https://www.scirp.org/journal/jbise
J. Biomedical Science and Engineering, 2024, Vol. 17, (No. 1), pp: 1-12

Using Cross Entropy as a Performance Metric for Quantifying Uncertainty in DNN Image Classifiers: An Application to Classification of Lung Cancer on CT Images

Eri Matsuyama[1], Masayuki Nishiki[2], Noriyuki Takahashi[3], Haruyuki Watanabe[4]

[1]Faculty of Informatics, University of Fukuchiyama, Kyoto, Japan; [2]Graduate School of Radiological Sciences, International University of Health and Welfare, Tochigi, Japan; [3]School of Health Sciences, Fukushima Medical University, Fukushima, Japan; [4]School of Radiological Technology, Gunma Prefectural College of Health Sciences, Gunma, Japan

Correspondence to: Eri Matsuyama, matsuyama-eri@fukuchiyama.ac.jp
Keywords: Cross Entropy, Performance Metrics, DNN Image Classifiers, Lung Cancer, Prediction Uncertainty
Received: December 11, 2023 **Accepted:** January 14, 2024 **Published:** January 17, 2024

ABSTRACT

Cross entropy is a measure in machine learning and deep learning that assesses the difference between predicted and actual probability distributions. In this study, we propose cross entropy as a performance evaluation metric for image classifier models and apply it to the CT image classification of lung cancer. A convolutional neural network is employed as the deep neural network (DNN) image classifier, with the residual network (ResNet) 50 chosen as the DNN architecture. The image data used comprise a lung CT image set. Two classification models are built from datasets with varying amounts of data, and lung cancer is categorized into four classes using 10-fold cross-validation. Furthermore, we employ t-distributed stochastic neighbor embedding to visually explain the data distribution after classification. Experimental results demonstrate that cross entropy is a highly useful metric for evaluating the reliability of image classifier models. It is noted that for a more comprehensive evaluation of model performance, combining with other evaluation metrics is considered essential.

1. INTRODUCTION

In 2023, it is estimated that there will be nearly 2 million new cancer cases and approximately 610 thousand cancer-related deaths in the United States. The leading cause of cancer-related deaths for both men and women is lung cancer [1]. Lung cancer is broadly categorized into small cell lung cancer and non-small cell lung cancer (NSCLC) based on histological type. NSCLC accounts for approximately 80% - 85% of all lung cancer cases [2]. Further classification of NSCLC is based on histology, leading to subtypes

such as lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), large cell carcinoma (LULC), and others, each exhibiting unique characteristics. LUAD represents 85% of NSCLC cases, and most of the patients often face challenges in survival due to drug resistance and recurrence [2]. LUSC constitutes around 30% of all NSCLCs and is strongly linked to smoking, characterized by a high overall mutation rate of 8.1 mutations per megabase (1,000,000 base pairs long) and significant genomic complexity [3]. LULC has a molecular profile characteristic of adenocarcinoma, and this profile is more similar to adenocarcinoma than squamous cell carcinoma [4]. Additionally, the prognosis is worse than other types of non-small cell lung cancer. Even within the broad category of NSCLC, the characteristics vary depending on the subtype. Therefore, early identification of the histological type is crucial for treatment strategies and reducing mortality.

Low-dose computed tomography (LDCT) screening proves valuable for early lung cancer detection [5-9]. Nevertheless, the rapid evolution of computed tomography (CT) equipment has led to the identification of numerous microscopic nodules, intensifying the workload for radiologists. Consequently, the implementation of computer-aided diagnosis (CAD) systems is anticipated to help radiologists ease their burden. Broadly, CAD is categorized into two types: computer-aided detection (CADe), focusing on lesion detection (presence diagnosis), and computer-aided diagnosis (CADx), aiming to analyze lesions (definitive diagnosis, such as benign/malignant differentiation). Extensive research in chest CT CAD, dating back to the 1960s, has been conducted for nodules and lung diseases, yielding some positive outcomes [10-14]. However, challenges persist, including a higher rate of false positives compared to physicians [11] and limitations in enhancing system accuracy [13]. Conversely, image recognition using deep neural networks (DNN) has exhibited significant advancements in the past decade.

In recent years, it has been reported that amazing recognition accuracy can be obtained with the attention mechanism developed for natural language processing [15] and vision transformer, which applies a transformer-like model to image processing [16]. These advancements have eliminated the need for the traditionally challenging feature extraction process in CAD research, thus enabling the development of highly accurate and robust designs. Consequently, research on artificial intelligence-assisted CAD systems targeting pulmonary diseases using deep neural networks (DNN) has progressed significantly [17-20].

