How Data Diversification Benefits the Reliability of Three-version Image Classification Systems

Mitsuho Takahashi, Fumio Machida, and Qiang Wen
Department of Computer Science
University of Tsukuba
Tsukuba, Japan
{takahashi.mitsuho@sd.cs, machida@cs, wen.qiang@sd.cs}.tsukuba.ac.jp

***Abstract*— Recently, we have witnessed an increased use of systems employing Machine learning (ML) models. The dependability of such systems is significantly impacted by the inference results of ML models that may not always be correct. N-version ML systems can be adopted for improving output reliability by detecting or correcting errors by combining multiple inference results. In this study, we focus on a three-version ML system that combines the inference results of an ML model on diversified input data to make the system reliable for ML image classification tasks. The three-version image classification system uses only one deep-learning classifier while generating three inference results in response to diversified inputs. We use several image transformation methods to generate diversified input data for inferences. The system reliability is evaluated by the coverage of errors and the certainty of accurate predictions, as those metrics are decision method agnostic. We confirm that appropriately combined inference results from diversified data can increase the coverage of errors and improve reliability while maintaining the certainty of accurate predictions. In order to search for such effective combinations of diversification methods, we propose the neuron coverage improvement rate (NCIR) as an indicator of data diversity. Through the experiments, we show that the NCIR tends to have correlations with the coverage of errors and the certainty of accurate predictions, indicating the usefulness of the indicator.**

Keywords*— *Data diversity, Machine learning system, Neuron coverage, N-version system, Reliability

I. INTRODUCTION

Our daily lives depend heavily on Information and Communication Technology (ICT) systems. In recent years, advances in Machine learning (ML) technology have led to increasing examples of ICT systems using ML, such as autonomous driving, drone autopilot, and face or object recognition. ML models are trained from data and make inferences and predictions on real-world input data, which play an important role in many ML-based systems. However, the inference results of ML models are not always correct. If input data is out of the distributions of the training data set, it may produce erroneous inference results or predictions. In systems that require safety and high reliability, there is a risk that erroneous output from ML models can cause serious problems. For example, a traffic accident can happen if a self-driving car misrecognizes a traffic sign. Therefore, when designing ML systems, it is imperative to consider how to deal with inference errors to avoid undesirable consequences.

To improve the system reliability against ML inference errors, we consider the N-version ML systems that use multiple input data to obtain diverse inference results [1]. We focus on image classification tasks for deep learning ML models and investigate how data diversification methods contribute to the reliability improvement of a three-version image classification system. We apply simple image transformation methods such as shift and rotation to obtain diversified input data from the original input data to be inferred. The obtained diversified input data is fed to the same ML model for inference, and multiple inference results are collected to determine the final output of the system. In this way, even if the ML model misclassifies the original input data, the misclassification can be invalidated or corrected by inference results from other diversified input data. Therefore, finding effective data diversification methods is key to improving system reliability by N-version ML system.

This study aims to investigate the effective combinations of data diversification methods and find a way to search for such desirable combinations in terms of system reliability. First, to analyze the effectiveness of different combinations of data diversifications, we conduct the experiments with an image classification task using two image data sets and two deep neural networks. The evaluation metrics used in the experiments are the coverage of errors and the certainty of accurate predictions that characterize the diversity of multiple inference results in N-version ML systems [2]. We confirm that some combinations are very efficient in improving the coverage of errors without significantly decreasing the certainty of accurate predictions. However, finding such a combination remains an essential challenge. Then, we devise a neuron coverage to score the combination of diversified input data for the N-version ML system using neural networks. Neuron coverage was originally proposed as the test criteria of the neural networks, which is calculated by the ratio of activated neurons for test input data [7]. We exploit the idea and define the neuron coverage improvement rate (NCIR) to measure the degree of diversity in the combination of input data for the neural network. By measuring the NCIR for the same data used in the reliability evaluation, we observe non-negligible correlations among the coverage of errors, the certainty of accurate predictions, and NCIR in most cases. We suggest NCIR as an indicator to find a good combination of data diversification methods for N-version machine learning system.

The rest of the paper is organized as follows. Section II describes the related work. Section III clarifies the motivation of the work. Section IV details the experimental configuration, data sets, and evaluation metrics. Section V explains the results of the reliability evaluation of a three-version image classification system using diversified input data. Section VI shows the results of NCIR and correlation analysis with the reliability metrics. Finally, Section VII describes the conclusion and future work.

II. Related work

This section describes related work on the reliability of N-version ML systems and data diversification techniques used in data augmentation and adversarial training.

A. N-version machine learning system

The N-version ML system is inspired by the concept of N-version programming (NVP). NVP is a software fault-tolerant design method in which N (≥ 2) functionally equivalent programs are independently developed and used from the same initial specification [3]. In programming, software version independence reduces the possibility of the same defect occurring in multiple software versions. N-version ML systems apply this concept to ML systems. While the previous work provides the reliability model for N-version ML systems with two versions, our system focuses on a three-version ML system for image classification tasks. A similar study evaluates an N-version configuration system using several deep neural networks [4]. The experiments in that work considered three independent factors: independent learning, independent networks, and independent training data. The approach using multiple ML models trained from a different batch of training data also resembles ensemble learning [21]. In contrast to these existing studies, we attempt to improve the system reliability by solely using input data diversification at inference time without using diverse ML models.

The reliability of the three-version system using a majority voting scheme has been investigated theoretically in the literature. Triple modular redundancy (TMR) is the low-level use of the 2-out-of-3 voting concept, which employs a voting system in which three modules simultaneously perform the same operation and output most of the same output [11]. If failures of individual versions are not independent of each other, the reliability model for the N-version system can be approximated by using the similarity parameter [20]. The reliability model of the triple-model with triple-input (TMTI) architecture is recently proposed and analyzed using diversity metrics [12]. Our study is complementary to the theoretical studies as we investigate the reliability of the three-version ML system using data diversification techniques through the experiments of image classifiers.

