameter optimisation for each model training. The tuning process involved using a Random Search strategy, where 20 trials were executed to explore the hyperparameter space. Early stopping was implemented to monitor the validation loss, with a patience of 10 epochs. If the validation loss did not improve for 5 consecutive epochs, a callback was employed to decrease the Adam optimiser learning rate logarithmically. Further details are presented in the experimental section.

D. ResNet [17]

Residual Networks (ResNet), is a highly influential deep learning architecture that uses skip connections, allowing for the training of very deep networks without the issues of vanishing or exploding gradients. The architecture consists of a series of stacked residual blocks, where each block includes identity mappings and convolutional layers. It is widely used in various medical image tasks [18]. Figure 2 is the idea of ResNet with 1 residual result at each layer. After training, ResNet achieved an overall MSE of 22.01, indicating a much smaller deviation between the predicted and actual values of the image quality metrics with a good overall R^2 of 0.85.

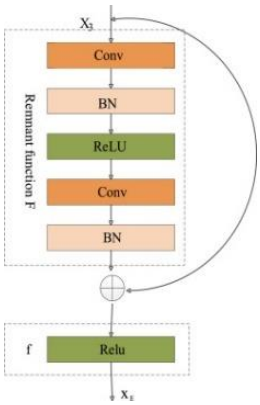

Figure 2. ResNet with 1 residual result at each layer [18]

E. RNN [19]

Recurrent Neural Networks (RNN) are a class of artificial neural networks where connections between nodes form a directed graph along a temporal sequence. The ability of RNN to maintain a memory of previous inputs is due to their feedback loops, which allow information to persist over time, thus enabling the network to capture patterns and dependencies that unfold across long sequences. In tasks related to medical imaging, particularly CT scanners, the RNN have been used as benchmarks for both GANs and deep learning networks [20]. After training, RNN achieved an MSE of 28.40 and a R^2 of 0.81. Figure 3 presents, the results of RNN performance during training for predicting SNR and CNR values from CT images of phantom holes in terms of overall R^2 for both training and validation datasets.

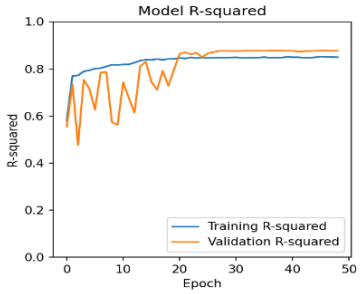

Figure 3. Performance of RNN during training

F. SE-ResNet [21]

This model enhances the traditional ResNet model by incorporating Squeeze-and-Excitation (SE) blocks, hence it is called SE-ResNet. The SE block operates by first applying global average pooling to squeeze global spatial information into a channel descriptor. This descriptor is then passed through a pair of fully connected layers to capture channel-wise dependencies, followed by a sigmoid activation to generate channel weights. It was used in various task of diagnosis from CT images [22]. After training, it achieved an MSE of 17.32 and a R^2 of 0.88. The results of SE-ResNet performance during training are presents in Figure 4 for both training and validation datasets.

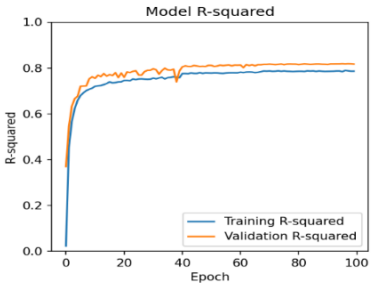

Figure 4. Performance of SE-ResNet during training

G. UNET-NILM [23]

This model leverages a one-dimensional U-NET-based Convolution Neural Networks (CNN) architecture for Non-Intrusive Load Monitoring (NILM), hence called UNET-NILM. It enables simultaneous appliance state detection and power consumption estimation [23]. By combining down-sampling and up-sampling blocks, it captures both local and global features of power signals effectively [24]. Figure 5 illustrating the structure of UNET-NILM with the idea of utilizing a U-Net architecture, originally designed for image segmentation, to effectively separate and identify individual electrical appliances' power usage from aggregated energy consumption data. By leveraging the encoder-decoder structure of U-Net, the model learns both local and global features, making it well-suited for accurately disaggregating energy signals at various levels of granularity. After training, UNET-NILM achieved an MSE of 20.24 and a R^2 of 0.86.