These papers discuss pattern detection of interstitial lung diseases [17, 18] and histological classification of lung cancer [19] using convolutional neural networks (CNNs). In both cases, to enhance the model's performance, the accuracy of the DNN model is evaluated from various perspectives, including accuracy, precision, recall, F-measure, receiver operating characteristic (ROC) curve, and area under the ROC (AUC), all considered gold standards. However, these existing evaluation metrics have problems such as lack of transparency in DNN inferences and inability to estimate uncertainty regarding results. As specific examples, there are issues such as uncertainty arising from facing out-of-distribution data [21], over-confident problems, and covariate shift [20].

Furthermore, in image classification tasks utilizing DNNs, it is common to employ the softmax function to represent the output as a probability value. However, one issue with DCNN is that calibration is often insufficient, making it difficult to interpret the model's output directly as a probabilistic measure [22].

Evaluating model uncertainty is essential for enhancing the transparency and reliability of predictions, improving data quality, and reducing misjudgments. Bayesian neural network (BNN) and Monte Carlo dropout (MCDO) [23] are recognized methods for estimating uncertainty in neural networks. BNN expresses the weights of a network model as a probability distribution, making it possible to estimate uncertainty in addition to prediction results. MCDO is a type of BNN, and is a method that enables approximate modeling of the probability distribution of weights by representing the weights of a network model using a Bernoulli distribution. However, both methods encounter the challenge of excessive computational costs when applied to DNNs.

In this study, we suggest employing cross entropy as a performance evaluation metric to quantify uncertainty in DNN image classifiers, applying it specifically to the classification of lung cancer in CT images. Cross entropy is typically utilized as a cost function during the training phase of DNN model con-

struction. However, in this study, we use it as one of the performance evaluation metrics for the classification model.

2. MATERIALS AND METHOD

In this study, we use a CNN as a DNN image classifier and perform finetuning. Two classification models are constructed using two data sets with different numbers of data. Each model undergoes a 10-fold cross-validation to perform a four-class classification task. In this experiment, alongside computing the proposed cross-entropy metric, we also calculate existing evaluation metrics for comparison. Additionally, we visualize the data distribution after classifying the classes.

2.1. Image Date Sets

The data used comprise a lung CT image set classified into four classes: LUAD, LULC, LUSC, and normal. This dataset is publicly available on the web for non-profit purposes, as provided by the research community [24]. Consequently, ethical concerns do not arise in this study, and obtaining informed consent is not necessary. An illustration of the image data is presented in **Figure 1**.

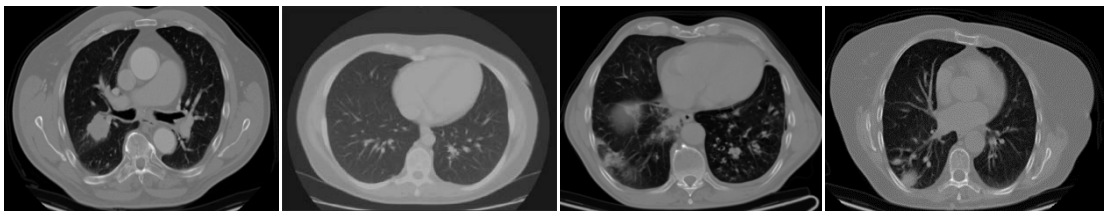
In the experiment, two models were constructed: "Model A", which was trained on a total of 1000 images with imbalanced data counts for each lesion, and "Model B", which was trained on a total of 748 images with balanced data counts for each lesion. Both models undergo a 10-fold cross-validation, where 90% of the data is allocated for training and the remaining 10% for validation. The distribution and total numbers of the data are detailed in **Table 1**.

2.2. Multioutput Classification Model Used

In this study, we employ the residual network (ResNet) 50 architecture [25] for the deep convolutional neural network (DCNN) and conduct learning through fine-tuning. Typically, in DCNNs, accuracy does not improve unless the number of stacked layers is sufficiently large. However, when the number of layers surpasses a certain threshold, the vanishing gradient problem arises, leading to a deterioration in accuracy. In ResNet, the introduction of a mechanism called shortcut connection solved the vanishing gradient problem by directly adding the input of the preceding layer to the subsequent layer [25]. Consequently, this allows for the realization of a deep network, and ResNet50 is considered highly effective for medical imaging applications [26].