B. Data diversification

Data diversification is a widely used technique to improve the accuracy or robustness of machine learning models. Data augmentation encompasses techniques that enhance the size and quality of training datasets so that better ML models can be built. It is a commonly adopted training method that uses diversified input data at random. AutoAugument is designed to find the best policy to achieve the highest validation accuracy on a target dataset [17]. Mixup aims to train a neural network on convex combinations of pairs of examples and their labels, which increases the robustness against adversarial examples [18]. A mutation-based fuzzing to augment the training data of DNNs is proposed for enhancing robust accuracy. [19]. In our study, data diversification methods are used for the input data instead of training data.

The study of adversarial examples has recently attracted attention in research on the behavior of ML models with intentionally perturbated input data. Adversarial examples are first introduced in an L-BFGS method to fool deep neural networks [5]. It shows that adversarial examples could be generalized to different models and training datasets. The Fast Gradient Sign Method (FGSM) is proposed as a technique to generate adversarial examples faster [13]. Besides, an efficient saliency adversarial map called Jacobian-based Saliency Map Attack (JSMA) is developed so that changes in a small portion of input features could fool the neural network [14][15]. In order to make adversarial examples more natural to human, Generative Adversarial Networks (GANs) is utilized as a part of the approach to generate adversarial examples of images and texts [16]. While adversarial examples are crafted for fooling ML models, in our study, on the contrary, we apply image data transformation to make ML models output correct inference results.

III. Motivation

The previous study has shown that the coverage of errors can be improved by combining different inference results from diversified input data without using different ML models [2]. However, it is not well investigated which diversification methods contribute to improving reliability more significantly. Furthermore, while the reliability improvement also depends on the combination of different diversification methods, there is no existing study to evaluate the impact of the combination of the diversified input data in an N-version ML system. This limitation motivates us to conduct experiments to evaluate the reliability of a three-version ML system using deep neural networks for image classification tasks with various combinations of diversified input data. In particular, we try to answer the following research questions through the experiments.

RQ1: How do input data diversification methods by image transformations impact the reliability of a three-version image classification system?

RQ2: How can we find effective combinations of data diversification methods in terms of output reliability?

To address research question 1, we use two well-known image data sets, MNIST and CIFAR-10, and apply image transformation methods (shift and rotation) to generate diversified input data and obtain three different inference results by the same neural network. The reliability of the three-version classification system is evaluated by the coverage of errors and the certainty of accurate predictions, which are

neutral to the decision logic (i.e., the decision rule is not restricted to majority voting). The experimental results are shown in Section V. On the other hand, to address research question 2, we look at neuron coverage as an indicator of the diversity of input data combinations. We examine how the proposed metric NCIR correlates with the coverage of errors and the certainty of accurate predictions. The results are explained in Section VI.

IV. Experiment Preparation

A. Three-version image classification system

This study evaluates the reliability of a three-version ML system that combines the inference results of three ML modules shown in Figure 1. Each ML module deploys a neural network ML model that performs an image classification task. For input data for classification, handwritten digits from MNIST [8] and the images from CIFAR-10 [9] are used. Two different diversification methods are applied to the MNIST and CIFAR-10 images. In Figure 1, the original data set is denoted as x and the corresponding diversified data sets are represented as x' and x''. The ML modules m_1, m_2, and m_3 classify the given input data independently, and finally integrate the classification results to determine the output using a certain decision method (e.g., majority voting). In this paper, we do not adhere to a specific decision method and instead focus on the diversity and correctness of inference results.

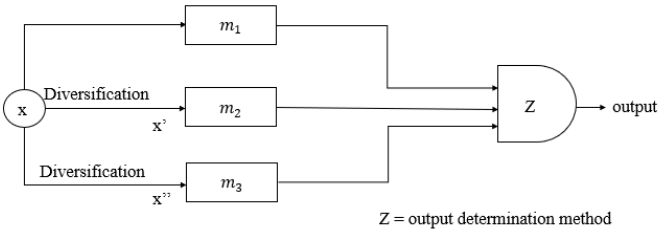

Figure 1. The 3-version image classification system used in the experiment

B. Image classification model

The ML models used in the 3-version image classification system are LeNet [10] and AlexNet [22]. LeNet is the traditional convolutional neural network for image classification tasks. LeNet has a network structure where the convolutional and pooling layers are repeated and connected to all coupled layers. On the other hand, AlexNet is a well-known deep neural network for image classification. Both LeNet and AlexNet are trained on MNIST with 60000 training samples, 128 batch sizes, 10 epochs, and RMSProp as the optimizer. On the other hand, the models for CIFAR-10 are trained with 50000 training samples, 128 batch sizes, 20 epochs, and the optimizer RMSProp. The accuracies of the trained models in the corresponding test data sets are shown in TABLE I.

TABLEI The accuracy of test data for LeNet and AlexNet

	LeNet	AlexNet
MNIST	0.9864	0.9834
CIFAR-10	0.5366	0.4489

C. Input data diversification methods

Two image transformation methods are used to diversify the input data: vertical and horizontal shifting of image data and rotation of image data. The three output results inferred with different input data are combined to evaluate the output results of the entire three-version image classification system. Figure 2 shows samples of transformed MNIST images, and Figure 3 shows the samples of transformed CIFAR-10 images. The first row shows the original image, the second row shows the image shifted vertically and horizontally, and the third row shows the rotated image.

