https://www.scirp.org/journal/jbise
J. Biomedical Science and Engineering,
2024, Vol. 17, (No. 9), pp: 157-170

# Performance Comparison of Vision Transformer- and CNN-Based Image Classification Using Cross Entropy: A Preliminary Application to Lung Cancer Discrimination from CT Images

**Eri Matsuyama[1], Haruyuki Watanabe[2], Noriyuki Takahashi[3]**

[1]Faculty of Informatics, The University of Fukuchiyama, Kyoto, Japan; [2]School of Radiological Technology, Gunma Prefectural College of Health Sciences, Gunma, Japan; [3]School of Health Sciences, Fukushima Medical University, Fukushima, Japan

**Correspondence to:** Eri Matsuyama, matsuyama-eri@fukuchiyama.ac.jp
**Keywords:** Lung Cancer Classification, Vision Transformers, Convolutional Neural Networks, Cross Entropy, Deep Learning
**Received:** August 26, 2024 **Accepted:** September 23, 2024 **Published:** September 26, 2024



## ABSTRACT

This study evaluates the performance and reliability of a vision transformer (ViT) compared to convolutional neural networks (CNNs) using the ResNet50 model in classifying lung cancer from CT images into four categories: lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), large cell carcinoma (LULC), and normal. Although CNNs have made significant advancements in medical imaging, their limited capacity to capture long-range dependencies has led to the exploration of ViTs, which leverage self-attention mechanisms for a more comprehensive global understanding of images. The study utilized a dataset of 748 lung CT images to train both models with standardized input sizes, assessing their performance through conventional metrics—accuracy, precision, recall, F1 score, specificity, and AUC—as well as cross entropy, a novel metric for evaluating prediction uncertainty. Both models achieved similar accuracy rates (95%), with ViT demonstrating a slight edge over ResNet50 in precision and F1 scores for specific classes. However, ResNet50 exhibited higher recall for LULC, indicating fewer missed cases. Cross entropy analysis showed that the ViT model had lower average uncertainty, particularly in the LUAD, Normal, and LUSC classes, compared to ResNet50. This finding suggests that ViT predictions are generally more reliable, though ResNet50 performed better for LULC. The study underscores that accuracy alone is insufficient for model comparison, as cross entropy offers deeper insights into the reliability and confidence of model predictions. The results highlight the importance of incorporating cross entropy alongside traditional metrics for a more comprehensive evaluation of deep learning models in medical image

**classification, providing a nuanced understanding of their performance and reliability. While the ViT outperformed the CNN-based ResNet50 in lung cancer classification based on cross-entropy values, the performance differences were minor and may not hold clinical significance. Therefore, it may be premature to consider replacing CNNs with ViTs in this specific application.**

## 1. INTRODUCTION

Lung cancer is the leading cause of cancer-related deaths globally, with 2.5 million new cases diagnosed in 2022 [1]. Lung cancer is divided into small cell lung cancer and non-small cell lung cancer (NSCLC). NSCLC includes subtypes such as lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), and large cell carcinoma (LULC). LUAD accounts for 85% of NSCLC cases, and patients often face challenges like drug resistance and recurrence. LUSC, which is linked to smoking, has a high mutation rate and genomic complexity. LULC has a molecular profile more similar to adenocarcinoma than squamous cell carcinoma and generally has a poorer prognosis than other NSCLC types. Identifying the histological type early is crucial for effective treatment and reducing mortality. Diagnosing lung cancer remains a significant challenge, primarily due to the asymptomatic nature of the disease in its early stages and the inherent difficulty in distinguishing between benign and malignant lesions. Although computed tomography (CT) screening, particularly with its high sensitivity and detailed imaging capabilities, holds promise, its widespread effectiveness is limited by challenges such as high false-positive rates and disparities in access to screening. These limitations underscore the need for ongoing research and innovation in diagnostic techniques to enhance early detection and minimize unnecessary interventions.

CT screening is useful for early lung cancer detection. However, advancements in CT technology have led to the detection of many microscopic nodules, increasing radiologists' workload [2, 3]. The primary features of CT screening include its high sensitivity, detailed imaging capabilities, rapid processing, and associated radiation exposure. While CT screening is essential for diagnosing lung cancer, particularly in high-risk populations, it is crucial to balance the benefits of early detection with the risks of false positives and radiation exposure. Ongoing advancements in imaging technologies, coupled with the development of artificial intelligence (AI)-based diagnostic tools, are expected to address these challenges in the future. Computer-aided diagnosis (CAD) systems can assist in easing this burden. While chest CT CAD research has shown promise, challenges remain, such as a higher rate of false positives compared to physicians and limitations in improving system accuracy.

In recent years, AI techniques, particularly deep learning models, have become essential in automating image processing and have been increasingly recognized in the domain of medical imaging [4, 5]. Convolutional neural networks (CNNs), a key AI technology, have brought about a transformation in medical imaging by effectively learning complex patterns, enabling the automated identification of diseases and abnormalities [6-9]. These advancements have led to significant improvements in a variety of medical imaging applications and modalities [10-12]. The typical architecture of CNNs consists of three key components: convolutional layers, pooling layers, and fully connected layers. Upon receiving an input image, the convolutional layers extract features from it. Next, the pooling layers, often using max-pooling, reduce the size of the feature maps. Finally, fully connected layers, stacked at the end of the network, perform classification using a specific function, such as SoftMax [13-17]. Despite their impressive performance, CNNs have inherent limitations and cannot naturally model explicit long-distance dependencies because of the constrained receptive field of their convolutional kernels [18].

Inspired by the remarkable success of transformer architectures in natural language processing, these techniques have become widely used in modern computer vision models. Since the introduction of vision transformers (ViTs), transformers have proven to be effective alternatives to CNNs for a variety of tasks, including image recognition, object detection, image segmentation, and image classification [18-21]. One significant theoretical advantage of ViT models over CNNs is their use of a self-attention mechanism, which allows them to gain a global understanding of an image rather than merely focusing on local features, as

CNNs do.

ViTs have several advantages, including the ability to capture long-range dependencies, adapt to different input sizes, and enable parallel processing, making them well-suited for image-related tasks. However, ViTs also face challenges, such as high computational demands, large model sizes, scalability issues with large datasets, interpretability, and generalization performance. These factors underscore the importance of comparing ViTs with established CNN models [22].