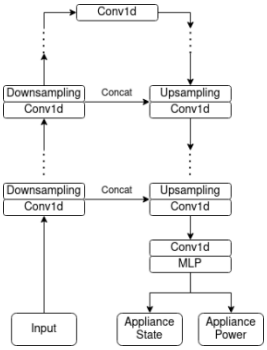

Figure 5. Architecture of UNetNILM [24]

H. SqueezeNet [25]

The model designed to achieve AlexNet-level accuracy on ImageNet with 50 times fewer parameters, reducing the model size to less than 0.5MB. The architecture accomplishes this through the use of "Fire modules," which combine 1x1 and 3x3 filters and late down-sampling to maximize accuracy while minimizing parameter count. This makes processing low-resolution images efficient. The efficiency of SqueezeNet for low-resolution medical images proved by Zhang et al. [26]. After training, SqueezeNet achieved an MSE of 17.55 and a R^2 of 0.88. Figure 6 presents, the results of SqueezeNet performance during training for both training and validation datasets.

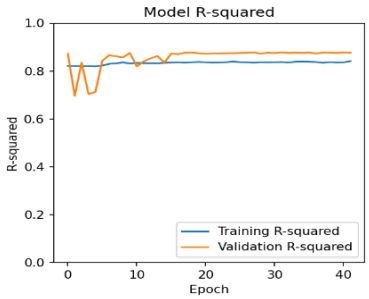

Figure 6. Performance of SqueeNet during training.

I. FastViT [27]

Fast Vision Transformer (FastViT) is a hybrid vision transformer architecture that combines the efficiency of CNN with the global context modeling capabilities of transformers. The key innovation of FastViT lies in its use of the RepMixer block, a novel token-mixing operator that employs structural parameterization and capacity to learn complex patterns. After training, FastViT achieved an MSE of 18.51 and a R^2 of 0.87.

J. The optimal model

The above developments for ample number of the CT input images with small sizes (6 to 9 pixels), imposed a model architecture that performs well with a low number of input parameters. It means the model must be able to capture both global and local features in large volume of datasets, which was also suggested by Talab et al. [28] for low-resolution images. As a result, we proposed SynQ-ViT, as optimal model that focuses on learning both local and global features linearly. For local features, we employed dense blocks from DenseNet [29] in it due to their ability to effectively learn through feature reuse. Adding attention mechanism from Vision Transformers (ViT) aids the model to capture long-range dependencies and contextual information across the entire images [30] or global features. RepMixer [27] was also introduced in our model for achieving efficient token mixing, reducing computational overhead through structural reparameterization, and enhancing the model's capacity to learn complex patterns. These are illustrated in Figure 7 showing SynQ-ViT model in train-inference phases with the dense block (Fig. 7a), transition block (Fig. 7b), RepMixer block (Fig 7c.1 in training mode and Fig 7c.2 in inference mode) and attention block (Fig 7d). These have also been explained in our other study [5] in some details. The model learns its parameters from data by passing outputs sequentially through its layers during training. During inference, by structural reparameterization the Batch Norm and skip connections are removed for simplifying RepMixer structure. This reduces computational overhead and memory access costs as shown by Weng et al. [31] that removing skip connections can improve computational efficiency and reduce resource requirements without significantly compromising accuracy. These are important for real-time applications and hardware deployments, especially when processing large volumes of CT images. The output layer is designed to predict SNR and CNR values by using a dense layer to transform the final feature representations into them. After training, it achieved an MSE of 16.03 and a R^2 of 0.89. Figure 8 presents, the results of SynQ-ViT performance during training for both training and validation datasets.