In the fine-tuning process of this experiment, we utilize the pre-trained ResNet50 model on natural images, retraining the entire network using lung CT images. In other words, fine-tuning is executed without placing a frozen (no weight updates) layer, and a four-class classification is conducted. Consequently, the final fully connected layer and the last classification layer are replaced and trained with new configurations tailored to the number of categories. To meet the structural requirements of ResNet50, the input data size needs to be 224×224. Therefore, bicubic interpolation is employed to standardize the overall image size. The mini-batch size is set to 10, and the optimizer used is Adam (combining momentum SGD + RMSprop). In the retraining with CT images, parameters are adjusted so that the learning rate increases in the newly replaced fully connected layer, decreases in the transfer layer, and decreases after completion of every 5 epochs. To prevent overfitting, an L2 regularization term is incorporated into the cost function (loss function). The number of epochs is determined by evaluating accuracy validation after each iteration. Retraining is halted if the accuracy falls below the highest accuracy achieved in the last 5 consecutive validations.

In this experiment, with a focus on model interpretability, we visualize the distribution of post-classification data. To achieve this, we employ t-distributed stochastic neighbor embedding (t-SNE) [27]. t-SNE is a dimension reduction method that condenses data into a low-dimensional space while preserving distances in high-dimensional data, allowing for nonlinear mapping. In this experiment, all high-dimensional activation data points in the final softmax layer are visualized through a two-dimensional mapping.



(a) (b) (c) (d)

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

Table 1. Breakdown of the image dataset used.

Class	Model A	Model B
LUAD (adenocarcinoma)	338	187
LULC (large cell carcinoma)	187	187
LUSC (squamous cell carcinoma)	260	187
Normal	215	187
Total	1000	748

2.3. Cross Entropy

Cross entropy serves as a metric for gauging the dissimilarity between two probability distributions [28-33]. In the realm of machine learning and deep learning, it is commonly used to assess the gap between the predicted probability distribution produced by a model and the true, ground truth probability distribution. Fundamentally, cross entropy quantifies the degree of disparity between these two distributions.

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

$$H(p,q) = -\sum_x p(x)\log_e q(x) \tag{1}$$

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

Cross entropy indicates the amount of information lost when utilizing the predicted distribution to infer the real one [28-33]. Essentially, it offers insights into the effectiveness of a classification model that provides probabilities ranging from 0 to 1. Put simply, it reveals the proximity of the predicted distribution to the actual one. A perfect match results in zero cross entropy, while significant differences yield a higher value. Consequently, cross entropy serves as a versatile metric for evaluating the performance of classification models.

The following provides a simplified numerical example of utilizing cross entropy for the quality evaluation of a deep learning classifier in a multi-class classification context [31, 33]. Let's consider a case where we have a deep learning classifier trained for a multi-class classification problem with three classes: apple, orange, and pear. The model has undergone training, and now our objective is to assess its performance using cross entropy.

Assume we have a small test dataset with three samples and the true class labels are:

Sample 1: True label = apple.

Sample 2: True label = orange.

Sample 3: True label = pear.

Now, let's say the model's predictions for these samples produce the following class probabilities:

Sample 1: Predicted probabilities = [0.7, 0.15, 0.15]

(70% confidence in apple, 15% in orange, 15% in pear).

Sample 2: Predicted probabilities = [0.1, 0.8, 0.1]
(10% confidence in apple, 80% in orange, 10% in pear).
Sample 3: Predicted probabilities = [0.25, 0.25, 0.5]
(25% confidence in apple, 25% in orange, 50% in pear).

Using Equation (1), we calculate the cross entropy for each sample and then the average cross entropy for the entire test dataset:

Sample 1—True label: [1, 0, 0], Predicted label: [0.7, 0.15, 0.15],
Cross entropy = $-(1)\log_e(0.7) = 0.3567$.
Sample 2—True label: [0, 1, 0], Predicted label: [0.1, 0.8, 0.1],
Cross entropy= $-(1)\log_e(0.8) = 0.2231$.
Sample 3—True label: [0, 0, 1], Predicted label: [0.25, 0.25, 0.5],
Cross entropy = $-(1)\log_e(0.5) = 0.6931$.
Then, we calculate the average cross entropy for the entire test dataset:
Average cross entropy = $(0.3567 + 0.2231 + 0.6931)/3 = 0.4243$.

In this example, the average cross entropy for the test dataset is approximately 0.4243. A lower cross entropy implies that the model's predicted probabilities are closer to the true class probabilities, indicating better model performance.

When evaluating the classification performance of two CNN models using cross entropy, the entropy values for both models are compared. A lower entropy suggests that the model is more confident in its predictions, leading to higher accuracy. Conversely, a higher entropy indicates more uncertainty and lower accuracy.