Recent studies have reported comparisons of the performance evaluation of Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) in medical image classification across different modalities [23-32]. Among these studies, some have focused on lung cancer. Fanizzi *et al.* [25] compared ViTs and CNNs for predicting the recurrence of non-small cell lung cancer. Based on their preliminary experimental results, they concluded that ViTs do not contribute to improving predictive performance for the addressed problem. Additionally, Gai *et al.* [26] conducted a comparative study on the performance of ViTs and CNNs for the automatic identification of lung cancer using a dataset of medical images. Their findings revealed that CNNs are more effective than ViTs when the dataset size is insufficient. In these two studies, commonly used evaluation metrics such as accuracy, specificity, precision, recall (or sensitivity), F1 score, and the area under the receiver operating characteristic curve (AUC) were employed to assess the performance of ViT and CNN models.

However, existing evaluation metrics have limitations, such as a lack of transparency in CNN inferences and difficulty in estimating uncertainty in the results. Specifically, issues like uncertainty arising from outlier data, overconfident predictions, and covariate shift can potentially affect the reliability of the models. In image classification tasks using ViTs and CNNs, the softmax function is commonly employed to represent the output as a probability value. However, these outputs are often poorly calibrated, making it challenging to directly interpret the model's predictions as probabilistic measures. Evaluating model uncertainty is crucial for enhancing the transparency and reliability of predictions, improving data quality, and minimizing misjudgments. Methods such as Bayesian Neural Networks and Monte Carlo Dropout are widely recognized for estimating uncertainty in neural networks. However, these approaches face the challenge of excessive computational costs when applied to ViTs and CNNs.

In this study, we propose the inclusion of cross entropy, commonly used as a cost function during the training of CNN models, as an additional metric to quantify uncertainty. This metric will be used alongside traditional evaluation matrices to specifically compare the classification performance of ViT and CNN models in detecting lung cancer in CT images.

## 2. MATERIALS AND METHODS

### 2.1. Dataset

In our study, we utilized a publicly available dataset provided by the research community for non-profit purposes [33]. The dataset consists of 748 lung CT images, with 187 images each for LUAD, LULC, LUSC, and normal cases. Since all data were in the public domain and did not involve human patients, there were no ethical concerns, and obtaining informed consent was not required. An illustration of the image data is presented in **Figure 1**.

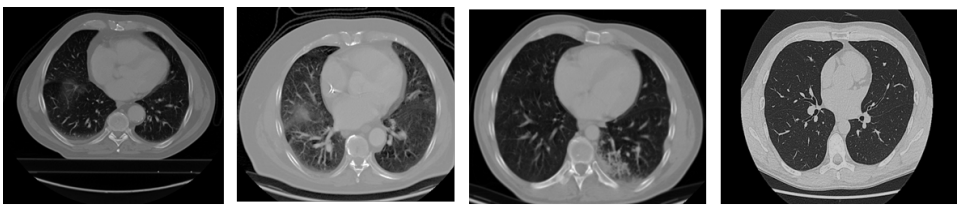


**Figure 1. An example of image data. (a) LUAD (adenocarcinoma); (b) LULC (large cell carcinoma); (c) LUSC (squamous cell carcinoma); (d) Normal.**

## 2.2. Model Architecture

In recent years, many deep CNN architectures, such as VGG [34], GoogLeNet [35], and ResNet [36], have demonstrated excellent performance in image classification tasks. In this study, we chose ResNet50 as the network model. This choice is based on our use of the same network in a previous study [37], which facilitates a direct comparison of results with ViT.

### 2.2.1. CNN-Based ResNet

Residual networks (ResNets), introduced by He *et al.* [36], represent a significant advancement in deep convolutional neural network (CNN) architectures. Among these, ResNet50 is particularly well-known. It comprises 16 residual blocks, each containing several convolutional layers with residual connections, as well as pooling layers, fully connected layers, and a softmax output layer for classification.

In this study, we selected the CNN-based ResNet50, which is widely used in medical imaging, and conducted learning through fine-tuning. Specifically, we utilized a pre-trained ResNet50 model initially trained on natural images and retrained the entire network using lung CT images. During fine-tuning, we did not freeze any layers, allowing all weights to be updated. A four-class classification was performed by replacing and training the final fully connected layer and the last classification layer with new configurations suited to the number of categories.

To conform to ResNet50's structural requirements, the input data size was standardized to 224 × 224 pixels using bicubic interpolation. The mini-batch size was set to 10, and the Adam optimizer (which combines momentum SGD and RMSprop) was employed. During retraining with CT images, parameters were adjusted such that the learning rate increased in the newly replaced fully connected layer, decreased in the transfer layers, and decreased after every 5 epochs.

To prevent overfitting, an L2 regularization term was incorporated into the loss function. The number of epochs was determined by evaluating validation accuracy after each iteration. Retraining was halted if the accuracy did not surpass the highest accuracy achieved in the last 5 consecutive validations.

### 2.2.2. Vision Transformer

In this study, we classify lung CT images into "LUAD," "LULC," "LUSC," and "normal" categories using the vision transformer (ViT) model. Specifically, we adopt the B-16 variant of the ViT model without modifications. This variant comprises 12 stacked transformer encoder blocks and uses a patch size of 16 × 16. The overall architecture is illustrated in **Figure 2**, and the details of the network are as follows.

The ViT processes image data by dividing it into small patches. The initial layer of the network is the patch encoder, which transforms the input image into multiple flattened patches. To retain structural and spatial information, positional embeddings are then added to these patches. This sequence is supplemented with a [class] embedding and fed into the transformer encoder.

The transformer encoder, following the architecture proposed by Vaswani *et al.* [38], consists of multi-headed self-attention layers and multilayer perceptron blocks, with layer normalization applied before each block. This normalization helps reduce training time and improve generalization performance. The encoder outputs feature vectors corresponding to the input patches. As with standard methods, we use the first feature vector related to the [class] embedding to represent the entire sequence.

Finally, a learnable linear layer processes this feature vector, generating a binary output vector, which is then passed to the softmax activation function.

## 2.3. Evaluation Metrics

In this work, in addition to the commonly used standard metrics, we employ cross entropy as a performance evaluation metric to quantify uncertainty in image classifiers.

