Test-Time Mixup Augmentation for Data and Class-Specific Uncertainty Estimation in Multi-Class Image Classification

Hansang Lee, Member, IEEE, Haeil Lee, Helen Hong, Member, IEEE, and Junmo Kim, Member, IEEE

Abstract—Uncertainty estimation of the trained deep learning network provides important information for improving the learning efficiency or evaluating the reliability of the network predictions. In this paper, we propose a method for the uncertainty estimation for multi-class image classification using test-time mixup augmentation (TTMA). To improve the discrimination ability between the correct and incorrect prediction of the existing aleatoric uncertainty, we propose the data uncertainty by applying the mixup augmentation on the test data and measuring the entropy of the histogram of predicted labels. In addition to the data uncertainty, we propose a class-specific uncertainty presenting the aleatoric uncertainty associated with the specific class, which can provide information on the class confusion and class similarity of the trained network. The proposed methods are validated on two public datasets, the ISIC-18 skin lesion diagnosis dataset, and the CIFAR-100 real-world image classification dataset. The experiments demonstrate that (1) the proposed data uncertainty better separates the correct and incorrect predictions than the existing uncertainty measures thanks to the mixup perturbation, and (2) the proposed class-specific uncertainty provides information on the class confusion and class similarity of the trained network for both datasets.

Index Terms—Uncertainty estimation, Mixup augmentation, Multi-class classification, Skin lesion diagnosis, Convolutional neural network

I. INTRODUCTION

CLASSIFICATION of medical images is an essential task for various applications such as disease diagnosis, surgical decisions, and predictions of treatment responses. Deep neural networks (DNNs) have achieved state-of-the-art accuracy in various image classification tasks and have even shown better performance than human experts for some applications [1], [2]. However, due to the lack of both transparency and explainability of some neural network models and their inability to provide a reliability measure for network predictions, deploying DNNs in real-world applications is limited. In order to improve the reliability of deep learning, uncertainty estimation, which refers to the task of measuring the degree of confidence of a deep learning model regarding its decisions, has received attention. In uncertainty estimation, the uncertainty is calculated as a probability value or score representing how inaccurate the trained network can be with regard to decisions based on test data.

Several methods have been devised to estimate the uncertainty for DNNs [3]. Kendal and Gal proved that the confidence of predictions for a Bayesian neural network directly represents the degree of uncertainty [4]. Gal and Ghahramani suggested the Monte-Carlo drop-out (MCDO) method, which estimates the network-based uncertainty, or the epistemic uncertainty, by perturbing the network through drop-out and measuring the entropy of the prediction results [5]. Wang et al. proposed the test-time-augmentation (TTA) method to measure the data-based uncertainty, or the aleatoric uncertainty, by perturbing the test data through affine transformation and measuring the entropy of the prediction results [6]. In addition to the methods that involve perturbations on the network or the data, Mukhoti et al. proposed the method for measuring the aleatoric and epistemic uncertainty by estimating the density of the features from the feature space [7]. Ghesu et al. trained a network to predict the uncertainty from the network and the data [8].

In this paper, we propose a method for estimating the uncertainty of multi-class image classification with convolutional neural networks using the test-time mixup augmentation (TTMA). Mixup augmentation on training data plays an essential role in regularizing the learner by emphasizing the boundary regions between different classes in the latent feature space [9]. We suggest two uncertainty measures using the TTMA method: the data uncertainty and the class-specific uncertainty. First, we propose a TTMA-based data uncertainty which is computed by performing mixup on the test data with partners uniformly sampled from all training classes and measuring the entropy of the predictions of the mixup data. Because the mixup involves more aggressive perturbation than the affine transformation, the proposed data uncertainty can more accurately distinguish correct predictions from incorrect predictions compared to the TTA-based uncertainty. Second, we suggest a TTMA-based class-specific uncertainty (CSU) which is measured by performing mixup on the test data with partners sampled from a specific training class and computing the entropy of predictions of the mixup data. Unlike the conventional aleatoric uncertainty, CSU can provide network information about the class confusion and class similarity by quantifying the data-based uncertainty associated with a specific class. To validate the effectiveness of the two proposed uncertainty measures, we conduct experiments on two multi-class image classification datasets with different character-
The first dataset is the ISIC-18 skin lesion dataset which shares characteristics with common medical imaging datasets, such as small amount of data, a class imbalance, significant in-class variance, and low between-class variance of the image appearances. The second dataset is the CIFAR-100 natural image dataset which shares characteristics with common natural image datasets, such as a very large amount of data, a large number of classes, and relatively distinctive appearances between different classes. In experiments, the proposed method showed consistent results on both datasets, even given their different characteristics. For both datasets, the proposed data uncertainty outperformed the existing TTA and MCDO uncertainty in terms of its ability to distinguish correct from incorrect prediction results, and the proposed CSU provides information about the class confusion and class similarity of the trained network.

The main contributions of this work are as follows.