2.4. Merits of Using Cross Entropy as an Evaluation Metric for Classification Models

Using cross entropy for quality evaluation of a deep learning classifier provides several advantages [28-34]:

- Cross entropy, rooted in information theory, can be perceived as a measure of information gain or loss. It quantifies the information gained when the true class label is disclosed, taking into account the predicted probabilities.
- Cross entropy is very sensitive to prediction errors. Incorrect predictions made with confidence are penalized more heavily than predictions closer to the correct answer. This sensitivity makes it a valuable indicator when accurate classification is a priority.
- Cross entropy takes into account the probability distribution predicted by the classifier. It assesses the dissimilarity between the predicted probabilities and the true class labels. This approach offers a more detailed evaluation of the model's confidence in its predictions.
- Cross entropy directly quantifies the likelihood of observed data based on predicted probabilities. This measurement evaluates how well the model's predicted probabilities match the actual class labels and aids in the probabilistic interpretation of the classifier's output.
- Cross entropy applies a logarithmic scaling to errors. This means that it penalizes even minor prediction errors, thus motivating the model to have greater confidence in its predictions.

3. RESULTS

The average accuracy of "Model A", trained on 1000 images with imbalanced data for each lesion, and that of "Model B", trained on 748 images with balanced data, was 0.974 and 0.954, respectively. The AUC values were 0.996 and 0.988 for "Model A" and "Model B", respectively. The confusion matrices for both models are presented in **Figure 2**. The confusion matrices in the figure are cross tables that count the results of 10 subsets of the 10-fold cross validation. **Table 2** and **Table 3** show the results of cross entropy and the existing evaluation metrics (precision, recall, F1, and specificity) when each lesion is considered positive for Models A and B, respectively. The values in the last row of the respective table represent the average values for each metric.

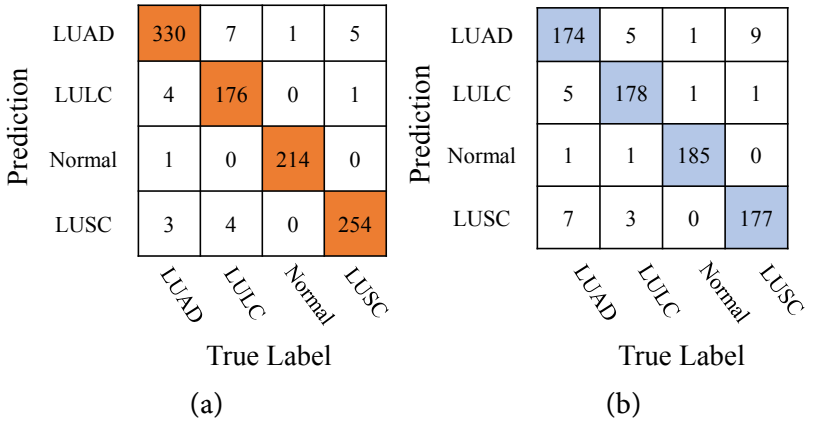


Figure 2. Confusion matrices. (a) Model A: accuracy 0.974, AUC 0.996, the value with the orange color is the number of correct answers. (b) Model B: accuracy 0.954, AUC 0.988, the value with the blue color is the number of correct answers.

Table 2. Cross entropy and existing evaluation metrics for Model A.

	Cross Entropy	Precision	Recall	F1	Specificity
LUAD	0.136	0.962	0.976	0.969	0.980
LULC	0.204	0.972	0.941	0.957	0.994
Normal	0.004	0.995	0.995	0.995	0.991
LUSC	0.122	0.973	0.977	0.975	0.991
Average	0.117	0.976	0.972	0.974	0.989

Table 3. Cross entropy and existing evaluation metrics for Model B.

	Cross Entropy	Precision	Recall	F1	Specificity
LUAD	0.313	0.921	0.930	0.926	0.921
LULC	0.270	0.962	0.952	0.957	0.988
Normal	0.139	0.989	0.989	0.989	0.996
LUSC	0.268	0.947	0.947	0.947	0.982
Average	0.247	0.955	0.955	0.955	0.972

Table 4 displays the accuracy and cross entropy for each of the 10 subsets in Model A. The "Average" in the last row represents the average cross entropy in each subset, while the "Average" in the rightmost column signifies the average cross entropy for each lesion prediction. Additionally, as an illustration of the existing evaluation metrics, Table 5 and Table 6 respectively present the calculated values for subsets No. 2 and No. 8, where accuracy was equivalent among the 10 subsets in Table 4. Figure 3 illustrates the dimensionality reduction data distribution map after class classification for subsets No. 2 and No. 8, respectively.