### 2.3.1. Standard Measures

The performance of classification models is typically assessed using standard metrics. A confusion

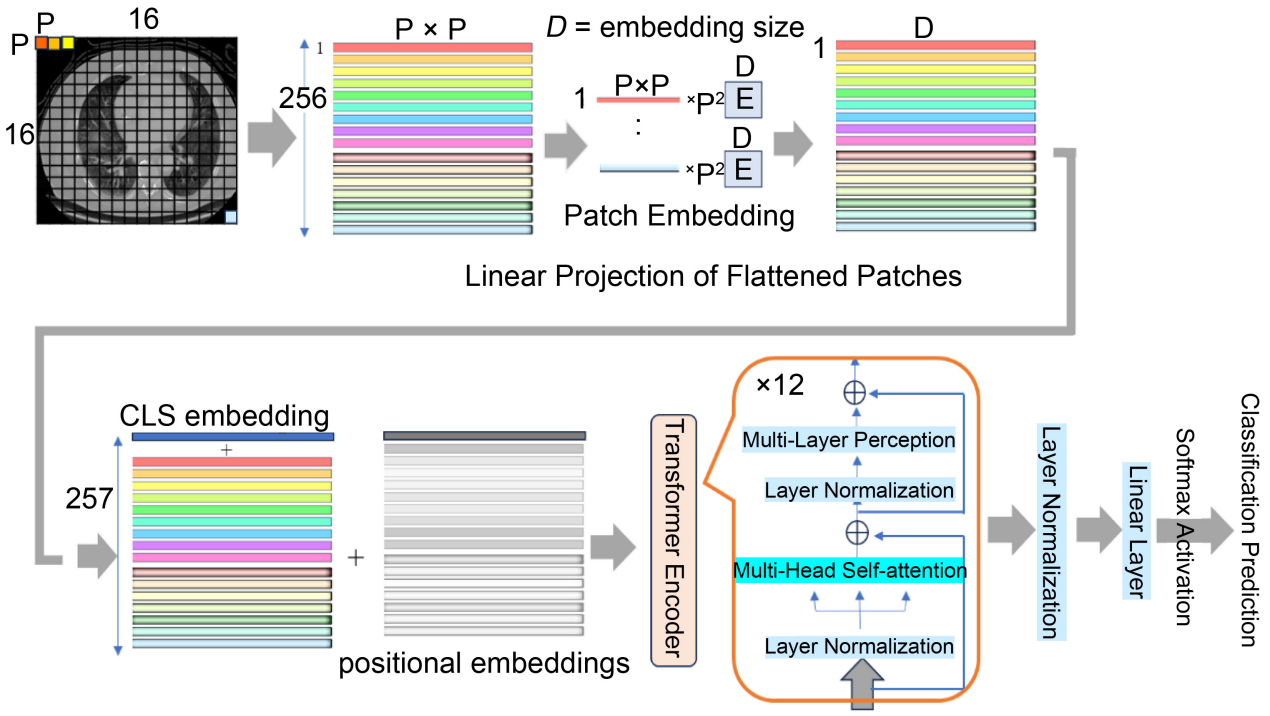


**Figure 2. Overview of the ViT model. The image is divided into 16 × 16 patches, and each patch is flattened. Learnable parameters are added through weighting (E), followed by class (CLS) embedding and position embedding, before being input into the transformer encoder.**

matrix is crucial for this evaluation, as it provides the basis for calculating these metrics [39]. It includes four possible outcomes: true positive (TP), false negative (FN), false positive (FP), and true negative (TN). The following standard metrics were used in this study, along with the area under the receiver operating characteristic curve (AUC).

Accuracy: Accuracy is a commonly used metric because it takes into account all values in the confusion matrix, *i.e.*, TP, FN, FP, and TN. It measures the proportion of correctly classified elements relative to all cases, defined as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TF} + \text{FP} + \text{FN}} \tag{1}$$

Precision: Precision compares the number of true positive values to all elements classified as positive, both true and false, and is defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

Recall: Recall (also called as sensitivity) measures the number of true positive values in relation to all positive cases and is defined as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

Specificity: Specificity measures how many cases classified as negative are actually negative and is defined as:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{4}$$

F1 score: The F1 Score is the harmonic mean of precision and recall, representing both values in a single metric, and is defined as:

$$\text{F1 score} = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}} \tag{5}$$

AUC: The AUC reflects how effectively the probabilities of the positive classes are distinguished from those of the negative classes.

### 2.3.2. Cross Entropy

Cross entropy measures the difference between two probability distributions [40-45]. In machine learning and deep learning, it's often used to evaluate how close the predicted probabilities from a model are to the actual probabilities. Essentially, cross entropy quantifies the gap between these two distributions.

The cross entropy between these two distributions is given by the following formula:

$$H(p,q) = -\sum_x p(x)\log_e q(x) \tag{6}$$

where $p$ is the true distribution, $q$ is the predicted distribution, and $x$ ranges over all possible outcomes.

Cross entropy measures the amount of information lost when using the predicted distribution to estimate the true one [40-45]. It helps evaluate the effectiveness of a classification model that outputs probabilities between 0 and 1. In simple terms, cross entropy shows how close the predicted distribution is to the actual one. A perfect match results in zero cross entropy, while larger differences result in higher values. Therefore, cross entropy is a useful metric for assessing the performance of classification models. When evaluating the classification performance of two models using cross entropy, the entropy values of both models are compared. A lower cross entropy value indicates that the model is more confident in its predictions, which typically corresponds to higher accuracy. Conversely, a higher cross entropy value signifies greater uncertainty in the predictions, leading to lower accuracy. A numerical example demonstrating the use of cross entropy for evaluating the quality of a deep learning classifier in a multi-class classification context is available in the literature [37].

Using cross entropy for evaluating the quality of a deep learning classifier offers several advantages [40-46]. First, cross entropy is highly sensitive to errors and penalizing confident but incorrect predictions more severely than predictions that are close to the correct answer. This sensitivity enhances its usefulness for accurate classification. Second, cross entropy thoroughly assesses the model's confidence in its predictions and measures how well the predicted probabilities match the actual data. This helps in interpreting the classifier output probabilistically. Third, the logarithmic scaling of cross entropy penalizes even small errors, encouraging the model to make more confident and accurate predictions.

## 3. RESULTS

In this study, we constructed ViT and ResNet50 models to classify lung cancer using CT images, evaluating their performance and visualizing data distribution to examine decision-making ambiguity. The experiments were conducted by setting the same number of training data for both ViT and ResNet50 models and performing 10-fold cross-validation. The total count of the 10 subsets is shown as a confusion matrix in **Figure 3**. The accuracy of each model, averaged over the 10 subsets, was 0.953 and 0.945, respectively. The AUCs were 0.991 and 0.988, respectively. The conventional evaluation metrics, calculated from **Figure 3** (assuming each lesion as positive), as well as the proposed metric, "cross entropy," are presented in **Table 1** and **Table 2**.