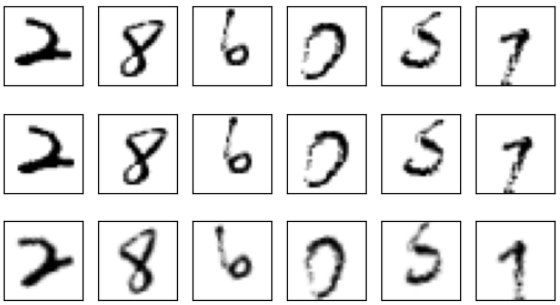

Figure 2. Diversified input data by image transformation in MNIST

Figure 3. Diversified input data by image transformation in CIFAR-10

D. Performance Indicators

We use the coverage of errors and the certainty of accurate predictions as reliability evaluation indices for a three-version image classification system. Coverage of errors is the probability that one or more of the ML modules in a three-version image classification system outputs correct results; even if one of the ML modules outputs incorrect results, the errors can be corrected or nullified using the output results of other ML modules. The higher coverage of errors leads to the lower probability that all ML modules output errors and the more reliable 3-version image classification system. The higher the number of versions becomes, the higher coverage of errors we get.

Let $m_i \in M$ denote the ML modules that comprise the 3-version image classification system. For a total set of input data S, denote $E_i \subset S$ as the data set that m_i outputs incorrectly. The coverage of errors Cov is given by

$$Cov = 1 - \frac{\left|\bigcap_{m_i \in M} E_i\right|}{|S|}. \quad (1)$$

where $\bigcap_{m_i \in M} E_i$ represents the intersection of E_i for $m_i \in M$.

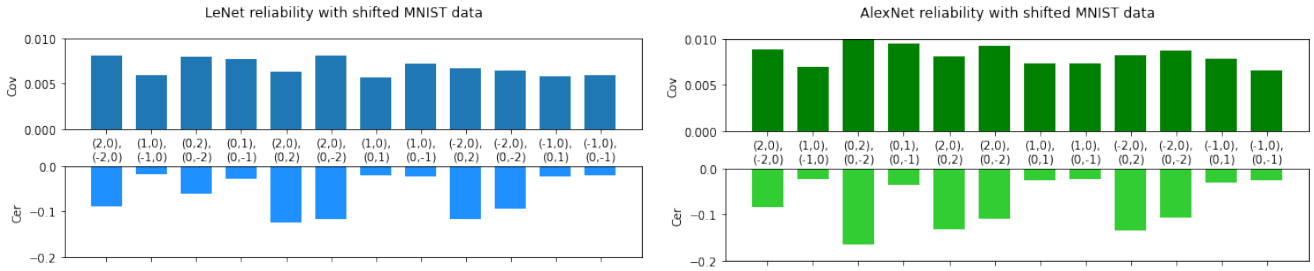

Figure 5. Coverage of errors (Cov) and certainty of accurate prediction (Cer) by suing two shifted MNIST data. The values show the differences from the baseline accuracies (0.9864 for LeNet and 0.9834 for AlexNet)

In contrast, the certainty of accurate predictions represents the probability that all ML models output correct results. Since a three-version image classification system has three ML modules, the certainty of accurate predictions decreases because of the increased probability that any one of the ML modules output an error. It is better that we have a smaller decrease in the certainty of accurate predictions. The certainty of accurate predictions Cer is given by

$$Cer = 1 - \frac{\left|\bigcup_{m_i \in M} E_i\right|}{|S|}. \quad (2)$$

where $\bigcup_{m_i \in M} E_i$ represents the union of E_i for $m_i \in M$. We use these metrics for evaluating the reliability of the output of three-version image classification systems because they are agnostic to the decision method.

V. Reliability Evaluation Results

The reliability evaluation of three-version image classification systems is conducted to address research question 1. We use MNIST and CIFAR-10 data sets and generate diversified data by applying the image transformation methods described above. For various combinations of diversified input data, we compute the coverage of errors and the certainty of accurate predictions.

A. Combination of shifted data

First, we conduct experiments on vertical and horizontal shift diversification methods. Figure 4 shows the results of LeNet's and AlexNet's classification of 10,000 vertically and horizontally shifted MNIST test images. Each bar represents the difference from the baseline accuracies presented in TABLE I. The label (x,y) represents the coordinate of the shift operation, where x and y represent horizontal and vertical shifts in pixels, respectively. As can be seen, the accuracies decrease when we use the shifted data. In particular, the accuracies considerably decrease with the images shifted by 2 pixels.

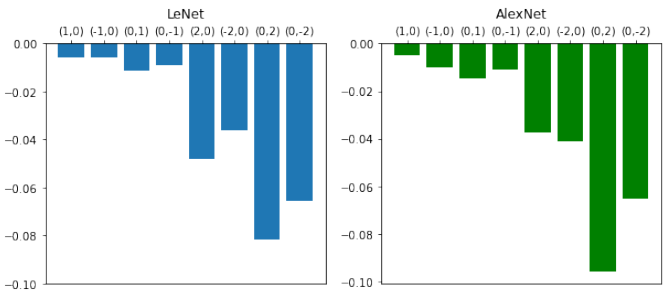

Figure 4 Classification accuracies for shifted MNIST data by LeNet and AlexNet: The values show the difference in the baseline accuracies.

Next, we evaluate a three-version image classification system that combines these inference results. Assuming a three-version image classification system, as shown in Figure 1, the original input data is input to m_1, vertically and horizontally shifted data are input to m_2 and m_3, respectively. All m_1, m_2 and m_3 install LeNet or AlexNet. The coverage of errors and the certainty of accurate predictions are computed and shown in Figure 5. The bar chart plots the differences in the baseline accuracies for individual combinations of input data. The labels on the x-axis represent the sets of coordinates of the selected shift operations. Compared to the baseline, a larger increase in the coverage of errors is favorable, while a smaller decrease in the certainty of accurate predictions is desirable. It can be seen that combining the inference results for the diversified data increases the coverage of errors and decreases the certainty of accurate predictions. There are two major combinations of inference results for shifted data. One is the combination of the inference on vertically or horizontally shifted data, such as up-and-down or right-and-left (e.g., the combination of (2,0) and (-2.0)). The other is the combination of the inference results on vertically shifted data with horizontally shifted data, such as up-and-right and down-and-left (e.g., the combination of (2.0) and (0.2)). The combination of vertically or horizontally shifted data tends to provide higher certainty of accurate predictions and the coverage of errors.