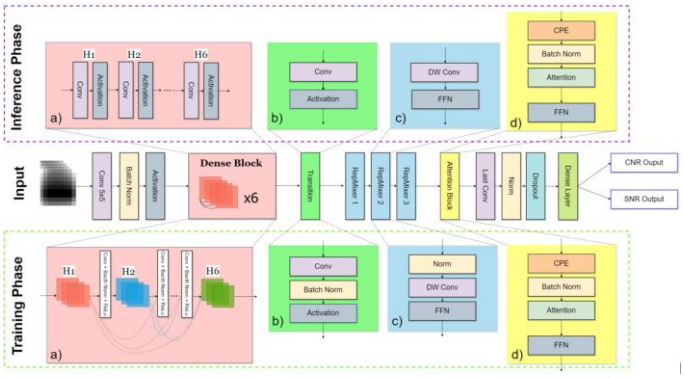

Fig 7. The optimal model architecture SynQ-ViT in training and inference phases: a) dense block, designed to maximize feature reuse by directly connecting each layer using H functions, which improves information flow and supports efficient learning with fewer parameters. b) transition block, reduces dimensionality, aiming to cut down computational overhead while retaining essential features for further processing. c) RepMixer block, plays a crucial role in optimizing the model's structure; during training, it incorporates skip connections for performance, but in the inference phase it removes both batch normalization and skip connections to reduce memory and computational costs. This structural reparameterization makes the model more efficient in real-time applications. d) attention block is used for token mixing, focusing on the most relevant parts of the input datasets and ensuring that the model captures key dependencies and patterns across the entire image.

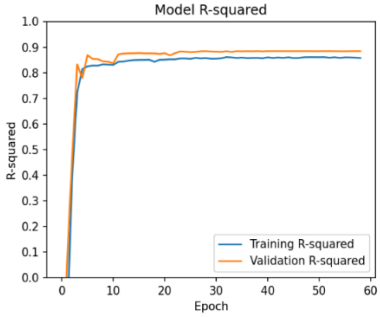

Fig 8. Performance of SynQ-ViT during training [5]

IV. Evaluation and Experimental Results

A. The Datasets

As stated earlier, the data acquired in this study consists of 45,500 holes images cropped from specifically designed Perspex phantom 500 CT images, with labels representing the SNR and CNR values that were manually calculated. The phantom was injected with AuNPs (0.005mg/ml) and scanned using a CT scanner (Biograph Vision 600 Siemens Definition Edge 128) under a variety of exposure settings to rigorously evaluate image quality metrics [5]. Figure 9 shows three randomly selected images from the dataset, which resized to 9x9 pixels to standardize the input dimensions. The presence of negative values in the SNR and CNR during data acquisition, led to implementing a filtering process to ensure their removal and consequence normalization to preserve the accuracy and reliability of the datasets.

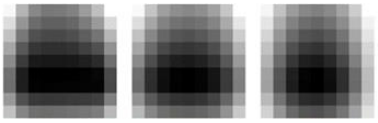

Fig 9. Three random sample images from the dataset in TIFF format. Each of these images represents a hole in phantom images [5].

B. The experiment design

After data acquisition and obtaining accurate and reliable datasets and exploring them, many models from a wide range of machine and deep learning methodologies were exploited to cover all possible models that satisfy the success criteria established in the sections III.A and III.C and to guide us toward an optimal model. The training and hyperparameter optimisation performed as mentioned in the section III.C If required in some models, the number of blocks within each stage was tuned between 1 and 4 with a growth rate between 12 and 48, the layer scale parameter was also adjusted logarithmically between 10^{-6} to10^{-4}, and both dropout and drop connect rates were varied between 0.0 and 0.5. After training, the performance of each model for predicting SNR and CNR values from CT images of phantom holes in terms of overall MSE and R^2 metrics were calculated and discussed in sections III.D to III.J, which are summarized in figure 10.

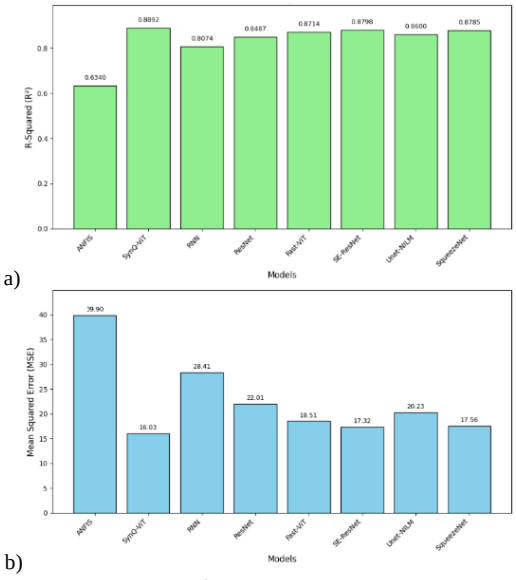

Fig 10. The comparison of a) R^2 and b) MSE performance across all models.