- We propose a TTMA-based data uncertainty, an aleatoric uncertainty measure that outperforms existing TTA-based uncertainty in evaluating the reliability of network predictions.
- We propose a TTMA-based class-specific uncertainty, an aleatoric uncertainty measure associated with a specific class, which can provide information about the class confusion and class similarity of the trained network.
- In experiments, we validate the effectiveness of the proposed method on two public datasets with different characteristics.

The remainder of this paper is organized as follows. In Section II, the details of our proposed method are described, with the experimental materials, results, and their evaluations are provided in Section III. In Section IV, further discussions and analyses of our proposed method and contributions are reported, and the paper then concludes with future research directions in Section V.

II. METHODS

A. Test-time mixup augmentation for data uncertainty estimation

The data uncertainty estimation method consists of three steps: (1) test data augmentation with mixup, (2) mixup label prediction and test label inference, and (3) data uncertainty estimation. The pipeline of the proposed data uncertainty estimation method is summarized in Fig. 1.

1) Test data augmentation with mixup: In data uncertainty estimation, we apply mixup augmentation to the test data to obtain perturbation-robust results and to estimate the uncertainty. For a training image-label pair \((s, t) \in S_{train}\) and a test image-label pair \((x, y) \in S_{test}\), we can form a mixup image-label pair \((x_{mk}, y_{mk})\) of a target test data \((x, y)\) and the training data \((s_{mk}, t_{mk})\) as follows. For a given test data \(x\), we form a mixed test data by mixing \(x\) with randomly selected training data \(x_{mk}\) as follows.

\[
x_{mk} = \lambda x + (1 - \lambda)s_{mk} \tag{1}
\]

\[
y_{mk} = \lambda y + (1 - \lambda)t_{mk} \tag{2}
\]

where \((x_{mk}, y_{mk})\) is a image-label pair of mixup data, \((s_{mk}, t_{mk})\) is a image-label pair of \(k\)-th randomly selected data in the training set of class \(m \in \{1, ..., M\}\), and \(\lambda\) is a mixup coefficient determined by the gamma distribution variable \(\lambda \sim \Gamma(\alpha, \alpha)\).
2) Mixup label prediction and test label inference: Our aim is to infer the label $\hat{y}$ of the test data $x$ from the predicted labels $\hat{y}_{mk} = f(x_{mk})$ of the mixup data $x_{mk}$, instead of directly finding the test label $\hat{y} = f(x)$ by applying the trained network to the test data $x$. From Eq. 2, we have

$$y = (y_{mk} - (1 - \lambda)t_{mk}) / \lambda$$  \hspace{1cm} (3)

We can rewrite Eq. 3 by replacing the true labels $y, y_{mk}$ with the inferred label $\hat{y}$ and the predicted label $\hat{y}_{mk} = f(x_{mk})$, respectively.

$$\hat{y}_{mk} = (f(x_{mk}) - (1 - \lambda)t_{mk}) / \lambda$$  \hspace{1cm} (4)

where $\hat{y}_{mk}$ is the inferred label $\hat{y}$ from the mixup data $(x_{mk}, y_{mk})$, and $f(x_{mk})$ is the network prediction of the mixup data $x_{mk}$.

3) Data uncertainty estimation: From Eq. 4, we can infer the test labels $\hat{Y}_{mk} = \{\hat{y}_{mk} | m = 1, ..., M, k = 1, ..., K\}$ from each of mixup data where $K$ is the number of selected training data from each class and $M$ is the number of classes. From the inferred labels of the mixup augmented test data, the final test label $\hat{y}$ can be obtained by majority voting by

$$\hat{y} = \arg \max_l P_l(\hat{Y}_{mk})$$  \hspace{1cm} (5)

where $P_l(\hat{Y}_{mk})$ is a label probability whose label $\hat{y}_{mk}$ is classified as class $l$ among $M \times K$ inferred labels.

The data uncertainty is then computed by the entropy of the distribution of inferred labels $\{\hat{Y}_{mk}\}$ as follows.

$$H(y) = - \sum_{l=1}^{M} P_l(\hat{Y}_{mk}) \ln \left( P_l(\hat{Y}_{mk}) \right)$$  \hspace{1cm} (6)

The proposed data uncertainty represents the instability of predictions of the test data due to perturbations from the mixup with various classes. Because the mixup involves much more aggressive perturbations than the conventional affine-based transformation, the proposed data uncertainty can provide information about the robustness of the trained networks against data-based perturbations in more extreme conditions than those of the existing TTA [6].

B. Test-time mixup augmentation for class-specific uncertainty estimation

The proposed class-specific uncertainty (CSU) estimation method consists of three steps: (1) test data augmentation with mixup, (2) mixup label prediction and test label inference, and (3) the CSU estimation. The pipeline of the proposed CSU estimation method is summarized in Fig. 2.