The accuracy and cross-entropy values for the 10 subsets obtained through 10-fold cross validation are shown in **Table 3** and **Table 4**. For each subset, the same data (image sets) were used in both models.

**Figure 4** shows the data distribution of Subset No. 9, which has an identical accuracy of 0.946 in both **Table 3** and **Table 4**, after dimensionality reduction using t-SNE. This represents the results of a 4-category

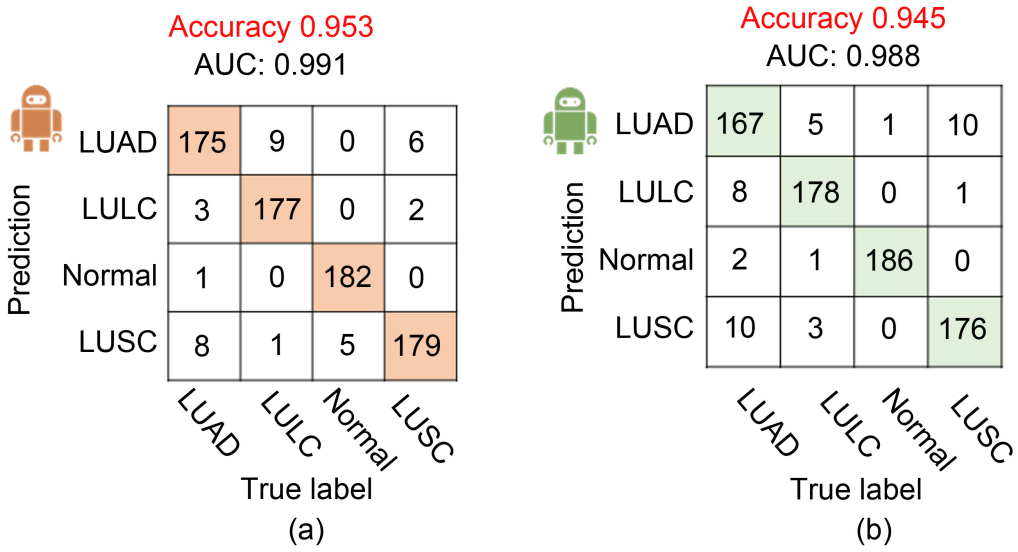


**Figure 3.** Confusion matrices: Total count of the 10-fold cross-validation. (a) Vit Model: accuracy 0.953, AUC 0.991; (b) ResNet 50 Model: accuracy 0.945, AUC 0.988.

**Table 1.** Cross entropy and existing evaluation metrics. Results of ViT when considering each lesion as positive.

| | Cross Entropy | Precision | Recall | F1 | Specificity |
|---|---|---|---|---|---|
| LUAD | 0.212 | 0.921 | 0.941 | 0.931 | 0.97 |
| LULC | 0.271 | 0.959 | 0.907 | 0.932 | 0.991 |
| Normal | 0.055 | 0.995 | 1 | 0.997 | 0.998 |
| LUSC | 0.149 | 0.947 | 0.947 | 0.947 | 0.98 |
| Average | 0.172 | 0.955 | 0.949 | 0.952 | 0.985 |

**Table 2.** Cross entropy and existing evaluation metrics. Results of ResNet 50 when considering each lesion as positive.

| | Cross Entropy | Precision | Recall | F1 | Specificity |
|---|---|---|---|---|---|
| LUAD | 0.527 | 0.913 | 0.893 | 0.903 | 0.971 |
| LULC | 0.242 | 0.952 | 0.952 | 0.952 | 0.984 |
| Normal | 0.032 | 0.984 | 0.995 | 0.989 | 0.995 |
| LUSC | 0.25 | 0.952 | 0.947 | 0.949 | 0.984 |
| Average | 0.263 | 0.95 | 0.947 | 0.948 | 0.984 |

classification, where the clusters are divided into four groups, but the misclassified data points appear as isolated points. The arrows and numbers in Figure 4 indicate the average information entropy (in bits) of the misclassified data, reflecting the uncertainty (ambiguity) in the 4-category classification. This allows for

Table 3. Cross entropy and accuracy of the Vit model.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Accuracy | 0.947 | 0.96 | 0.947 | 0.974 | 0.974 | 0.974 | 0.96 | 0.974 | 0.946 | 0.879 | 0.953 |
| | Subset No. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Average |
| Cross entropy | LUAD | 0.418 | 0.018 | 0.113 | 0.181 | 0.373 | 0.137 | 0.201 | 0.048 | 0.303 | 0.41 | 0.22 |
| | LULC | 0.143 | 0.511 | 0.776 | 0.307 | 0.051 | 0.062 | 0.216 | 0.031 | 0.378 | 0.195 | 0.267 |
| | Normal | 0.028 | 0.002 | 0.16 | 0.000 | 0.145 | 0.026 | 0.000 | 0.000 | 0.000 | 0.197 | 0.056 |
| | LUSC | 0.275 | 0.024 | 0.004 | 0.008 | 0.037 | 0.24 | 0.224 | 0.414 | 0.088 | 0.155 | 0.147 |
| | Average | 0.216 | 0.139 | 0.263 | 0.124 | 0.152 | 0.117 | 0.16 | 0.123 | 0.192 | 0.24 | 0.172 |

Table 4. Cross entropy and accuracy of the ResNet50 model.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Accuracy | 0.973 | 0.92 | 0.933 | 0.987 | 0.973 | 0.907 | 0.92 | 0.973 | 0.946 | 0.919 | 0.945 |
| | Subset No. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Average |
| Cross entropy | LUAD | 0.360 | 1.144 | 0.587 | 0.046 | 0.106 | 0.677 | 1.391 | 0.000 | 0.767 | 0.393 | 0.547 |
| | LULC | 0.015 | 0.708 | 0.541 | 0.028 | 0.003 | 0.064 | 0.009 | 0.004 | 0.591 | 0.422 | 0.238 |
| | Normal | 0.000 | 0.000 | 0.005 | 0.000 | 0.032 | 0.785 | 0.802 | 0.493 | 0.079 | 0.259 | 0.246 |
| | LUSC | 0.000 | 0.000 | 0.000 | 0.000 | 0.327 | 0.005 | 0.000 | 0.000 | 0.000 | 0.000 | 0.033 |
| | Average | 0.094 | 0.463 | 0.283 | 0.018 | 0.117 | 0.383 | 0.550 | 0.124 | 0.359 | 0.268 | 0.266 |

an evaluation of the model's performance, whether it outputs results with ambiguity or misclassifies data with high confidence.