Considering only overall MSE and R^2 metrics, SynQ-ViT was suggested as the optimal model in section III.J. However, other success criteria and metrics are also important for evaluating each model in other to confirm the optimal one for this application as presented in the next section.

C. Models Evaluation

This section discusses also other success criteria and optimality conditions such as having highest speed, minimum number of parameters, the ease in development and implementation in real medical working environments. These can collectively be considered as working experience with each model and are shown in Table 1, including metrics such as epochs, and training time. Please note that one can roughly state that the number of parameters is inversely proportional to the speed of model in inference phase, i.e., a smaller number of parameters is a metric of desired implementation and ease of use in real medical working environments.

Table 1. Experience of working with each model.

Architecture	Epochs	number of Parameters	Training time
ANFIS	10	-	1 hour
SynQ-ViT	61	145,902	~ 1 hours
RNN	64	334,082	~ 1 hour
Resnet	87	118,002	~ 20 mins
Fast-ViT	36	3,210,530	~ 8 hours
SE-ResNet	44	8,672,624	~ 1 day
Unet-NILM	25	2,032,578	~ 6 hours
SqueezeNet	32	113,026	20 mins

- Comparative Analysis for Predicative Performance:

To validate that SynQ-ViT offers an optimal solution among the other AI and machine/deep learning models developed in this study, we conducted a thorough comparative analysis. For this purpose, each model is considered to serve as a benchmark to evaluate the effectiveness of SynQ-ViT in predicting image quality metrics, ensuring it delivers superior and reliable results. Figures 10 already showed SynQ-ViT achieved an impressive R^2 value of 0.89 and an MSE of 16.03 after 61 epochs on the validation set, with a convergence occurring after the 5th epoch (Figure 8). While complex models like SE-ResNet, FastViT and Unet-NILM also demonstrated robust performance, they did not surpass SynQ-ViT in terms of R^2 and MSE. On the other hand, among simpler models only SqueezeNet showed a close predictive power to SynQ-ViT while Resnet and RNN were found to be less suitable for this task, exhibiting notably lower performance metrics.

- Comparative Analysis for Working Experience:

From working experience with each model point of view, while models like FastViT and SE-ResNet performed well indicate strong predictive performance, these models require significant computational resources, with largest parameter counts of 3,210,530 and 8,672,624, and longest training time of approximately 8 hours and one day respectively. This could be inefficient, challenging and disadvantageous when developing and implementing a system that needs to handle large input volumes and operate in real time in medical working environments. On the other hand, models with smaller parameter counts, such as ResNet and RNN (with

118,002 and 334,082 parameters and training time of approximately 20 minutes and 1hour respectively), did not perform well. UNET-NILM demonstrated average performance, with neither parameter count nor efficiency standing out significantly.

Among the models compared, only SqueezeNet approached the performance of SynQ-ViT, with much lower parameter size (113,026) and 3 times faster training time (approximately 20 minutes) than SynQ-ViT, could claim the optimal model title. This required further detailed analysis of the performance during training and validation results. Notably, SynQ-ViT, converged around the 5th epoch (blue line in Figure 8) producing a stable validation result quickly (orange line in Figure 8) while that is not the case with SqueezeNet (orange line in Figure 6). Indeed, compared to the training results of all models, SynQ-ViT has achieved the highest stability and convergence, offering top working experience performance among the best predictive models, yet it requires to pass the last evaluation below before establishing itself as the optimal model for this application.

- Transparency and Interpretability:

As mentioned earlier, any candidate model that fails transparency and interpretability criteria, will not be considered as the optimal model. The transparency and interpretability will be evaluated by the model's ability to highlight important areas such as the ROI for the CT images in this application. For this purpose, a method called Gradient-weighted Class Activation Mapping (Grad-CAM) by Selvaraju et al. [32] has been utilized to visualize and interpret a model's decision-making process without changing its parameters. The Grad-CAM addresses the lack of transparency and interpretability by producing heatmaps and highlighting the regions in the heatmaps that the model relied on for its predictions. In order to do so, Grad-CAM uses randomly selected images and the corresponding final layers of the model.