1) Test data augmentation with mixup: In the CSU estimation, we can define the uncertainty of test data for a specific mixup class $j \in \{1, ..., M\}$ by computing the entropy of class-specific inference results. For a given test data $x$, we form a mixed test data by mixing $x$ with randomly selected training data $s_{jk}$ as follows.

$$x_{jk} = \lambda x + (1 - \lambda)s_{jk}$$  \hspace{1cm} (7)

$$y_{jk} = \lambda y + (1 - \lambda)t_{jk}$$  \hspace{1cm} (8)

where $(x_{jk}, y_{jk})$ is a image-label pair of mixup data, $(s_{jk}, t_{jk})$ is a image-label pair of $k$-th randomly selected data in the training set of class $j$, and $\lambda$ is a mixup coefficient determined by the gamma distribution variable $\lambda \sim \Gamma(\alpha, \alpha)$. 

Fig. 2: A pipeline of the class-specific uncertainty estimation method.
2) **Mixup label prediction and test label inference**: Our aim is to infer the label \( \hat{y} \) of the test data \( x \) from the predicted labels \( \hat{y}_{jk} = f(x_{jk}) \) of the \( j \)-th class mixup data \( x_{jk} \). From Eq. 8 we have

\[
y = (y_{jk} - (1 - \lambda)t_{jk}) / \lambda \tag{9}
\]

We can rewrite Eq. 8 by replacing the true labels \( y, y_{jk} \) with the inferred label \( \hat{y} \) and the predicted label \( \hat{y}_{jk} = f(x_{jk}) \), respectively.

\[
\hat{y}_{jk} = (f(x_{jk}) - (1 - \lambda)t_{jk}) / \lambda \tag{10}
\]

where \( \hat{y}_{jk} \) is the inferred label \( \hat{y} \) from the mixup data \((x_{jk}, t_{jk})\), and \( f(x_{jk}) \) is the network prediction of the mixup data \( x_{jk} \).

3) **Class-specific uncertainty estimation**: From Eq. 10 we can infer the test labels \( \hat{Y}_{jk} = \{\hat{y}_{jk}, k = 1, \ldots, K\} \) from each of mixup data where \( K \) is the number of selected training data from the \( j \)-th class. From the inferred labels of the mixup augmented test data, the final test label \( \hat{y} \) associated with the class \( j \) can be obtained by majority voting by

\[
\hat{y}_j = \arg\max_l P_l(\hat{Y}_{jk}) \tag{11}
\]

where \( \hat{y}_j \) is the class-specific inferred label \( \hat{y} \) associated with the class \( j \), and \( P_l(\hat{Y}_{jk}) \) is a label probability whose label \( \hat{y}_{jk} \) is classified as class \( l \) among \( K \) inferred labels.

The CSU is then computed by the entropy of the distribution of inferred labels \( \{\hat{Y}_{jk}\} \) as follows.

\[
H(y)_j = -\sum_{l=1}^{M} P_l(\hat{Y}_{jk}) \ln \left( P_l(\hat{Y}_{jk}) \right). \tag{12}
\]

The CSU represents the instability of predictions of the test data due to perturbations from the mixup with a specific class.

4) **Interpretation of class-specific uncertainty**: To the best of our knowledge, the proposed CSU is the first uncertainty measure that has test data and a class as variables. Thus, it is essential to interpret which type of information the proposed CSU can provide and to verify it in an experiment. Here, we hypothesize that the CSU provides two types of information about the relationship between the test data and a specific class: the class confusion and class similarity.

First, in terms of class confusion, it can be defined that the network *confuses* two classes \( A \) and \( B \) when the network classifies a large number of class \( A \) data into class \( B \) and vice versa. For instance, a photo of a basketball can be confused with an orange, which has a similar color and shape. In the feature space, class confusion can also be defined as a phenomenon in which the distributions of two classes \( A \) and \( B \) overlap. When class confusion appears, the two classes will have a low average feature distance (AFD) in the feature space. However, the two classes with class confusion will have high CSU because the predictions of mixup data will be evenly distributed in both classes.

Second, in terms of class similarity, it can be defined that class \( A \) is more *class-similar* to class \( B \) than class \( C \) if \( A \) and \( B \) share more appearance or categorical similarity compared to \( A \) and \( C \). For instance, a photo of a dog has a higher class similarity with a cat belonging to the same animal category than an airplane. In the feature space, class similarity can also be defined as the negative of the AFD between the two feature distributions. If \( A \) is more class-similar to \( B \) than \( C \), the AFD between \( A \) and \( B \) will be lower than that between \( A \) and \( C \) in the feature space. The CSU between \( A \) and \( B \) will also be lower than that between \( A \) and \( C \), because the mixup with \( C \) is more aggressive to \( A \) and will more increase the prediction instability compared to the mixup with \( B \).

Class confusion is when the network cannot distinguish two classes, while class similarity is when the network recognizes two classes as similar but can clearly distinguish them. Recognizing and discriminating the class confusion and class similarity is essential in analyzing the performance and behavior of the prediction network. However, the AFD in feature space cannot distinguish class confusion from class similarity because the AFD has low values in both circumstances. The proposed CSU has high values for class confusion and low values for class similarity. Thus, the combination of AFD and CSU makes it possible to discriminate between class confusion and class similarity.