Since Figure 5 shows that the combination of up-and-down and right-and-left shifted data produced better results, we next evaluate combining the inference results of data shifted in the same direction, such as right-to-right, left-to-left, up-and-up, and down-and-down, The results are shown in Figure 6.

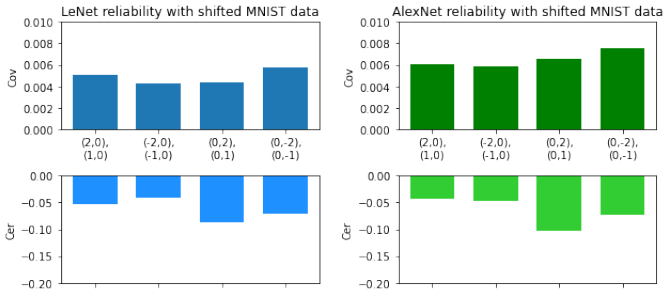

Figure 6 Coverage of errors (Cov) and certainty of accurate prediction (Cer) by using two shifted MNIST data. The values show the differences from the baseline accuracies (0.9864 for LeNet and 0.9834 for AlexNet).

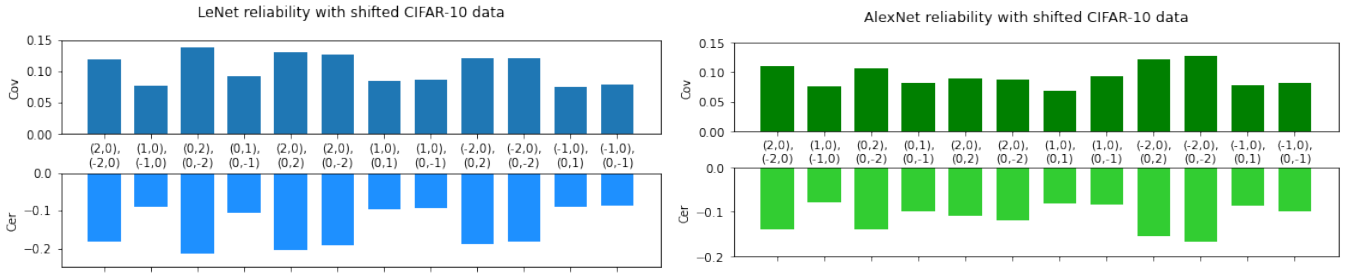

Figure 8. Coverage of errors (Cov) and certainty of accurate prediction (Cer) by suing two shifted CIFAR-10 data. The values show the differences from the baseline accuracies ((0.5366 for LeNet and 0.4489 for AlexNet)

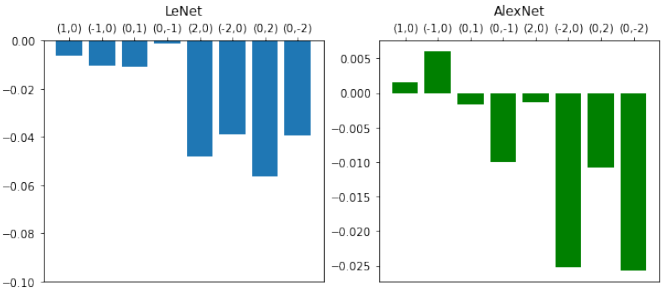

Figure 7 Classification accuracies for shifted CIFAR-10 data by LeNet and AlexNet: The values show the difference from the baseline accuracies.

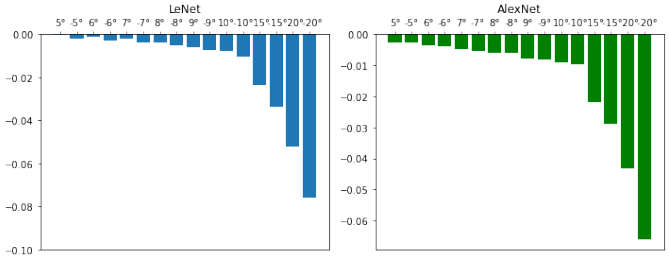

Figure 9 Classification accuracies for rotated MNIST data by LeNet and AlexNet: The values show the difference from the baseline accuracies.

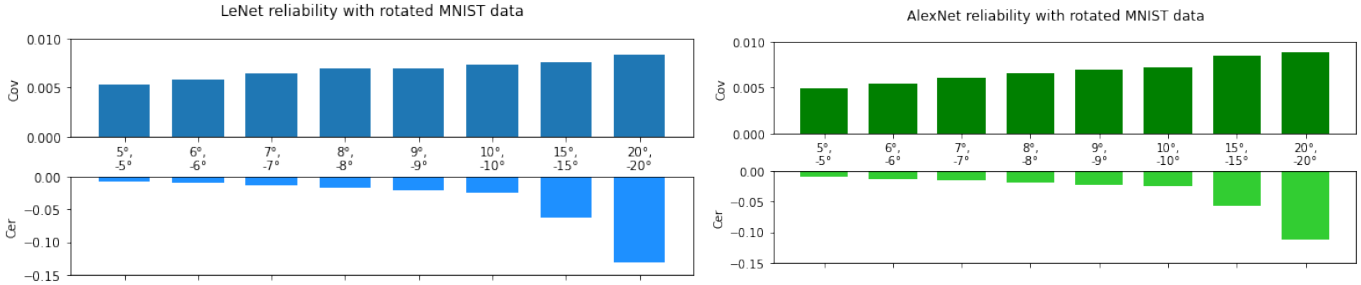

Figure 10. Coverage of errors (Cov) and certainty of accurate prediction (Cer) by suing two rotated MNIST data. The values show the differences from the baseline accuracies (0.9864 for LeNet and 0.9834for AlexNet)

Compared with the results in Figure 5, combining the inference results for data shifted in the same direction has no significant improvement in the coverage of errors.