4. DISCUSSION

Generally, classification results obtained using the DCNN model are computed by tallying the number of correct answers/incorrect answers and summarizing them into a confusion matrix. Subsequently, various evaluation metrics are calculated. In this experiment, we created a confusion matrix (Figure 2)

and computed various existing metrics (columns 3 to 6 of Table 2 and Table 3). In addition to accuracy and AUC metrics, as evident from these tables, all the average values of the existing metrics are higher for Model A than for Model B. Consequently, when all metrics exhibit high values, it is generally straightforward to conclude that Model A is more accurate. However, it is crucial to selectively use evaluation metrics based on the specific purpose of the classification task. For instance, in the case of a 4-class classification of lung cancer with LULC as the positive class, as shown in the third row of the two tables, the recall for Model A is 0.941, while for Model B, it is 0.952, indicating a higher value for Model B. Conversely, precision and specificity are higher for Model A, and the F1 score remains the same. In such cases, it becomes challenging to determine which model can accurately distinguish LULC.

Furthermore, when considering "Normal" as the positive class, as indicated in the fourth row of Table 2 and Table 3, Model B exhibits high specificity. In this case, Model B can be considered to be able to more accurately identify lung cancer from a group of lung cancer images. Consequently, relying solely on existing evaluation metrics derived from the confusion matrix has its limitations when assessing the performance of a model. Moreover, there is an issue wherein existing evaluation metrics do not furnish information about the distribution of probabilities, which represent the model outputs. Consequently, it is not possible to evaluate reliability based on confidence.

Table 4. Accuracy and cross entropy for Model A.

Subset No.	1	2	3	4	5	6	7	8	9	10	Average
Accuracy	0.970	0.980	0.960	0.940	1.000	1.000	0.990	0.980	0.990	0.930	0.974
LUAD	0.013	0.160	0.032	0.233	0.009	0.005	0.015	0.210	0.000	0.688	0.136
LULC	0.492	0.002	0.641	0.237	0.005	0.001	0.015	0.032	0.146	0.467	0.204
Normal	0.000	0.000	0.000	0.000	0.000	0.000	0.041	0.000	0.000	0.000	0.004
LUSC	0.603	0.108	0.026	0.185	0.000	0.000	0.012	0.028	0.000	0.256	0.122
Average	0.277	0.067	0.175	0.164	0.004	0.002	0.020	0.068	0.037	0.353	0.117

Table 5. Calculated values for subset No. 2 of the 10 subsets from Table 4, where the accuracy of the subset is 0.98.

	Cross Entropy	Precision	Recall	F1	Specificity
LUAD	0.160	0.971	0.971	0.971	0.985
LULC	0.002	0.950	1.000	0.974	0.988
LUSC	0.108	1.000	0.962	0.980	1.000
Normal	0.000	1.000	1.000	1.000	1.000
Average	0.067	0.980	0.983	0.982	0.993

Table 6. Calculated values for subset No. 8 of the 10 subsets from Table 4, where the accuracy of the subset is 0.98.

	Cross Entropy	Precision	Recall	F1	Specificity
LUAD	0.210	1.000	0.941	0.970	1.000
LULC	0.032	0.947	1.000	0.973	0.988
LUSC	0.028	0.963	1.000	0.981	0.986
Normal	0.000	1.000	1.000	1.000	1.000
Average	0.068	0.978	0.985	0.981	0.994