### III. Experiments

#### A. Datasets and implementation

To validate the proposed method, two public datasets, ISIC-18 and CIFAR-100, were used. ISIC-18 [10], [11], [12] is a medical image diagnosis dataset consisting of 10,208 dermoscopic skin lesion images [13], [14] (10,015 for training and 193 for validation) for seven skin disease classes. Table 1 summarizes the statistics of the amount of data for each class in the training and validation sets, and Fig. 3 shows examples of skin lesion images for each class. Classification of skin lesion images is known to be associated with two challenges [15]. First, as shown in Table 1, there is a risk of overfitting due to the small size of the dataset, as well as a risk of learning bias due to the class imbalance. Second, as shown in Fig. 3, the dataset contains examples of both inter-class similarity and intra-class variation [16], where there is a visual difference among the lesions in the same AKIEC class as well as visual similarities in the shape and color between BCC, BKL, and DF.

We trained a VGG-19 [17] on the ISIC-18 dataset for 300 epochs with a mini-batch size of 128. The initial learning rate was 0.01, and the learning rate decayed by 10 after 150 and 225 epochs. During the training process, both affine and mixup augmentation for the training data and drop-out for fully connected layers were applied. In affine data augmentation, a random horizontal flip, a random vertical flip, a random rotation with a degree of \(-45^\circ \leq \theta \leq +45^\circ\), a random translation with shift rates of \((0.1,0.1)\), and a random scaling with a factor of \(1 \leq \sigma \leq 1.2\) were applied. In mixup augmentation, the mixup hyper-parameter \(\alpha\) determining the mixup weights \(\lambda \sim \text{Beta}(\alpha,\alpha)\) was set to 0.2. In drop-out for fully connected layers, the drop-out probability was set to 0.5.

During the test-time, an augmentation method suitable in each case for TTA, MCDO, and TTMA was applied to the
test data, with the parameters of each augmentation method identical to those used during the training process. For TTA, affine augmentation with the same parameters used during training was applied to the test data to generate the augmented test data. For MCDO, drop-out for fully connected layers with a drop-out probability of 0.5 was applied to the trained network to generate the augmented test data prediction. For TTMA, mixup augmentation with $\alpha = 0.2$ was applied to the test data to generate the augmented test data, where the number of selected training data for each class to compute the mixup was set to $K = 30$. Because the total number of mixup augmented test data for TTMA is $MK = 210$, where $M = 7$ denotes the number of ISIC-18 classes, we generated 210 augmented test data per one original test data for TTA and MCDO.

CIFAR-100 [18] is a common natural image classification dataset consisting of a total of 60,000 color images (50,000 for training and 10,000 for validation) for 100 object classes. In the experiment here, a total of 1,000 images (10 images for each class) were randomly selected from 10,000 validation images, and were used for validation.

We trained a Wide Residual Network [19] with a depth of 28 and a widening factor of 10 (WRN-28-10) on the CIFAR-100 dataset for 200 epochs with a mini-batch size of 256. The initial learning rate was 0.1, and the learning rate decayed by 5 after 60, 120, and 160 epochs. During the training process, both affine and mixup augmentation for the training data and drop-out for fully connected layers were applied. In affine data augmentation, random cropping with a square size of 32, a random horizontal flip, a random rotation with a degree of $-45^\circ \leq \theta \leq +45^\circ$, a random translation with shift rates of $(0.1,0.1)$, and a random scaling with a factor of $1 \leq \sigma \leq 1.2$ was applied. In mixup augmentation, the mixup hyperparameter $\alpha$ determining the mixup weights $\lambda \sim \text{Beta}(\alpha, \alpha)$ was set to 0.2. In drop-out for fully connected layers, the drop-out probability was set to 0.3.

During the test-time, an augmentation method suitable in each case for TTA, MCDO, and TTMA was applied to the test data, with the parameters of each augmentation method identical to those used during the training process. For TTA, affine augmentation with the same parameters used during training was applied to the test data to generate the augmented test data. For MCDO, drop-out for fully connected layers with a drop-out probability of 0.3 was applied to the trained network to generate the augmented test data prediction. For TTMA, mixup augmentation with $\alpha = 0.2$ was applied to the test data to generate the augmented test data, where the number of selected training data for each class to compute the mixup was set to $K = 10$. Because the total number of mixup augmented test data in the TTMA is $MK = 1000$, where $M = 100$ denotes the number of CIFAR-100 classes, we generated 1,000 augmented test data per one original test data for TTA and MCDO. For both experiments, we used eight NVIDIA RTX 2080 Ti GPU machines. All algorithms were implemented with PyTorch.