We applied Grad-CAM to the final convolutional layers of SynQ-ViT as shown in Figure 11, to verify and visualize the areas of focus when SynQ-ViT making its decision. Figure 12 illustrates the heatmaps generated by Grad-CAM from randomly selected images and the corresponding final layers of SynQ-ViT, highlighting the regions that SynQ-ViT relied on for its predictions. Analysis of these heatmaps determine whether SynQ-ViT has been focusing on important areas such as the ROI or not. The red areas in the heatmap correspond to the ROI in the CT images. These visualizations confirm that this model accurately focuses on the critical areas of the input images, thereby validating its effectiveness and reliability in predicting SNR and CNR values as the optimal model.

V. Conclusion, Discussion and Further Work

In this paper we presented our research journey for finding an innovative optimal machine/deep learning methodology, which can evaluate SNR and CNR in CT images in the most efficient manner and at the same time can be transparent and interpretable. After acquiring the required datasets and defining the domain challenge in some details, main success criteria and metrics for achieving optimal predictive performance and working experience were defined. Considering the importance of acquired datasets in the training and evaluation phases, a rigorous preprocessing phase was implemented for ensuring the uniformity, fidelity, accuracy and reliability of the datasets by applying appropriate filtering, normalization, standardization and other procedures for image data in TIFF format, along with the associated SNR and CNR target values.

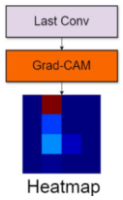

Fig 11. Showing how Grad-CAM was applied to the final convolutional layers in Figure 7 just before the model makes its prediction.

Based on the acquired datasets and the nature of the domain dynamics, eight models from a wide range of AI, machine and deep learning methodologies were consider as optimal candidate models. For successful training and to optimize the performance of our models, we conducted extensive hyperparameter optimisation experiments for each model training. The first choice in choosing an optimal model was the ANFIS model due to its inherent transparency interpretability and explainability properties. The ANFIS model also proved the optimal choice for working experience metrics because of its fast training (10 epochs in an hour with a rapider convergence) and smaller parameter set but underperformed predictively with an R-squared value of just 0.63, indicating its inadequacy for this application. These experiments revealed that each model has its own advantages and limitations when applied to CT image quality assessment. For instance, models like SE-ResNet and FastViT, despite their competitive R-squared values, require high computational resources, making them less feasible for real-time applications. On the other hand, simpler models such as RNN showed limited predicative performance, as reflected by their lower R-squared values. While models with smaller parameter counts like SqueezeNet and Unet-NILM came close to SynQ-ViT's predictive performance, they did not achieve the same level of stability and efficiency. This highlights the importance of balancing model complexity and efficiency, and SynQ-ViT demonstrates an optimal combination, achieving highest accuracy and stability with lower computational demands. Please note that although the model converged quite early at the 5th epoch, it was able to further reduce the error and that is why training continued until epoch 61. Finally, analyzing heatmaps created by Grad-CAM validated the interpretability of the SynQ-ViT's predications.

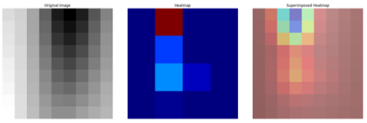

Fig 12. Example of Grad-CAM created heatmaps from a randomly selected original image (left). The middle heatmap indicating the red areas that the model focuses on to predict SNR and CNR, and the right image is the heatmap superimposed on the original image. highlighting the regions the model relied on for its predictions.

There is a discussion about how the optimal model obtained its predictive power and if it is possible to achieve higher

predictive performance while keeping other metrics at optimal. We introduced SynQ-ViT, as a model that combines DenseNet, attention mechanisms, and reparameterization techniques to efficiently learn both local and global features from CT images of phantom holes. The predictive performance and early convergence with low-resolution input images can be attributed to SynQ-ViT's architecture, which effectively reuses learned features from previous layers, allowing for effective feature extraction and model optimization. However, the question on the possibility of increasing its predictive performance, say a R^2 value of above 0.95, is an open research challenge. When compared to other advanced models, SynQ-ViT consistently achieved superior accuracy while maintaining a lower parameter count, demonstrating its efficiency, particularly in real-time applications involving large datasets. Its rapid convergence and ability to handle resource constraints make it an ideal candidate for clinical deployment, where timely processing is critical.