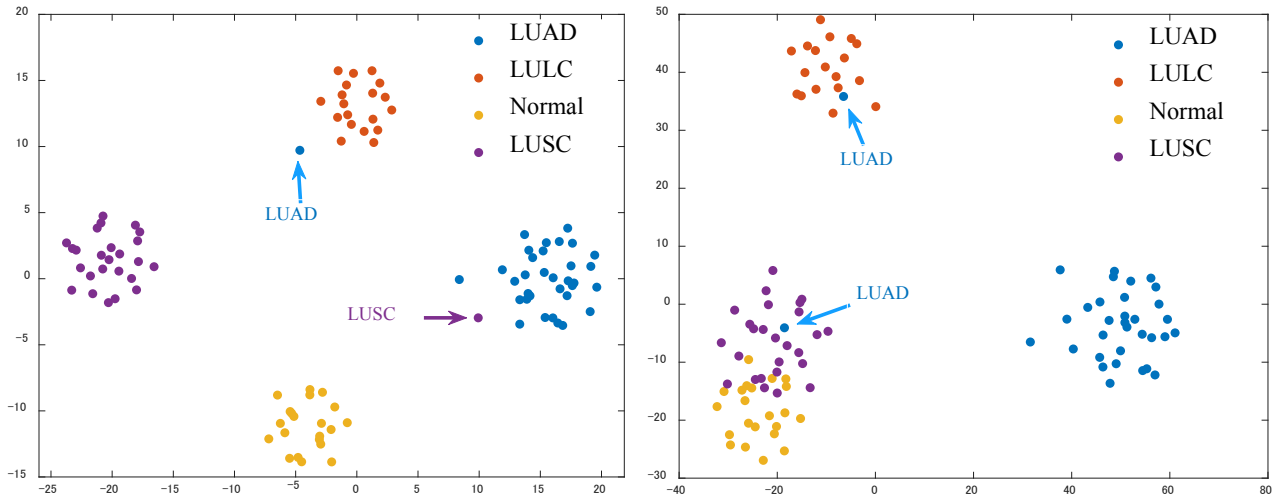


Figure 3. Two-dimensional data mapping after class classification. (a) Distribution of 2D data for subset No. 2. (b) Distribution of 2D data for subset No. 8.

The second column in Table 2 and Table 3 represents the cross entropy for each class calculated using the probability distribution output by the model. Cross entropy indicates how closely the predicted probability distribution aligns with the true distribution, with lower values indicating proximity to the true distribution and higher values indicating deviation from it. In essence, it quantitatively demonstrates the reliability of the model's predictions. From the two tables, the average cross-entropy values are 0.117 for Model A and 0.247 for Model B. This result suggests that Model A is closer to the true probability distribution, signifying lower uncertainty. Similarly, even when considering any of the lesions as positive, it can be stated that Model A exhibits lower uncertainty. By employing cross entropy as an evaluation metric in this manner, it becomes possible to compare the uncertainty between multiple models and assess their reliability.

In DCNN classification, the classification results may be influenced by the imbalance in the training data. Table 4 illustrates the cross entropy for each subset in the 10-fold cross-validation of Model A. In subset No. 1, the cross-entropy value (0.603) for LUSC is significantly higher compared to LUSC in the other subsets. Regarding LUAD, the value in subset No. 10 (0.688) is comparably high. These findings suggest that predictions for these lesions are uncertain (ambiguous), indicating a bias in the data. This outcome implies that using cross entropy as a metric can prompt a reevaluation of the data, leading to an improvement in data quality. For example, the accuracy for both subsets No. 5 and No. 6 is 1.0, but their respective cross entropy values differ. This result demonstrates that even if all class classifications are correct, varying levels of uncertainty exist. Table 5 and Table 6 compare subsets No. 2 and No. 8, both having an accuracy of 0.98. With existing evaluation metrics, interpreting which specific metrics should be used to assess performance becomes challenging.

On the contrary, utilizing cross entropy facilitates a straightforward comparative evaluation. In subsets No. 2 and No. 8, the average cross entropy for the 4-class classification is 0.067 and 0.068, respectively, indicating nearly equivalent performance. However, the cross-entropy values for each class of lesions differ, signifying distinct predictive uncertainties. For instance, in the prediction of LUAD, the cross-entropy value (0.210) in subset No. 8 (Table 6) is higher than that (0.160) in subset No. 2 (Table 5). A similar pattern is observed in the case of LULC, where the cross-entropy value is higher in subset No. 8. This implies that the predictions for LUAD and LULC are more ambiguous (with higher uncertainty) in subset No. 8, as compared to subset No. 2. As for LUSC, the cross-entropy value is higher in subset No. 2, indicating greater uncertainty in the predictions for this subset.

The visualization of these data distributions is presented in Figure 3. In subset No. 8 (Figure 3(b)), two LUAD data points (blue) are intertwined with the clusters of LULC (red) and LUSC (purple). The cross entropy for LUAD in this subset is 0.210, indicating the highest level of uncertainty in the predictions. With this mixture, it can be inferred that the uncertainty of LULC (red) and LUSC (purple) has increased. In subset No. 2 (Figure 3(a)), there are isolated points in LUAD (blue) and LUSC (purple) respectively. From the distribution of LUAD (blue) data points, it is apparent that there is little influence on other clusters, but there is some uncertainty in the predictions. On the other hand, the isolated points in LUSC (purple) indicate uncertainty in the predictions and may also affect the prediction of LUAD (blue). Thus, cross entropy has the capability to capture the uncertainty (ambiguity) in class-specific predictions, which cannot be determined by existing evaluation metrics. Based on these results, we believe that cross entropy is a highly useful metric for evaluating model reliability.

The accuracy of cross-entropy used for classifying lung cancer in CT images depends on various factors, including the quality of the dataset, the complexity of the model architecture, and the overall experimental setup. However, the main purpose of our paper is to use cross entropy as a performance metric for quantifying uncertainty in DNN image classifiers. Thus, we have refrained from delving into detailed discussion on this matter as it lies beyond the scope addressed in this work. Forecast uncertainty in medical image classification tasks can arise from various factors. For example, limited data availability, data quality and variability, class imbalance, artifact presence, model complexity, etc. Since our paper mainly focus on the application of cross entropy to classification of lung cancer, discussion on what factor contribute to prediction uncertainty was not detailed conducted.