For example, when the entropy is 2.0, it means that the model has no understanding of the 4-category classification (no confidence at all). When the entropy is 1.0, it indicates that the model has equal confidence between two categories (making it difficult to decide between them). With an entropy of 0.5, the model has slightly higher confidence in one category and lower confidence in the others. When the entropy is 0.2, the model has high confidence in one category, and when the entropy is 0, it can be interpreted as the model having absolute confidence in one category.

## 4. DISCUSSION AND CONCLUSIONS

In this study, we used the ViT model and the ResNet50 model to perform a four-class classification of lung cancer. As shown in Figure 3, the accuracy of both models is approximately 95%. Numerically, both models demonstrate equivalent performance. However, the confusion matrix reveals that the ResNet50 model has more misclassifications in the LUAD class, suggesting that its utility as a four-class lung cancer classification model may be limited. Thus, it is difficult to compare the usefulness of models based solely on accuracy. Therefore, conventional metrics such as recall, precision, and F1 score are employed for evaluating the performance of deep learning models.

Table 1 and Table 2 show the performance evaluation results of the classification models used in this experiment. The precision for each class is slightly higher in the ViT model, except for the LUSC class. In general, this suggests that the ViT model accurately detects the positive class without false detection. On the other hand, the ResNet model has a higher recall for the LULC class, meaning that it misses fewer instances of the LULC class. The F1 score, which balances recall and precision, is useful when one is high and the

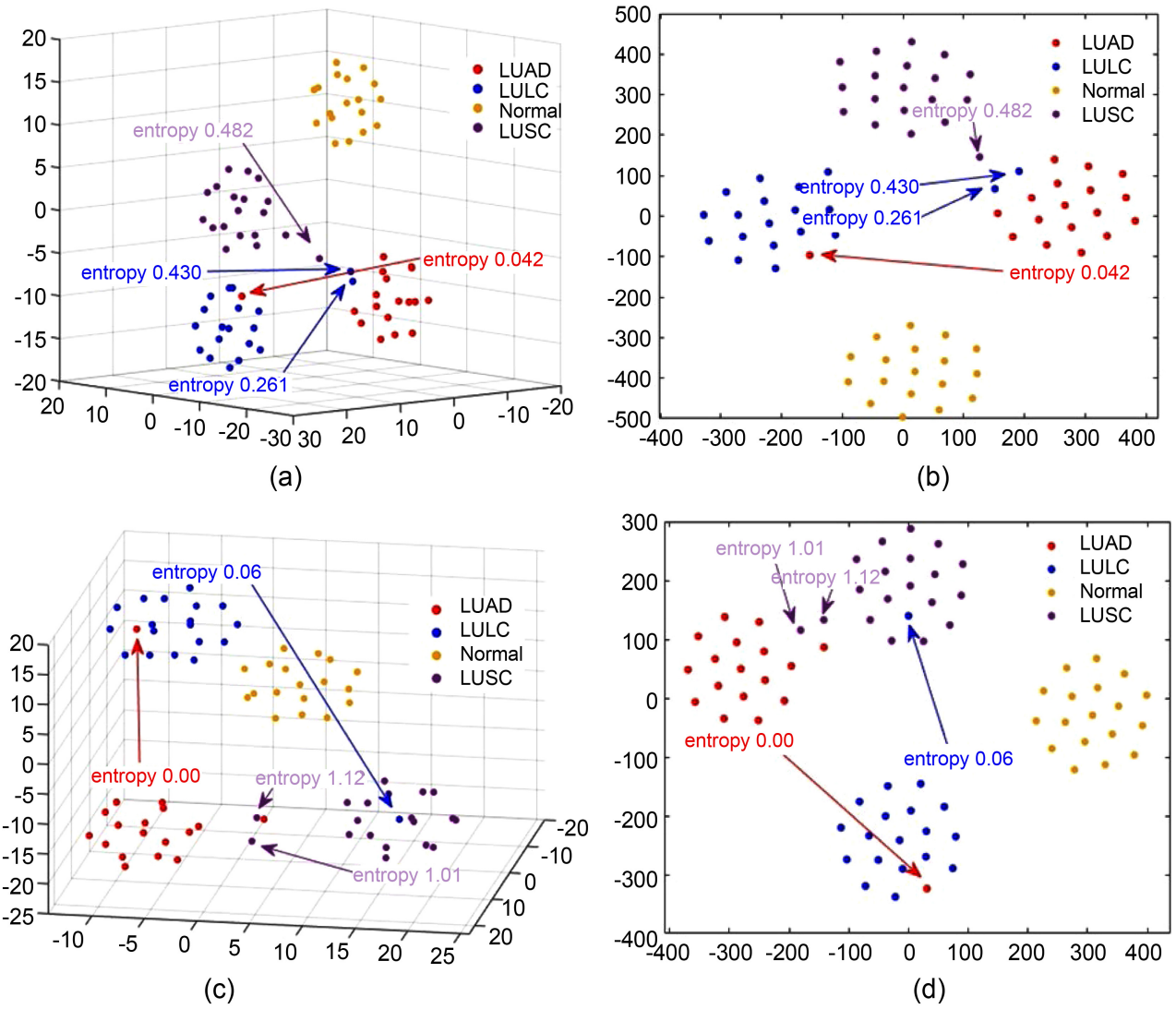


**Figure 4. Data distribution of Subset No. 9. The arrows and numbers indicate the misclassified data and their entropy (in [bits]). (a) 3D data distribution of the ViT model; (b) 2D data distribution of the ViT model; (c) 3D data distribution of the ResNet50 model; (d) 2D data distribution of the ResNet50 model.**

other is low. However, in this experiment, the F1 score is higher for the ViT model in the LUAD and Normal classes, while it is higher for the ResNet50 model in the LULC and LUSC classes. Specificity is a metric that evaluates how well false detections are minimized. In this experiment, the ViT model has higher specificity in the LULC and Normal classes, while the ResNet model is higher in the LUAD and LUSC classes. These results make it difficult to compare the performance of the two models. Thus, using only conventional evaluation metrics derived from the confusion matrix has limitations in comparing the performance of deep learning models.