The same experiment is performed for CIFAR-10. Figure 7 shows the accuracies of LeNet and AlexNet's classifications for 10,000 test images of CIFAR-10 shifted vertically and horizontally. While the results of LeNet are similar to those observed for MNIST, the accuracies of AlexNet are increased in some cases (shift with (1,0) and (-1,0)). We can expect improved system reliability by using these diversified images as input data.

Figure 8 shows the reliabilities of a 3-version image classification system with different combinations of shifted CIFAR-10 images. Similar to the results for MNIST, combining inference results from diversified input data increases the coverage of errors and decreases the certainty of accurate predictions. However, the low certainty of accurate predictions indicates that the system becomes less reliable.

B. Combinations of rotated data

Next, we conduct experiments on diversification methods that rotate the images by a certain angle. Figure 9 shows the results of LeNet's and AlexNet's classification of 10,000 rotated MNIST test images. As can be seen, the classification accuracies decrease when using the rotated data. However, compared with the shifted data, the decrease is smaller, and the number of correct outputs has not decreased significantly when the rotation angle is small.

Next, we evaluate the three-version image classification system that combines these inference results. Assuming a three-version image classification system, as shown in Figure 1, the original input data is input to m_1, right and left rotated data are input to m_2 and m_3, respectively. All m_1, m_2 and m_3 install LeNet or AlexNet. The coverage of errors and the certainty of accurate predictions are computed and shown in Figure 10. It shows that the combined inference results for the rotated data increase coverage of errors and decrease the certainty of

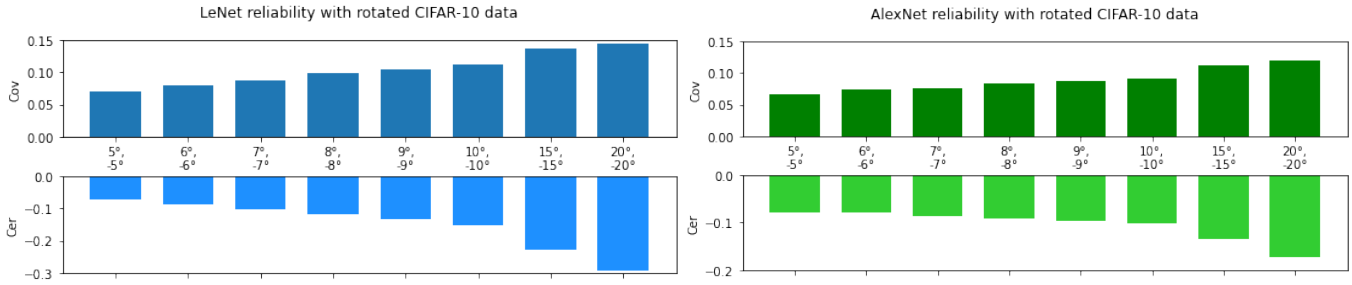

Figure 12. Coverage of errors (Cov) and certainty of accurate prediction (Cer) by suing two rotated CIFAR-10 data. The values show the differences from the baseline accuracies ((0.5366 for LeNet and 0.4489 for AlexNet)

accurate predictions. When combining the rotated images larger than 10° or smaller than -10°, the certainty of accurate predictions decreases significantly.

The same experiment is then performed for CIFAR-10. Figure 11 shows the accuracies of LeNet and AlexNet's for 10,000 test images of CIFAR-10 rotated data. The accuracies are generally lowered by using the rotated data like the cases of vertically and horizontally shifted data. Interestingly, however, the accuracies of AlexNet slightly improve in some cases, especially when the degree of rotation is not so large.

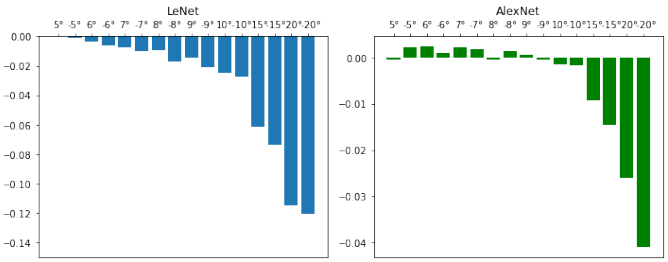

Figure 11 Classification accuracies for shifted CIFAR-10 data by LeNet and AlexNet: The values show the difference in the baseline accuracies.

Finally, we evaluate the three-version image classification system that combines these inference results. The results are shown in Figure 12. As can be observed, combining the inference results with the rotated data increases the coverage of errors and decreases the certainty of accurate predictions. However, similar to the vertically and horizontally shifted data, the certainty of accurate predictions is too low to be considered reliable, especially when the rotation angle is large.

C. Summary

In this experiment, we use image transformation methods to generate diversified versions of MNIST and CIFAR-10 images and investigate the improved reliabilities by the three-version image classification systems using LeNet and AlexNet. We confirm that different combination of data diversification methods has different effects on the increasing coverage of errors and decreasing the certainty of accurate predictions. For shift operation, combining the inference results of vertical and horizontal directions can improve the coverage of errors while maintaining high certainty of accurate predictions, considered good combinations. However, combining inference results for shifted data in the same direction, such as up-and-up or right-and-right, does not make a significant increase in the coverage of errors. As for the rotation method, it is found that the coverage of errors increases as the angle of rotation increases. However, when the rotation angle increases more than 10° or -10°, the certainty of accurate prediction tends to decrease significantly. Compared with the results of CIFAR-10, the evaluation results of MNIST show smaller decreases in the certainty of accurate predictions, implying that the three-version architecture is more effective when applied to MNIST.