### Table I: Statistics on the amount of data by disease class for the training and validation sets in the ISIC-18 dataset.

| Classes   | Training | Validation |
|-----------|----------|------------|
| AKIEC     | 327      | 8          |
| BCC       | 514      | 15         |
| BKL       | 1099     | 22         |
| DF        | 115      | 1          |
| MEL       | 1113     | 21         |
| NV        | 6705     | 123        |
| VASC      | 142      | 3          |
| Total     | 10015    | 193        |

**B. Evaluation and comparison**

For both data and class-specific uncertainty, we evaluated and compared the proposed methods with different types of uncertainty estimation methods to validate their effectiveness. In terms of the data uncertainty, we compared the aleatoric uncertainty measured by the proposed TTMA method with (1) the aleatoric uncertainty measured by the conventional TTA method [6], and (2) the epistemic uncertainty measured by the conventional MCDO method [5]. We also conducted comparisons of two internal parameters for the proposed TTMA, $\alpha$ and $K$ to analyze parameter sensitivity and optimal parameters, where $\alpha$ is the mixup weight hyperparameter for $\lambda \sim \text{Beta}(\alpha, \alpha)$ and $K$ is a number of selected training data per class for the mixup. With the ISIC-18-trained VGG-19, parameter comparisons were performed for
\( \alpha \in \{0, 0.2, 0.4, 0.6, 0.8, 1\} \) and \( K \in \{10, 20, 30, 40, 50\} \). In the CIFAR-100-trained WRN-28-10, parameter comparisons were performed for \( \alpha \in \{0, 0.2, 0.4, 0.6, 0.8, 1\} \) and \( K \in \{5, 10, 15, 20\} \).

To validate whether the proposed data uncertainty better discriminates correct predictions from incorrect predictions compared to the conventional aleatoric uncertainty, we observed (1) the uncertainty histograms of correct/incorrect test samples, and (2) the accuracy-rejection curves. The uncertainty histograms of correct/incorrect test samples represent two uncertainty distributions of correct and incorrect prediction results. It can be considered that the smaller the overlap between the two test samples in the histogram, the better the uncertainty distinguishes correct predictions from incorrect predictions. In the accuracy-rejection curves, the test data of the top \( T \) uncertainty level were rejected, and the accuracy of the remaining test data with lower uncertainty was measured, where the rejection rates are \( T \in \{0\%, 5\%, 10\%, \ldots, 90\%, 95\%\} \). For instance, 0\% rejection accuracy is the accuracy of the entire test data, whereas 95\% rejection accuracy is the accuracy of the 5\% subgroup of the test data with the lowest uncertainty. The accuracy-rejection curve represents a change in the rejection accuracy according to the rejection rate \( T \). It can be considered that the more the curve monotonically increases and the steeper the slope, the better the uncertainty distinguishes correct predictions from incorrect predictions.

To validate whether the proposed CSU provides information about the class confusion and class similarity as we hypothesized, we compared the CSU for each class with the AFD in feature space. The AFD between one test data and one class was computed as the average of the cosine feature distances between the test data and the training data belonging to the class. We observed the boxplots of CSU and AFD with the feature distributions through tSNE [20] to verify whether the proposed CSU can assist the AFD in discriminating the class confusion and class similarity.

C. Results: Skin lesion diagnosis on ISIC-18

Fig. 4 shows histograms of the aleatoric uncertainty for correct and incorrect test data for (a) TTA, (b) MCDO, and (c) the proposed TTMA data uncertainty. An ideal distribution is one in which the two distributions can be distinguished clearly from each other; The distribution of the correct samples is concentrated in the low uncertainty area, and the distribution of the incorrect samples is concentrated in the high uncertainty area. For TTA and MCDO, the uncertainty distributions of correct and incorrect test samples overlap considerably, making it difficult to differentiate between them. In contrast, the proposed TTMA shows distributions of correct and incorrect samples that are relatively more distinguishable than other methods. Moreover, given that there are only correct samples in the region where the normalized uncertainty is less than 0.3, we can obtain 100\% accurate prediction results from the TTMA data uncertainty through uncertainty thresholding.

Fig. 5 shows the accuracy-rejection curves with (a) different uncertainty estimation methods, (b) different values of \( \alpha \), and (c) different values of \( K \) with the proposed TTMA. Table II summarizes the rejection accuracy for the proposed and comparative methods in Fig. 5. In Fig. 5 (a), TTA shows a curve that is saturated at 96.1\% accuracy at 35\% rejection, whereas MCDO shows a curve that is saturated at a relatively low 91.7\% accuracy rate with 20\% rejection. Starting with 0\% rejection accuracy of 83.9\%, the proposed TTMA shows a curve that increases monotonically as the rejection rate increases, achieving 100\% accuracy at 50\% rejection. In Fig. 5 (b), it can be observed that TTMA with a mixup weight parameter \( \alpha = 0.2 \) shows a curve saturated to 100\% accuracy most rapidly at 50\% rejection, compared to the other choices of \( \alpha \). Fig. 5 (c) indicates that TTMA with \( K = 30 \) results in a curve that is saturated to 100\% accuracy most rapidly at 50\%
Fig. 5: Accuracy-rejection curves of ISIC-18 skin lesion classification results with (a) different uncertainty estimation methods, (b) different values of mixup hyper-parameter $\alpha$, and (c) different values of mixup sampling number $K$ for TTMA.

Fig. 6: tSNE feature distributions of training (marked as o) and validation (marked as x) data in the ISIC-18 skin lesion classification results.

rejection compared to the other choices of $K$.