The second column of Table 1 and Table 2 shows the cross entropy for each class, calculated from the probability distributions output by the models. Cross entropy is used as a metric to measure the discrepancy between the model's output probability distribution and the true class distribution (target distribution). A smaller value (closer to 0) indicates that the model's predictions are closer to the target distribution and are therefore more accurate. In other words, it quantitatively represents how reliable the model's predictions

are. From Table 1 and Table 2, the average cross entropy value is 0.172 for the ViT model and 0.263 for the ResNet50 model. This result suggests that the ViT model is closer to the true probability distribution and is a less uncertain model. Similarly, assuming each lesion as positive, the ViT model can be interpreted as having lower uncertainty than the ResNet50 model in the classification of LUAD, Normal, and LUSC, while the ResNet50 model has lower uncertainty for LULC. Thus, by using cross entropy as an evaluation metric, it is possible to compare uncertainty between multiple models and across classes within a single model, allowing for assessment from the perspective of reliability.

On the other hand, deep learning models are affected by the quality and quantity of the training data. Table 3 and Table 4 show the cross entropy of the 10 subsets in the 10-fold cross validation for the ViT and ResNet50 models. In Table 4, subset No. 7 shows a cross-entropy value of 1.391 for the LUAD class, which is significantly higher compared to the LUAD values in other subsets. This value indicates that the predictions for the LUAD class are uncertain (ambiguous), suggesting that there is room for improvement. These results imply that by using cross entropy as a metric, it is possible to revise the training process and improve the quality of the data.

Additionally, Table 3 and Table 4 show that even when accuracy is the same, the average cross entropy is not equivalent. This suggests that the uncertainty within the models differs, even if the classification accuracy is identical. For instance, in subset No. 9, the accuracy for both models is 0.946, but their cross-entropy values differ. In subset No. 9, the cross entropy for the LUAD, LULC, and Normal classes is lower in the ViT model compared to the ResNet50 model. This indicates that the ViT model is closer to the target distribution (more accurate with lower uncertainty) than the ResNet50 model. Conversely, in the LUSC class, the higher cross entropy value for the ViT model suggests that the ResNet50 model is more accurate.

Figure 4 visualizes these data distributions. Each point in the figure represents a single image data, and the four color-coded clusters indicate the classification results of the four lesion classes by the model. Misclassified data appear as isolated points. Figure 4(a) and Figure 4(b) show the data distribution of subset No. 9 for the ViT model, while Figure 4(c) and Figure 4(d) show the data distribution of subset No. 9 for the ResNet50 model. Since the model outputs are probabilities, the mean information (entropy) was calculated here as a measure of data distribution and model interpretability.

In the ViT model shown in Figure 4(a) and Figure 4(b), one point (corresponding to one image) with a true label of LUSC (purple) is located near the LUAD (red) cluster, with an entropy value of approximately 0.5. This indicates that the ViT model, while having relatively high confidence that the image belongs to the LUAD (red) class (and lower confidence in other classes), made an uncertain decision and misclassified it. A similar interpretation can be made for the point where LULC (blue) was misclassified as LUAD (red) with an entropy value of 0.430. On the other hand, another point where LULC (blue) was misclassified as LUAD (red) with an entropy value of 0.261 suggests that the model had high confidence in the LUAD (red) classification (with only slight uncertainty) when making the incorrect classification. On the other hand, the point where LUAD (red) was misclassified as LULC (blue) with an entropy value of 0.042 can be interpreted as a confident error. From this perspective, in the ResNet50 model shown in Figure 4(c) and Figure 4(d), the data where LUAD (red) was misclassified as LULC (blue) with an entropy value of 0, and the data where LULC (blue) was misclassified as LUSC (purple) with an entropy value of 0.06, can also be considered as errors made with strong confidence. This indicates that the model is over-confident (outputting a high score despite being incorrect in its prediction).

In the two images where LUSC (purple) was misclassified as LUAD (red) in Figure 4(c) and Figure 4(d), the entropy values for both are above 1.0. This indicates that the model is equally confident between the two classes out of the four, resulting in indecision. Naturally, uncertainty is also present in the correct predictions. The cross-entropy method proposed in this study quantitatively represents these and is used for performance evaluation.

Current ViTs face several limitations that affect their performance and usability [47-51]. They require substantial computational and memory resources, making them challenging to deploy in resource-limited environments. ViTs also demand large amounts of training data, which raises concerns about their efficiency

compared to traditional models. The fixed-size patch grid used in ViTs limits their flexibility in real-world applications, and quantization can lead to performance degradation, complicating their use in low-resource settings. Additionally, ViTs may produce underconfident predictions due to nonlinear responses, which can hinder decision-making. These challenges underscore the need for further research to improve the efficiency and adaptability of ViTs.

Our study has some limitations. First, due to computational resource constraints, the ViT and ResNet50 models were fine-tuned using pre-trained models on the ImageNet 2012 dataset (with a resolution of 384 × 384 and over 14 million images). If the models are pre-trained on a larger dataset (on the order of billions), the prediction results and uncertainty could significantly change. Therefore, the usefulness of the proposed method, which evaluates performance using cross-entropy, needs to be further validated. Second, we used an open-source dataset, which is easily accessible but may have labeling errors that we couldn't completely eliminate. Third, the potential bias within a single training dataset may limit the generalizability of the study findings to other datasets. Fourth, the relatively small dataset size of 748 CT images may have impacted the generalizability of the deep learning model results. To further optimize the model applications, it is important to validate the models using larger and more diverse datasets. Future research should focus on this aspect.

In conclusion, while the ViT outperformed the CNN-based ResNet50 in lung cancer classification from chest CT images, the performance differences were small and may not be clinically significant. Although ViTs have shown superior results in general imaging datasets, these findings suggest it may be premature to replace CNNs with ViTs in this specific application. Nonetheless, this study highlights the potential of ViTs and lays the groundwork for future advancements in the field.

## CONFLICTS OF INTEREST

The authors declare no conflicts of interest regarding the publication of this paper.