Answer to RQ1: Diversification of input data by image transformation has improved the reliability of three-version image classification systems in MNIST and CIFAR-10. However, excessive image transformation leads to a significant decrease in the certainty of accurate predictions.

In order to find an effective combination of diversification methods for improving the reliability of the three-version image classification system, one must conduct these experiments to evaluate the coverage of errors and the certainty of accurate predictions for any possible combinations. Such experiments are not efficient in a practical scenario. Any efficient methods to search for a good combination of diversification methods are needed. To this end, we look into the internal states of neural networks when predicting the labels with the diversified images, as we focus on deep neural networks as classifiers in this study. Inspired by the previous study that evaluates the neuron coverage of diversified testing samples [6], we leverage the neuron coverage to measure the diversity of inferences for a combination of diversified input data. Suppose the coverage of errors and the certainty of accurate predictions have any relation to the neuron coverage. In that case, we may exploit the neuron coverage to search for effective combinations for three-version image classification systems. Thus, we focus on the evaluation of neuron coverage in the next section.

VI. Evaluation of Neuron Coverage Improvement Rate

This section investigates the relationship between the reliability improvement of data diversification by image transformation methods and the neuronal coverage improvement rate.

A. Neuron coverage

In neural network models, when given input data, the neurons in the network output a certain vector value for that

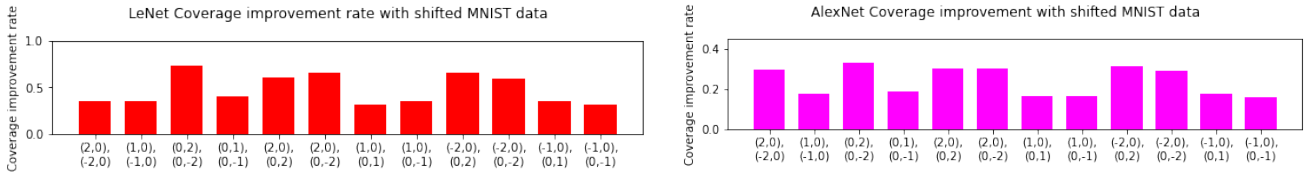

Figure 13. NCIR for the three-version image classification system using shifted MNIST data

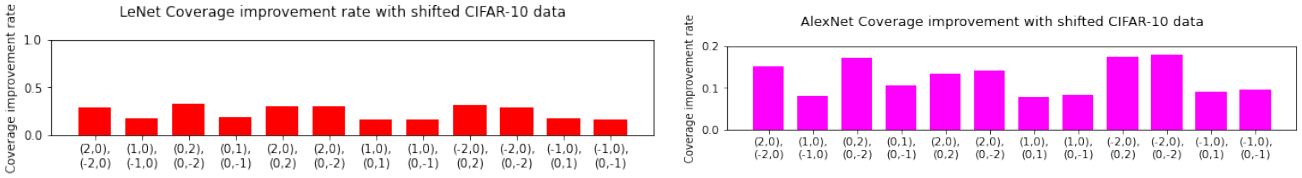

Figure 14. NCIR for the three-version image classification system using shifted CIFAR-10 data

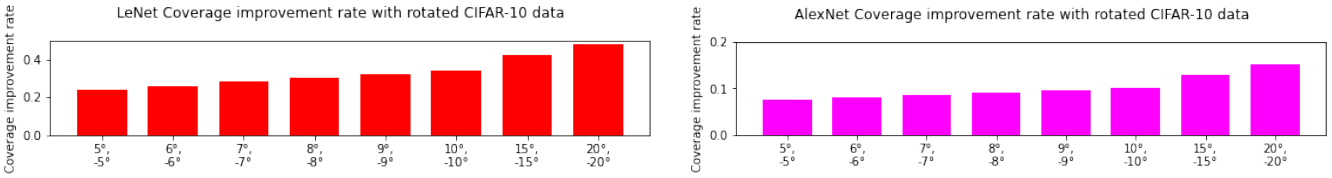

Figure 15. NCIR for the three-version image classification system using rotated MNIST data

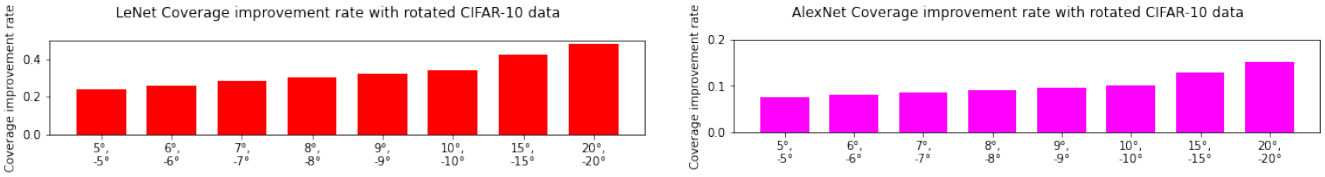

Figure 16. NCIR for the three-version image classification system using rotated CIFAR-10 data

input data. When this vector value exceeds a threshold value, the neuron is considered to be activated. As defined below, the ratio of activated neurons to all neurons for given test samples is called neuron coverage.

$$Neuron\ Coverage = \frac{Activated\ neurons}{All\ neurons}. \quad (3)$$

Neuron coverage has attracted attention as an indicator for evaluating the robustness of neural networks and has been utilized in neural network testing methods [7]. Meanwhile, recent studies have doubted its usefulness in testing [23][24].

B. Neuron coverage improvement rate

It is known that testing neural networks with diversified input data improves neuronal coverage [6]. Therefore, the internal states of neuron coverage potentially indicate the degree of diversity for given diversified input data sets. We consider using neuron coverage to measure the diversity of input data combinations used in a three-version ML system. We investigate how different diversified data combinations affect neuronal coverage's value. To measure the difference in the neuronal coverage before and after adding inferences with diversified input data, we define the neuron coverage improvement rate (NCIR). The NCIR is computed by the following procedure.