Cross entropy is considered a useful evaluation metric for the performance of a multi-class classifier. However, it does have some limitations. First, it is somewhat sensitive to class imbalance. If there is a significant imbalance in the distribution of classes in the dataset, the model may be biased towards the majority class. Second, while cross entropy provides a measure of how well the predicted probabilities match the true distribution of classes, it does not offer direct interpretability. Third, cross entropy assumes that the predictions for each class are independent of each other. In some real-world scenarios, classes may be correlated, and this assumption may not hold. In spite of these limitations, cross entropy is considered a valuable model evaluation metric due to its simplicity and effectiveness. However, it is essential to complement its use with other evaluation metrics for a more comprehensive assessment.

5. CONCLUSION

In this study, we proposed the utilization of cross entropy, known as the loss function for DNN models, as one of the performance evaluation metrics for the models. We applied this metric to the classification of lung cancer in CT images. As a result, we demonstrated that it is possible to quantitatively depict the uncertainty of predictions based on the differences in probability distributions in the model's output. Particularly in multi-class classification tasks, it was possible to demonstrate uncertainty for each class. Furthermore, by mapping the class classification results into two-dimensional data, we were able to visually interpret the prediction uncertainty indicated by cross-entropy values. Based on these results, cross entropy is considered a very useful metric for evaluating model reliability. However, for a more comprehensive evaluation of DNN model performance, it is essential to use cross entropy in conjunction with other evaluation metrics.

ACKNOWLEDGEMENTS

This work was supported in part by JSPS KAKENHI (Grant-in-Aid for Scientific Research) Grant Number 18K15641.

CONFLICTS OF INTEREST

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

REFERENCES

1. Siegel, R.L., Miller, K.D., Wagle, N.S. and Jemal, A. (2023) Cancer Statistics, 2023. *CA*: *A Cancer Journal for Clinicians*, **73**, 17-48. https://doi.org/10.3322/caac.21763
2. Xu, R., Lu, T., Wang, C., Li, Q., Peng, B., Zhao, J., *et al.* (2023) Single-Cell Data Analysis of Malignant Epithelial Cell Heterogeneity in Lung Adenocarcinoma for Patient Classification and Prognosis Prediction. *Heliyon*, **9**, e20164. https://doi.org/10.1016/j.heliyon.2023.e20164
3. Niu, Z., Jin, R., Zhang, Y. and Li, H. (2022) Signaling Pathways and Targeted Therapies in Lung Squamous Cell Carcinoma: Mechanisms and Clinical Trials. *Signal Transduction and Targeted Therapy*, **7**, 353. https://doi.org/10.1038/s41392-022-01200-x
4. Copin, M.-C. (2016) Carcinome à Grandes Cellules, Carcinome Lymphoepithelioma-Like, Carcinome NUT Large Cell Carcinoma, Lymphoepithelioma-Like Carcinoma, NUT Carcinoma. *Annales de Pathologie*, **36**, 24-33. https://doi.org/10.1016/j.annpat.2015.11.006
5. The National Lung Screening Trial Research Team (2011) Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. *The New England Journal of Medicine*, **365**, 395-409. https://doi.org/10.1056/NEJMoa1102873
6. Aberle, D.R., DeMello, S., Berg, C.D., Black, W.C., Brewer, B., Church, T.R., *et al.* (2013) Results of the Two Incidence Screenings in the National Lung Screening Trial. *The New England Journal of Medicine*, **369**, 920-931. https://doi.org/10.1056/NEJMoa1208962
7. The National Lung Screening Trial Research Team (2013) Results of Initial Low-Dose Computed Tomographic Screening for Lung Cancer. *The New England Journal of Medicine*, **368**, 1980-1991. https://doi.org/10.1056/NEJMoa1209120
8. Kramer, B.S., Berg, C.D., Aberle, D.R. and Prorok, P.C. (2011) Lung Cancer Screening with Low-Dose Helical CT: Results from the National Lung Screening Trial (NLST). *Journal of Medical Screening*, **18**, 109-111. https://doi.org/10.1258/jms.2011.011055
9. Midthun, D.E. (2011) Screening for Lung Cancer. *Clinics in Chest Medicine*, **32**, 659-668. https://doi.org/10.1016/j.ccm.2011.08.014
10. Goo, J.M. (2011) A Computer-Aided Diagnosis for Evaluating Lung Nodules on Chest CT: The Current Status and Perspective. *Korean Journal of Radiology*, **12**, 145-155. https://doi.org/10.3348/kjr.2011.12.2.145
11. Suzuki, K. (2012) A Review of Computer-Aided Diagnosis in Thoracic and Colonic Imaging. *Quantitative Imaging in Medicine and Surgery*, **2**, 163-176.
12. El-Baz, A., Beache, G.M., Gimel'farb, G., Suzuki, K., Okada, K., *et al.* (2013) Computer-Aided Diagnosis Systems for Lung Cancer: Challenges and Methodologies. *International Journal of Biomedical Imaging*, **2013**, Article ID: 942353. https://doi.org/10.1155/2013/942353
13. Retico, A. (2013) Computer-Aided Detection for Pulmonary Nodule Identification: Improving the Radiologist's Performance? *Imaging in Medicine*, **5**, 249-263. https://doi.org/10.2217/iim.13.24
14. Firmino, M., Morais, A.H., Mendoca, R.M., Dantas, M.R., Hekis, H.R. and Valentim, R. (2014) Computer-Aided Detection System for Lung Cancer in Computed Tomography Scans: Review and Future Prospective. *Biomedical Engineering Online*, **13**, 1-16. https://doi.org/10.1186/1475-925X-13-41
15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., *et al.* (2017) Attention Is All You Need.
16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., *et al.* (2020) An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale.
17. Anthimopoulos, M., Christodoulidis, S., Ebner, L., Christe, A. and Mougiakakou, S. (2016) Lung Pattern Classi-