Predicting SNR and CNR for each hole in the phantom image is a crucial first step toward creating a robust quality assessment tool for CT calibration. Moving forward, our goal is to expand this model into a comprehensive system that aggregates these predictions to provide a holistic quality evaluation of the entire phantom image. This tool would help automate imaging protocol optimization in clinical settings, advancing medical imaging and improving patient care.

ACKNOWLEDGMENT

The first author, Mr Khanh Quoc Man, would like to acknowledge Dr Mohammed M. Abdelsamea from Department of Computer Science at University of Exeter, for introducing him to the team.

Authors' Contributions

The authors' contributions to the paper are equal.

Statement of Conflicts of Interest

There is no conflict of interest between the authors.

Statement of Research and Publication Ethics

The authors declare that this study complies with Research and Publication Ethics

REFERENCES

[1] Soufian, M., Robinson F.V.P., and Soufian M.,1996, Fuzzy Logic Controller for Whole Body NMR imaging. IEE Colloquium on Fuzzy Logic Controllers in Practice (Digest No. 96/200). London, U.K, DOI: 10.1049/ic:19961125.

[2] Bright J. J., Claydon M. A., Soufian M., and Gordon D. B., 2002, Rapid typing of bacteria using Matrix-Assisted Laser Desorption Ionisation Time-of-Flight Mass Spectrometry and Pattern Recognition Software, Journal of Microbiological Methods, Vol. 48, Issue 2-3, pp 127-138, https://doi.org/10.1016/S0167-7012(01)00317-7 . PMID:11777563

[3] Kumar, R. and Rattan, M., 2012. Analysis of various quality metrics for medical image processing. *International Journal of Advanced Research in Computer Science and Software Engineering*, *2*(11), pp.137-144.

[4] Wang, C.L., Wang, C.M., Chan, Y.K. and Chen, R.T., 2012. Image-quality figure evaluator based on contrast-detail phantom in radiography. *The International Journal of Medical Robotics and Computer Assisted Surgery*, *8*(2), pp.169-177.

[5] Khanh, Q. M., Alsaeedi, A. M., Soufian, M., Fulford, J. and Razak, A. H., 2024. SynQ-ViT: Synthetic Image Quality Assessment for CT Calibration with Vision Transformer, Submitted to IEEE-Engineering in Medicine & Biology Society Conference on Biomedical Engineering and Science (IECBES2024), Penang, Malaysia.

[6] Valdes, G., Scheuermann, R., Hung, C.Y., Olszanski, A., Bellerive, M. and Solberg, T.D., 2016. A mathematical framework for virtual IMRT QA using machine learning. Medical physics, 43(7), pp.4323-4334.

[7] Sharma, M. and Mukharjee, S., 2012. Artificial neural network fuzzy inference system (ANFIS) for brain tumor detection. arXiv preprint arXiv:1212.0059, pp.1-5.R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, "High-speed digital-to-RF converter," U.S. Patent 5 668 842, Sept. 16, 1997.

[8] Hossam, A., Fawzy, A., Elnaghi, B.E. and Magdy, A., 2022. An intelligent model for rapid diagnosis of patients with COVID-19 based on ANFIS. In Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2021 (pp. 338-355). Springer International Publishing.

[9] Bahonar, B.M., Changizi, V., Ebrahiminia, A. and Baradaran, S., 2023. Prediction of breast dose in chest CT examinations using adaptive neuro-fuzzy inference system (ANFIS). *Physical and Engineering Sciences in Medicine*, *46*(3), pp.1071-1080.

[10] Lee, J.H., Grant, B.R., Chung, J.H., Reiser, I. and Giger, M., 2018, March. Assessment of diagnostic image quality of computed tomography (CT) images of the lung using deep learning. In *Medical Imaging 2018: Physics of Medical Imaging* (Vol. 10573, pp. 399-405). SPIE.

[11] Shi, Y., Xia, W., Wang, G. and Mou, X., 2024. Blind ct image quality assessment using ddpm-derived content and transformer-based evaluator. *IEEE Transactions on Medical Imaging*.