Fig. 6 shows tSNE feature distributions of the sampled training and the validation data of the ISIC-18 dataset. It can be seen that the feature distributions of four classes, AKIEC, BCC, BKL, and DF, nearly overlap in the lower center, while MEL, NV, and VASC appear densely distributed in different corners. Thus, it can be expected that the four overlapping classes will have high CSU and low AFD values and that the remaining three classes will have CSU distributions similar to those of the AFD.

Fig. 7 shows boxplots of (a) CSU and (b) AFD for the AKIEC disease class. It can be observed that the three classes overlapping with the AKIEC, i.e., MEL, NV, and VASC, the low-to-high value order of AFD is NV < VASC < MEL, which is consistent with that of CSU. Thus, it can be confirmed that the CSU not only provides information about class similarity like AFD but also can distinguish class confusion from class similarity through the combination of CSU and AFD.

D. Results: Natural image classification on CIFAR-100

Fig. 8 shows histograms of the aleatoric uncertainty for correct and incorrect test data for (a) TTA, (b) MCDO, and (c) the proposed TTMA data uncertainty. For TTA and MCDO, many correct samples are distributed in the lowest uncertainty region at [0,0.05], whereas many correct samples are also distributed in the higher uncertainty region. As a result, the uncertainty distributions of correct and incorrect test samples overlap greatly, making them difficult to differentiate. In contrast, the proposed TTMA shows almost no correct samples

| Methods    | 0   | 25  | 50  | 75  | 95  |
|------------|-----|-----|-----|-----|-----|
| Single     | 83.4| 83.4| 83.4| 83.4| 83.4|
| TTA        |     |     |     |     |     |
| MCDO       |     |     |     |     |     |
| TTMA ($\alpha = 0.0$) | 82.4| 89.7| 96.1| 96.1| 96.1|
| TTMA ($\alpha = 0.2$) | 83.9| 89.7| 100 | 100 | 100 |
| TTMA ($\alpha = 0.4$) | 84.5| 90.3| 96.9| 100 | 100 |
| TTMA ($\alpha = 0.6$) | 81.3| 92.4| 96.9| 100 | 100 |
| TTMA ($\alpha = 0.8$) | 79.8| 86.2| 96.9| 100 | 100 |
| TTMA ($\alpha = 1.0$) | 82.9| 91.0| 99.0| 100 | 100 |

TABLE II: Performance evaluation and comparisons of ISIC-18 skin lesion classification with different uncertainty estimation methods. Accuracy (%) was evaluated on the rejected test data for various rejection rates $T \in \{0\%, 25\%, 50\%, 75\%, 95\%\}$.
Fig. 7: Boxplots of (a) class-specific uncertainty (CSU) and (b) average feature distance (AFD) for the AKIEC class in the ISIC-18 classification results.

Fig. 8: Histograms of data uncertainty for correct and incorrect test samples for (a) TTA, (b) MCDO, and (c) TTMA methods.

in the lowest uncertainty region at [0,0.05) and indicates a shift of the overall distribution to the higher uncertainty region, but the distributions of the correct and incorrect samples are relatively distinguishable compared to those of the other methods. Moreover, because there are only correct samples in the region where the normalized uncertainty is less than 0.2, we can obtain 100% accurate prediction results from the TTMA data uncertainty through uncertainty thresholding.

Fig. 9 shows the accuracy-rejection curves for (a) the different uncertainty estimation methods, and with (b) different values of $\alpha$, and (c) different values of $K$ with the proposed TTMA method. Table III summarizes the rejection accuracy for the proposed and comparative methods shown in Fig. 9.

In Fig. 9 (a), TTA shows a curve that increases with a steep slope from low 0% rejection accuracy, whereas MCDO shows a curve that is saturated at 50% rejection accuracy. Starting with 0% rejection accuracy similar to MCDO, the proposed TTMA shows a curve that increases monotonically as the rejection rate increases, achieving the highest 95% rejection accuracy. In Fig. 9 (b), it can be observed that TTMA with a mixup weight parameter $\alpha = 0.2$ achieves the highest 95% rejection accuracy compared to other choices of $\alpha$. Fig. 9 (c) indicates that the accuracy-rejection characteristics of the proposed TTMA data uncertainty are less affected by the number of mixup samples $K$.

Fig. 10 shows tSNE feature distributions of the sampled training and the validation data in CIFAR-100. Unlike ISIC-18, the feature distribution of CIFAR-100 has almost no overlap between the classes, and the data of each class are densely distributed. Thus, it can be expected that the CSU will only provide information about class similarity, not class confusion.

Fig. 11 presents the images of the five low CSU and the five high CSU classes for the five CIFAR-100 test data. It can be observed that the classes with low CSU (1) belong to the same super-class, e.g., lion-tiger and baby-boy, or (2) belong to different classes but have similar appearances in
Fig. 9: Accuracy-rejection curves of CIFAR-100 classification results with (a) different uncertainty estimation methods, (b) different values of mixup hyper-parameter $\alpha$, and (c) different values of mixup sampling number $K$ for TTMA.