## REFERENCES


1. World Health Organization (2024) Global Cancer Burden Growing, amidst Mounting Need for Services. https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounting-need-for-services
2. Kwee, T.C. and Kwee, R.M. (2021) Workload of Diagnostic Radiologists in the Foreseeable Future Based on Recent Scientific Advances: Growth Expectations and Role of Artificial Intelligence. *Insights into Imaging*, **12**, Article No. 88. https://doi.org/10.1186/s13244-021-01031-4
3. Harolds, J.A., Parikh, J.R., Bluth, E.I., Dutton, S.C. and Recht, M.P. (2016) Burnout of Radiologists: Frequency, Risk Factors, and Remedies: A Report of the ACR Commission on Human Resources. *Journal of the American College of Radiology*, **13**, 411-416. https://doi.org/10.1016/j.jacr.2015.11.003
4. Lewis, S.J., Gandomkar, Z. and Brennan, P.C. (2019) Artificial Intelligence in Medical Imaging Practice: Looking to the Future. *Journal of Medical Radiation Sciences*, **66**, 292-295. https://doi.org/10.1002/jmrs.369
5. Rajpurkar, P., Chen, E., Banerjee, O. and Topol, E.J. (2022) AI in Health and Medicine. *Nature Medicine*, **28**, 31-38. https://doi.org/10.1038/s41591-021-01614-0
6. Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2017) Imagenet Classification with Deep Convolutional Neural Networks. *Communications of the ACM*, **60**, 84-90. https://doi.org/10.1145/3065386
7. Goodfellow, I., Bengio, Y. and Courville, A. (2016) Deep Learning. MIT Press.
8. Chen, M. (2024) Classification with Convolutional Neural Networks in Mapreduce. *Journal of Computer and Communications*, **12**, 174-190. https://doi.org/10.4236/jcc.2024.128011
9. Ren, J. and Wang, Y. (2022) Overview of Object Detection Algorithms Using Convolutional Neural Networks. *Journal of Computer and Communications*, **10**, 115-132. https://doi.org/10.4236/jcc.2022.101006

10. Lundervold, A.S. and Lundervold, A. (2019) An Overview of Deep Learning in Medical Imaging Focusing on MRI. *Zeitschrift für Medizinische Physik*, **29**, 102-127. https://doi.org/10.1016/j.zemedi.2018.11.002

11. Lakhani, P. and Sundaram, B. (2017) Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. *Radiology*, **284**, 574-582. https://doi.org/10.1148/radiol.2017162326

12. Iqbal, T. and Ali, H. (2018) Generative Adversarial Network for Medical Images (MI-GAN). *Journal of Medical Systems*, **42**, Article No. 231. https://doi.org/10.1007/s10916-018-1072-9

13. Comes, M.C., Fanizzi, A., Bove, S., Didonna, V., Diotaiuti, S., La Forgia, D., *et al.* (2021) Early Prediction of Neoadjuvant Chemotherapy Response by Exploiting a Transfer Learning Approach on Breast DCE-MRIs. *Scientific Reports*, **11**, Article No. 14123. https://doi.org/10.1038/s41598-021-93592-z

14. Comes, M.C., Fucci, L., Mele, F., Bove, S., Cristofaro, C., De Risi, I., *et al.* (2022) A Deep Learning Model Based on Whole Slide Images to Predict Disease-Free Survival in Cutaneous Melanoma Patients. *Scientific Reports*, **12**, Article No. 20366. https://doi.org/10.1038/s41598-022-24315-1

15. Bove, S., Fanizzi, A., Fadda, F., Comes, M.C., Catino, A., Cirillo, A., *et al.* (2023) A CT-Based Transfer Learning Approach to Predict NSCLC Recurrence: The Added-Value of Peritumoral Region. *PLOS ONE*, **18**, e0285188. https://doi.org/10.1371/journal.pone.0285188

16. Sakamoto, T., Furukawa, T., Lami, K., Pham, H.H.N., Uegami, W., Kuroda, K., *et al.* (2020) A Narrative Review of Digital Pathology and Artificial Intelligence: Focusing on Lung Cancer. *Translational Lung Cancer Research*, **9**, 2255-2276. https://doi.org/10.21037/tlcr-20-591

17. Silva, F., Pereira, T., Neves, I., Morgado, J., Freitas, C., Malafaia, M., *et al.* (2022) Towards Machine Learning-Aided Lung Cancer Clinical Routines: Approaches and Open Challenges. *Journal of Personalized Medicine*, **12**, Article No. 480. https://doi.org/10.3390/jpm12030480

18. Azad, R., Kazerouni, A., Heidari, M., Aghdam, E.K., Molaei, A., Jia, Y., *et al.* (2024) Advances in Medical Image Analysis with Vision Transformers: A Comprehensive Review. *Medical Image Analysis*, **91**, Article ID: 103000. https://doi.org/10.1016/j.media.2023.103000

19. Dosovitskiy, A., Beyer, L., Kolesnikov, A., *et al.* (2021) An Image Is Worth16x16 Words: Transformers for Image Recognition at Scale. https://arxiv.org/abs/2010.11929

20. Zhu, X., Su, W., Lu, L., *et al.* (2020) Deformable DETR: Deformable Transformers for End-to-End Object Detection. https://doi.org/10.48550/arXiv.2010.11929

21. 21. Chen, J., Lu, Y., Yu, Q., *et al.* (2021) TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. https://doi.org/10.48550/arXiv.2102.04306

22. Maurício, J., Domingues, I. and Bernardino, J. (2023) Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. *Applied Sciences*, **13**, Article No. 5521. https://doi.org/10.3390/app13095521

23. Gheflati, B. and Rivaz, H. (2022) Vision Transformers for Classification of Breast Ultrasound Images. 2022 44*th Annual International Conference of the IEEE Engineering in Medicine & Biology Society* (*EMBC*), Glasgow, 11-15 July 2022, 480-483. https://doi.org/10.1109/embc48229.2022.9871809

24. Wu, Y., Qi, S., Sun, Y., Xia, S., Yao, Y. and Qian, W. (2021) A Vision Transformer for Emphysema Classification Using CT Images. *Physics in Medicine & Biology*, **66**, Article ID: 245016. https://doi.org/10.1088/1361-6560/ac3dc8

25. Fanizzi, A., Fadda, F., Comes, M.C., Bove, S., Catino, A., Di Benedetto, E., *et al.* (2023) Comparison between Vision Transformers and Convolutional Neural Networks to Predict Non-Small Lung Cancer Recurrence. *Scientific Reports*, **13**, Article No. 20605. https://doi.org/10.1038/s41598-023-48004-9