1. Measure the output value $v(x)$ of each neuron when classifying a set of input data x.
2. Compute the neuron coverage $N(x)$ with the threshold τ for the neural network with N neurons.
$$N(x) = \frac{1}{N}\sum_{v(x)\geq\tau} 1.$$
3. Generate the diversified input data sets x' and x'' from x. Then, measure the output values $v(x')$ and $v(x'')$ of each neuron when classifying x' and x'', respectively.
4. Compute $v_{\max}(x,x',x'') = \max\big(v(x), v(x'), v(x'')\big)$.
5. Compute the neuron coverage $N(x,x',x'')$ for the three-version image classifier using x, x', x'' as input data.
$$N(x,x',x'') = \frac{1}{N}\sum_{v_{\max}(x,x',x'')\geq\tau} 1.$$
6. Compute NCIR by comparing $N(x)$ and $N(x,x',x'')$.
$$NCIR = \frac{N(x,x',x'') - N(x)}{N(x)}.$$

TABLE II. CORRELATION MATRICES BY VERTICAL AND HORIZONTAL SHIFTED IMAGE DATA

LeNet - MNIST

	NCIR	Cov	Cer
NCIR	1		
Cov	0.410167	1	
Cer	-0.76677	-0.3687	1

AlexNet - MNIST

	NCIR	Cov	Cer
NCIR	1		
Cov	0.692934	1	
Cer	-0.96591	-0.68374	1

LeNet – CIFAR-10

	NCIR	Cov	Cer
NCIR	1		
Cov	0.985996	1	
Cer	-0.98808	-0.98946	1

LeNet – CIFAR-10

	NCIR	Cov	Cer
NCIR	1		
Cov	0.890488	1	
Cer	-0.948481	-0.939934	1

TABLE III. CORRELATION MATRICES BY ROTATED IMAGE DATA

LeNet - MNIST

	NCIR	Cov	Cer
NCIR	1		
Cov	0.922147	1	
Cer	-0.94954	-0.7973	1

AlexNet - MNIST

	NCIR	Cov	Cer
NCIR	1		
Cov	0.95038	1	
Cer	-0.968351	-0.85388	1

LeNet – CIFAR-10

	NCIR	Cov	Cer
NCIR	1		
Cov	0.985971	1	
Cer	-0.99642	-0.96877	1

LeNet – CIFAR-10

	NCIR	Cov	Cer
NCIR	1		
Cov	0.979287	1	
Cer	-0.995203	-0.955229	1

C. Results

First, the NCIRs are computed for three-version image classification systems using vertically and horizontally shifted data. We use the same configurations as explained in section V-A. The neuron coverages are computed with respect to the threshold $\tau = 0.2$ (following the previous literature [6]). The computed NCIRs for MNIST and CIFAR-10 are shown in Figure 13 and Figure 14, respectively. We observe that the combinations using 1px shifted data generally have smaller NCIRs than the combinations using 2px shifted data, meaning that more neurons are activated by the data with larger changes from the original data. This trend has a certain similarity to what was observed in the evaluation of the coverage of errors and the certainty of accurate predictions presented in Figure 5 and Figure 8.

Next, the NCIRs are computed for three-version image classification systems using rotated data of MNIST and CIFAR-10. We use the same configurations as explained in section V-B. The computed NCIRs for MNIST and CIFAR-10 are shown in Figure 15 and Figure 16, respectively. Similar to the trends observed in the coverage of errors and the certainty of accurate predictions shown in Figure 10 and Figure 12, the NCIRs tend to increase as the angle of rotation increases. The results motivate us to carry out the correlation analysis between the coverage of errors, the certainty of accurate predictions, and the NCIR.

D. Correlation analysis

Following the observations in the NCIR results, we conduct the correlation analysis among the coverage of errors, the certainty of accurate predictions, and NCIR. The correlation matrix is summarized in TABLE II and TABLE III. Each entry of the matrix shows the correlation coefficient between two measures for comparison. At a glance, we can see strong correlations in most entries as their absolute values are larger than 0.5, except for the case with the LeNet applied to MNIST. As the coverage of errors and NCIR have positive correlations, we can use NCIR as the indicator to choose the diversified data for improving the coverage of errors. On the other hand, negative correlations are observed between the certainty of accurate predictions and NCIR. This implies that NCIR also gives an indicator of the certainty of accurate predictions such that a higher NCIR may compromise the certainty. The relation is also confirmed by the negative correlation between the coverage of errors and the certainty of accurate predictions, which reveals the trade-off between the two measures. The best trade-off may depend on the application as well as the decision method. We can conclude that the NCIR effectively represents the coverage of errors and the certainty of accurate predictions observed in three-version image classification systems using the diversified input data.

Answer to RQ2: NCIR can be used as an indicator to search for effective combinations of diversified input data for three-version image classification systems. The higher NCIR likely improves the coverage of errors and decreases the certainty of accurate predictions.

Considering MNIST data, Figure 13 shows that the highest NCIRs are observed when the inference results for the original data, the data shifted up 2 pixels and the data shifted down 2 pixels are combined (the bar labeled with ((0,2),(0,-2))). This value is higher than the NCIR at the inference results of the data with the largest rotation combined (the bar labeled with (20°,-20°) in Figure 15). The certainty of accurate predictions drops significantly when the angle of rotation is large, as observed in Figure 10. In contrast to this, Figure 5 shows that the combination of data shifted 2 pixels up and 2 pixels down have no significant decrease in the certainty of accurate predictions. From this observation, for MNIST data, it is considered the best option to combine the inference results of the original data with those of the data shifted 2 pixels up and 2 pixels down, as it increases the coverage of errors while minimizing the decline in the certainty of accurate predictions.