IV. DISCUSSION

In this paper, we proposed a method for estimating data and class-specific uncertainty in multi-class image classification using TTMA. Our contributions can be summarized as follows. First, we proposed a data uncertainty estimation method using TTMA and confirmed that the proposed TTMA outperforms the existing uncertainty estimation methods in evaluating the reliability of network predictions. Second, we proposed a novel uncertainty estimation method, termed CSU, using mixup characteristics and confirmed that the proposed CSU represents the class confusion and class similarity of the trained network. Third, we validated the proposed method for two different multi-class image classification datasets and observed that the proposed method showed consistent characteristics in the results for the various datasets.

Table IV shows the expected calibration error (ECE) for the proposed and comparative methods on the ISIC-18 and CIFAR-100 datasets. The ECE is computed as the difference between the prediction confidence and the accuracy [21].
Fig. 11: Examples of CIFAR-100 test classes (a) and the classes with the lowest five CSU (b-f) and the classes with the highest five CSU (g-k). Five example test classes in (a) are orange, lion, baby, leopard, and rose (top to bottom).

Fig. 12: Boxplots of (a) class-specific uncertainty (CSU) and (b) average feature distance (AFD) for "orange" class in the CIFAR-100 classification results. Classes are sorted according to the median values. The ten classes with the lowest medians and the ten classes with the highest medians are shown on the left and right sides, respectively. The classes corresponding to both of the lowest/highest ten in-class CSU and in-class AFD are shown in bold.
ECE is zero when the trained network is calibrated, and the ECE increases as the network becomes more over-confident or more under-confident. In Table V it can be observed that the proposed TTMA shows even more greatly increased ECEs than TTA and MCDO. This is consistent with recent work on the relationship between mixup augmentation, ensemble learning, and network calibration [22]. Wen et al. reported that mixup and ensemble have the effect of lowering the confidence of the over-confident deep neural network and that when both mixup and ensemble are applied, the effects are accumulated, and the network becomes under-confident. This under-confidence characteristic of TTMA can also be observed from the uncertainty distributions in Fig. 6 and Fig. 8. Because the proposed TTMA has a smooth confidence distribution relative to those of TTA and MCDO, the entropy-based uncertainty is also distributed at higher values than those of TTA and MCDO.

In order to explain the reliability of network predictions, one of the critical functions of uncertainty estimations is to distinguish between correct and incorrect predictions based on the estimated uncertainty without the ground truth. An ideal uncertainty estimation method provides low uncertainty values for all correct predictions and high uncertainty values for all incorrect predictions. However, CNN is known to have very severe over-confidence characteristics and thus shows low uncertainty values for most decisions. In order to alleviate this low uncertainty issue, it is necessary to lower the confidence and calibrate the model by applying perturbations to the input data or network or by ensembling the multiple networks for uncertainty estimation. Both TTA and MCDO utilize perturbations and the ensemble through affine augmentation of the data and dropout on the network, respectively, but as shown in the experimental results for both datasets, they could not fully overcome the low uncertainty issue. However, the proposed TTMA lowers the network confidence and increases the uncertainty distribution relative to the other thanks to mixup augmentation, which is a much more aggressive perturbation than an affine transformation in TTA.

In addition to the proposed data uncertainty, which improved the performance of the existing aleatoric uncertainty, we also proposed a novel concept of an uncertainty measure termed CSU and analyzed its behavior. In TTA, affine transformation is applied only to the data alone, whereas in TTMA, mixup augmentation is performed by calculating a mixture with data of different classes. Based on these mixup characteristics, in addition to the case of calculating a mixture of target data and training data from all classes, the case of calculating a mixture of target data and training data from one specific class can be naturally considered. The former will play the same role as the existing aleatoric uncertainty, but the latter can be expected to provide class-specific information related to a partner training class to be mixed with. As expected, in experiments on the two datasets, it was confirmed that the CSU appears to provide information about class confusion and class similarity. In the ISIC-18 dataset with heavy class confusion, the CSU provided class similarity information similar to the AFD for the non-overlapped classes but also showed a different trend from AFD for the overlapped class so that class confusion could be distinguished from class similarity. In the CIFAR-100 dataset with nearly no class confusion, the CSU provided mainly class similarity information similar to the AFD by showing low CSU values for similar semantic or visual classes. It was confirmed that the proposed CSU not only provides information about class similarity similar to the AFD but also enables distinguishing class confusion from class similarity through combination with the AFD.