26. Gai, L., Xing, M., Chen, W., Zhang, Y. and Qiao, X. (2023) Comparing CNN-Based and Transformer-Based Models for Identifying Lung Cancer: Which Is More Effective? *Multimedia Tools and Applications*, **83**, 59253-59269. https://doi.org/10.1007/s11042-023-17644-4

27. Uparkar, O., Bharti, J., Pateriya, R.K., Gupta, R.K. and Sharma, A. (2023) Vision Transformer Outperforms Deep Convolutional Neural Network-Based Model in Classifying X-Ray Images. *Procedia Computer Science*, **218**, 2338-2349. https://doi.org/10.1016/j.procs.2023.01.209

28. Goh, J.H.L., Ang, E., Srinivasan, S., Lei, X., Loh, J., Quek, T.C., *et al.* (2024) Comparative Analysis of Vision Transformers and Conventional Convolutional Neural Networks in Detecting Referable Diabetic Retinopathy. *Ophthalmology Science*, **4**, Article ID: 100552. https://doi.org/10.1016/j.xops.2024.100552

29. Oh, S., Kim, N. and Ryu, J. (2024) Analyzing to Discover Origins of CNNs and Vit Architectures in Medical Images. *Scientific Reports*, **14**, Article No. 8755. https://doi.org/10.1038/s41598-024-58382-3

30. Murphy, Z.R., Venkatesh, K., Sulam, J. and Yi, P.H. (2022) Visual Transformers and Convolutional Neural Networks for Disease Classification on Radiographs: A Comparison of Performance, Sample Efficiency, and Hidden Stratification. *Radiology*: *Artificial Intelligence*, **4**, e220012. https://doi.org/10.1148/ryai.220012

31. Cantone, M., Marrocco, C., Tortorella, F. and Bria, A. (2023) Convolutional Networks and Transformers for Mammography Classification: An Experimental Study. *Sensors*, **23**, Article No. 1229. https://doi.org/10.3390/s23031229

32. Nishigaki, D., Suzuki, Y., Watabe, T., Katayama, D., Kato, H., Wataya, T., *et al.* (2024) Vision Transformer to Differentiate between Benign and Malignant Slices in $^{18}$F-FDG PET/CT. *Scientific Reports*, **14**, Article No. 8334. https://doi.org/10.1038/s41598-024-58220-6

33. Chest CT-Scan Images Dataset. https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images

34. Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. https://doi.org/10.48550/arXiv.1409.1556

35. Szegedy, C., Liu, W., Jia, Y., *et al.* (2015) Going Deeper with Convolutions. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 7-12 June 2015, 1-9. https://doi.org/10.1109/cvpr.2015.7298594

36. He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/cvpr.2016.90

37. Matsuyama, E., Nishiki, M., Takahashi, N. and Watanabe, H. (2024) Using Cross Entropy as a Performance Metric for Quantifying Uncertainty in DNN Image Classifiers: An Application to Classification of Lung Cancer on CT Images. *Journal of Biomedical Science and Engineering*, **17**, 1-12. https://doi.org/10.4236/jbise.2024.171001

38. Vaswani, A., Shazeer, N., Parmar, N., *et al.* (2017) Attention Is All You Need. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 6000-6010. https://doi.org/10.48550/arXiv.1706.03762

39. Powers, D.M. (2020) Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. https://doi.org/10.48550/arXiv.2010.16061

40. Shan, B. and Fang, Y. (2020) A Cross Entropy Based Deep Neural Network Model for Road Extraction from Satellite Images. *Entropy*, **22**, Article No. 535. https://doi.org/10.3390/e22050535

41. Kurian, N.C., Meshram, P.S., Patil, A., Patel, S. and Sethi, A. (2021) Sample Specific Generalized Cross Entropy for Robust Histology Image Classification. 2021 *IEEE* 18*th International Symposium on Biomedical Imaging* (*ISBI*), Nice, 13-16 April 2021, 1934-1938. https://doi.org/10.1109/isbi48211.2021.9434169

42. Mannor, S., Peleg, D. and Rubinstein, R. (2005) The Cross Entropy Method for Classification. *Proceedings of the* 22*nd International Conference on Machine Learning—ICML*'05, Bonn, 7-11 August 2005, 561-568.

https://doi.org/10.1145/1102351.1102422

43. Brownlee, J. (2020) A Gentle Introduction to Cross-Entropy for Machine Learning. https://machinelearningmastery.com/cross-entropy-for-machine-learning/
44. Mao, A., Mohri, M. and Zhong, Y. (2023) Cross-Entropy Loss Functions: Theoretical Analysis and Applications. *Proceedings of the* 40*th International Conference on Machine Learning*, Honolulu, Vol. 202, 23803-23828. https://proceedings.mlr.press/v202/mao23b/mao23b.pdf
45. Nova (2023) A Comprehensive Guide to Cross Entropy in Machine Learning. https://aitechtrend.com/a-comprehensive-guide-to-cross-entropy-in-machine-learning/
46. Sheikh, I. (2023) Understanding Cross-Entropy Loss and Its Role in Classification Problems. https://medium.com/@l228104/understanding-cross-entropy-loss-and-its-role-in-classification-problems-d2550f2caad5
47. Chen, F., Luo, Z., Zhou, L., Pan, X. and Jiang, Y. (2024) Comprehensive Survey of Model Compression and Speed up for Vision Transformers. *Journal of Information*, *Technology and Policy*, 1-12. https://doi.org/10.62836/jitp.v1i1.156
48. Li, Y., Xu, S., Lin, M., Cao, X., Liu, C., Sun, X., *et al.* (2024) Bi-Vit: Pushing the Limit of Vision Transformer Quantization. *Proceedings of the AAAI Conference on Artificial Intelligence*, **38**, 3243-3251. https://doi.org/10.1609/aaai.v38i4.28109
49. Kim, J., Park, J., Kim, S. and Lee, J. (2024) Curved Representation Space of Vision Transformers. *Proceedings of the AAAI Conference on Artificial Intelligence*, **38**, 13142-13150. https://doi.org/10.1609/aaai.v38i12.29213
50. Pardyl, A., Kurzejamski, G., Olszewski, J., Trzci'nski, T. and Zieli'nski, B. (2023) Beyond Grids: Exploring Elastic Input Sampling for Vision Transformers. https://arxiv.org/abs/2309.13353
51. Pandey, L., Wood, S.M.W. and Wood, J.N. (2023) Are Vision Transformers More Data Hungry than Newborn Visual Systems? https://arxiv.org/abs/2312.02843