VII. CONCLUSION

This paper investigated how data diversification by image transformations can improve the reliability of three-version image classification systems. Experiments are conducted to evaluate the coverage of errors and the certainty of accurate predictions for the images of MNIST and CIFAR-10. The results show that combining the inference results with diversified input data leads to higher reliability of the system. However, a significant transformation (e.g., 20° of rotation) considerably reduces the certainty of accurate predictions,

resulting in the ineffectiveness of the three-version image classification system. In order to search for effective combinations of data diversification methods, we proposed NCIR as an indicator to represent the diversity of input data combinations. Our correlation analysis revealed that NCIR had correlations with the coverage of errors and the certainty of accurate predictions. Therefore, NCIR can be used as the indicator to choose the effective combinations of diversification methods.

We can extend our work in several directions. It is possible to conduct similar experiments on more realistic image classification tasks such as traffic signs recognition, object recognition, and face recognition. Data diversification methods can also be extended by considering more natural perturbations. For example, the effects of various input data (cloudiness, cracks, reflections due to light, etc.) could be applied. While the reliability can be improved by using multiple inference results, the trade-off with other quality measures, such as performance and energy consumption, might be a major concern in system engineering. Designing high-performance and energy-efficient systems while maintaining system reliability is also a future challenge.

ACKNOWLEDGEMENT

This work is partly supported by JSPS KAKENHI Grant Numbers 19K24337 and 22K17871.

REFERENCES

[1] F. Machida, N-version machine learning models for safety critical systems, In Proc. of the DSN Workshop on Dependable and Secure ML, pp. 48-51, 2019.

[2] F. Machida, On the diversity of machine learning models for system reliability, In Proc. of the 24th Pacific Rim International Symposium on Dependable Computing, pp. 276-285, 2019.

[3] A. Avizienis, The methodology of n-version programming, Software fault tolerance, Vol. 3, pp. 23-46, John Wiley & Sons, New York, 1995.

[4] H. Xu, Z. Chen, W. Wu, Z. Jin, S. Kuo, M. Lyu, NV-DNN: Towards Fault-Tolerant DNN System with N-Version Programming, In Proc. of the DSN Workshop on Dependable and Secure machine learning, pp. 44-47, 2019.

[5] I. Goodfellow, J. Shlens and C. Szegedy, Explaining and harnessing adversarial examples, arXiv preprint arXiv:1412.6572, 2014.

[6] Y. Tian, K .Pei, S. Jana and B. Ray, DeepTest: Automated Testing of Deep-Neural-Network-driven Autonomous Cars, In Proc. of the 40th International Conference on Software Engineering, pp 303-314, 2018.

[7] K. Pei, Y. Cao, J. Yang, and S. Jana, DeepXplor: Automated Whitebox Testing of Deep Learning Systems, In Proc. of the 26th Symposium on Operating Systems Principles, pp. 1-18,2017.

[8] Y. LeCun, C. Cortes, and C. Burges, The MNIST database of handwritten digits, http://yann.lecun.com/exdb/mnist/, retrieved in May, 2019.

[9] A. Krizhevsky. Learning multiple layers of features from tiny images. Master's thesis, Department of Computer Science, University of Toronto, 2009.

[10] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In Proc of the IEEE, 1998.

[11] L. Robert E and W. Vanderkulk, The use of triple-modular redundancy to improve computer reliability, IBM journal of research and development, Vol. 6, No. 2, pp. 200-209, 1962.

[12] Q. Wen, F. Machida, Reliability Models and Analysis for Triple-model with Triple-input Machine Learning Systems, In Proc. of the 5th IEEE Conference on Dependable and Secure Computing, 2022

[13] I. J. Goodfellow, J. Shlens, and C. Szegedy, Explaining and harnessing adversarial examples, arXiv preprint arXiv:1412.6572, 2014.

[14] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, The limitations of deep learning in adversarial settings, In 2016 IEEE European symposium on security and privacy (EuroS&P), pp. 372-387, 2016.

[15] N. Carlini and D. Wagner, Towards evaluating the robustness of neural networks, In Proc. of the 2017 IEEE Symposium on Security and Privacy (S&P), pp. 39–57, 2017.

[16] Z. Zhao, D. Dua, and S. Singh, Generating natural adversarial examples, arXiv preprint arXiv:1710.11342, 2017.

[17] E. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. Le, Autoaugment: Learning augmentation strategies from data, In Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 113-123, 2019

[18] Z. Hongyi, C. Moustapha, N. Yann, and L. David, Mixup: Beyond empirical risk minimization, In Proc. of the International Conference on Learning Representations (ICLR), 2017.

[19] X. Gao, R. Saha, M. Prasad, and A. Roychoudhury, Fuzz testing based data augmentation to improve robustness of deep neural networks. In Proc. of 42nd International Conference on Software Engineering (ICSE), pp. 1147-1158, 2022.

[20] M. Ege, A. Eyler M and MU. Karakas, Reliability analysis in N-version programming with dependent failures, In Proc. of IEEE EUROMICRO Conference, pp. 174-181, 2001.

[21] T. Dietterich, Ensemble methods in machine learning, In Proc. of the international workshop on multiple classifier systems, pp. 1-15, 2000.

[22] A. Krizhevesky, I. Sutskever, and G. E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, In Proc. of NIPS, 1097-1105, 2012.

[23] F. Harel-Canada, L. Wang, M. A. Gulzar, Q. Gu, and M. Kim, Is neuron coverage a meaningful measure for testing deep neural networks?, In Proc. of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE), pp. 851-862, 2020.

[24] S. Yan, G. Tao, X. Liu, J. Zhai, S. Ma, L. Xu, and X. Zhang, Correlations between deep neural network model coverage criteria and model quality, In Proc. of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE), pp. 775-787, 2020.