[12] Jensen, C.T., Liu, X., Tamm, E.P., Chandler, A.G., Sun, J., Morani, A.C., Javadi, S. and Wagner-Bartak, N.A., 2020. Image quality assessment of abdominal CT by use of new deep learning image reconstruction: initial experience. *American Journal of Roentgenology*, *215*(1), pp.50-57.

[13] Jang, J.S., 1993. ANFIS: adaptive-network-based fuzzy inference system. *IEEE transactions on systems, man, and cybernetics*, *23*(3), pp.665-685.

[14] Korchiyne, R., Farssi, S.M., Sbihi, A., Touahni, R. and Alaoui, M.T., 2014. A combined method of fractal and GLCM features for MRI and CT scan images classification. *arXiv preprint arXiv:1409.4559*.

[15] Ramamurthy, B. and Chandran, K.R., 2012. Content based medical image retrieval with texture content using gray level co-occurrence matrix and k-means clustering algorithms. *Journal of Computer Science*, *8*(7), p.1070.

[16] Soufian M., Molaei M., and Nefti S., 2017, Adaptive clustering based inclusion and computational intelligence for fed-batch fermentation process control. In IEEE Development in eSystem Engineering (DeSE), Paris, France. DOI: 10.1109/DeSE.2017.45.

[17] Targ, S., Almeida, D. and Lyman, K., 2016. Resnet in resnet: Generalizing residual architectures. *arXiv preprint arXiv:1603.08029*.

[18] Xu, W., Fu, Y.L. and Zhu, D., 2023. ResNet and its application to medical image processing: Research progress and challenges. *Computer Methods and Programs in Biomedicine*, *240*, p.107660.

[19] Grossberg, S., 2013. Recurrent neural networks. *Scholarpedia*, *8*(2), p.1888.

[20] Zhang, H. and Qie, Y., 2023. Applying deep learning to medical imaging: a review. *Applied Sciences*, *13*(18), p.10521.

[21] Thiruppathi, K., Selvakumar, K. and Shenbagavel, V., 2023. SE-RESNET: Monkeypox Detection Model. *International Journal of Advanced Computer Science and Applications*, *14*(9).

[22] Abdelrahman, A. and Viriri, S., 2023. FPN-SE-ResNet model for accurate diagnosis of kidney tumors using CT images. *Applied Sciences*, *13*(17), p.9802.

[23] Faustine, A., Pereira, L., Bousbiat, H. and Kulkarni, S., 2020, November. UNet-NILM: A deep neural network for multi-tasks appliances state detection and power estimation in NILM. In *Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring* (pp. 84-88).

[24] Virtsionis Gkalinikis, N., Nalmpantis, C. and Vrakas, D., 2023. Variational regression for multi-target energy disaggregation. *Sensors*, *23*(4), p.2051.

[25] Iandola, F.N., 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. *arXiv preprint arXiv:1602.07360*.

[26] Zhang, W., Li, J. and Qiu, X., 2019, December. SAR image super-resolution using deep residual SqueezeNet. In *Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing* (pp. 1-5).

[27] Vasu, P.K.A., Gabriel, J., Zhu, J., Tuzel, O. and Ranjan, A., 2023. FastViT: A fast hybrid vision transformer using structural

reparameterization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 5785-5795).

[28] Talab, M.A., Awang, S. and Ansari, M.D., 2020. A Novel Statistical Feature Analysis-Based Global and Local Method for Face Recognition. *International Journal of Optics*, *2020*(1), p.4967034.

[29] Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q., 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700-4708).

[30] Li, Y., Wang, J., Dai, X., Wang, L., Yeh, C.C.M., Zheng, Y., Zhang, W. and Ma, K.L., 2023. How does attention work in vision transformers? A visual analytics attempt. *IEEE transactions on visualization and computer graphics*, *29*(6), pp.2888-2900.

[31] Weng, O., Marcano, G., Loncar, V., Khodamoradi, A., Sheybani, N., Meza, A., Koushanfar, F., Denolf, K., Duarte, J.M. and Kastner, R., 2024. Tailor: Altering skip connections for resource-efficient inference. *ACM Transactions on Reconfigurable Technology and Systems*, *17*(1), pp.1-23.

[32] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and 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).