fication for Interstitial Lung Diseases Using a Deep Convolutional Neural Network. *IEEE Transactions on Medical Imaging*, **35**, 1207-1216. https://doi.org/10.1109/TMI.2016.2535865

18. Gao, M., Bagci, U., Lu, L., Wu, A., Buty, M., Shin, H.C., *et al.* (2018) Holistic Classification of CT Attenuation Patterns for Interstitial Lung Diseases via Deep Convolutional Neural Networks. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, **6**, 1-6. https://doi.org/10.1080/21681163.2015.1124249
19. Matsuyama, E., Lee, Y., Takahashi, N. and Tsai, D.Y. (2019) A Wavelet Coefficient-Based Convolutional Neural Network for Histological Classification of Lung Cancer in CT Images. *Japanese Journal of Imaging and Information Sciences in Medicine* (*In Japanese*), **36**, 64-71.
20. Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J. and Oermann, E.K. (2018) Variable Generalization Performance of a Deep Learning Model to Detect Pneumonia in Chest Radiographs: A Cross-Sectional Study. *PLOS Medicine*, **15**, e1002683. https://doi.org/10.1371/journal.pmed.1002683
21. Ovadia, Y., Fertig, E., Ren, J., Nado. Z., Sculley, D., Nowozin, S., *et al.* (2019) Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty under Dataset Shift. 33*rd International Conference on Neural Information Processing Systems*, Vancouver, 8-14 December 2019, 13969-13980.
22. Guo, C., Pleiss, G., Sun Y. and Weinberger, K.Q. (2017) On Calibration of Modern Neural Networks. *Proceedings of the* 34*th International Conference on Machine Learning*, Sydney, Vol. 70, 1321-1330. https://proceedings.mlr.press/v70/guo17a.html
23. Gal, Y. and Ghahramani, Z. (2016) Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *Proceedings of the* 33*rd International Conference on Machine Learning*, Vol. 48, 1050-1059.
24. Chest CT-Scan Images Dataset. https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images
25. He, K., Zhang, X., Ren, S. and Sun, J. (2015) 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
26. Narayanan, B.N., De Silva, M.S., Hardie, R.C., Kueterman, N.K. and Ali, R. (2019) Understanding Deep Neural Network Predictions for Medical Imaging Applications.
27. van der Maaten, L.J.P. and Hinton, G.E. (2008) Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research*, **9**, 2579-2605.
28. 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
29. 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
30. Mannor, S., Peleg, D. and Rubinstein, R. (2005) The Cross Entropy Method for Classification. *Proceedings of the* 22*nd International Conference on Machine Learning*, Bonn, 7-11 August 2005, 561-568. https://doi.org/10.1145/1102351.1102422
31. Brownlee, J. (2020) A Gentle Introduction to Cross-Entropy for Machine Learning. https://machinelearningmastery.com/cross-entropy-for-machine-learning/
32. 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
33. Nova (2023) A Comprehensive Guide to Cross Entropy in Machine Learning. https://aitechtrend.com/a-comprehensive-guide-to-cross-entropy-in-machine-learning/

34. 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