## Table IV: Expected calibration error (ECE) evaluation and comparison of ISIC-18 and CIFAR-100 classification results with different uncertainty estimation methods.

| Methods | Datasets | ISIC-18 | CIFAR-100 |
|---------|----------|---------|-----------|
| Single  |          | 0.0729  | 0.0992    |
| TTA     |          | 0.0878  | 0.1475    |
| MCDO    |          | 0.1316  | 0.1638    |
| TTMA ($\alpha = 0.0$) |          | 0.1686  | 0.3733    |
| TTMA ($\alpha = 0.2$) |          | 0.2196  | 0.3641    |
| TTMA ($\alpha = 0.4$) |          | 0.1663  | 0.3687    |
| TTMA ($\alpha = 0.6$) |          | 0.1529  | 0.3586    |
| TTMA ($\alpha = 0.8$) |          | 0.0958  | 0.3538    |
| TTMA ($\alpha = 1.0$) |          | 0.1750  | 0.3378    |

## Conclusion

In this paper, we proposed an uncertainty estimation method for multi-class deep learning classification using test-time mixup augmentation. The contributions of the proposed method can be summarized as follows. First, the proposed data uncertainty showed improved reliability evaluations of network predictions compared to existing aleatoric uncertainty, e.g., TTA and MCDO, thanks to the aggressive perturbation characteristics of mixup augmentation. Second, we proposed a novel uncertainty measure called the class-specific uncertainty (CSU), and we confirmed that the CSU provides information about the class confusion and class similarity of the trained network. Third, we validated the proposed method on two different multi-class image classification datasets and confirmed that the proposed method showed consistent performances and behaviors for both datasets. Our future works include improving the performance of TTMA itself by applying mixup variants, e.g., CutMix [23] or AugMix [24], and utilizing the TTMA uncertainty to improve the accuracy of network predictions.

## Acknowledgment

This work was supported by the National Research Foundation of Korea Grant funded by the Korea government (MSIT) (No. 2020R1A2C1102140), and the Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, the Ministry of Food and Drug Safety) (Project Number: 9991007550, KMDF_PR_20200901_0269)
REFERENCES

[1] J. G. Richens, C. M. Lee, and S. Johri, “Improving the accuracy of medical diagnosis with causal machine learning,” Nature Communications, vol. 11, no. 1, p. 3923, Aug 2020.

[2] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O’Donoghue, D. Visentin, G. van den Driessche, B. Lakshminarayanan, C. Meyer, F. Mackinder, S. Bouton, K. Ayoub, R. Chopra, D. King, A. Karthikesalingam, C. O. Hughes, R. Raine, J. Hughes, D. A. Sims, C. Egan, A. Tufail, H. Montgomery, D. Hassabis, G. Rees, T. Back, P. T. Khaw, M. Suleyman, J. Cornebise, P. A. Keane, and O. Ronneberger, “Clinically applicable diagnosis and referral for retinal disease,” Nature Medicine, vol. 24, no. 9, pp. 1342–1350, Sep 2018.

[3] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi, “A review of uncertainty quantification in deep learning: Techniques, applications and challenges,” Information Fusion, 2021.

[4] A. Kendall and Y. Gal, “What uncertainties do we need in bayesian deep learning for computer vision?” in Advances in Neural Information Processing Systems, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, 2017.

[5] Y. Gal and Z. Ghahramani, “Dropout as a bayesian approximation: Representing model uncertainty in deep learning,” in Proceedings of The 33rd International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48, New York, New York, USA, 20–22 June 2016, pp. 1050–1059.

[6] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren, “Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks,” Neurocomputing, vol. 338, pp. 34–45, 2019.

[7] J. Mukhoti, A. Kirsch, J. van Amersfoort, P. H. S. Torr, and Y. Gal, “Deterministic neural networks with appropriate inductive biases capture epistemic and aleatoric uncertainty,” CoRR, vol. abs/2012.11582, 2021.

[8] F. C. Ghesu, B. Georgescu, A. Mansoor, Y. Yoo, E. Gibson, R. Vishwanath, A. Balachandran, J. M. Balter, Y. Cao, R. Singh, S. R. Digumarthy, M. K. Kalra, S. Gribic, and D. Comaniciu, “Quantifying and leveraging predictive uncertainty for medical image assessment,” Medical Image Analysis, vol. 68, p. 101855, 2021.

[9] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” in International Conference on Learning Representations, 2018.

[10] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. K. Mishra, H. Kittler, and A. Halpern, “Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isci),” CoRR, vol. abs/1710.05006, 2017.

[11] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, “Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isci),” in 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 2018, pp. 168–172.

[12] N. C. F. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. W. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. A. Marchetti, H. Kittler, and A. Halpern, “Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isci),” CoRR, vol. abs/1902.03368, 2019.

[13] P. Tschandl, C. Rosendahl, and H. Kittler, “The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions,” Scientific Data, vol. 5, no. 1, p. 180161, Aug 2018.

[14] M. Cobalja, N. C. F. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A. C. Halpern, S. Puig, and J. Malvehy, “Ben20000: Dermoscopic lesions in the wild,” 2019.

[15] J. Zhang, Y. Xie, Y. Xia, and C. Shen, “Attention residual learning for skin lesion classification,” IEEE Transactions on Medical Imaging, vol. 38, no. 9, pp. 2092–2103, 2019.

[16] L. Yu, H. Chen, Q.Dou, J. Qin, and P.-A. Heng, “Automated melanoma recognition in dermoscopy images via very deep residual networks,” IEEE Transactions on Medical Imaging, vol. 36, no. 4, pp. 994–1004, 2017.