Evaluating the Performance of the Model Selection with Average ECE and Naive Calibration in Out-of-Domain Generalization Problems for Binary Classifiers

Anxian Liu
Supervisors: Jesse Krijthe, Rickard Karlsson, Stephan Bongers
EEMCS, Delft University of Technology, The Netherlands

June 19, 2022

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

Abstract

Out-of-domain (OOD) generalization refers to learning a model from one or more different but related domain(s) that can be used in an unknown test domain. It is challenging for existing machine learning models. Several methods have been proposed to solve this problem, and multi-domain calibration is one of these methods. Model selection with the average expected calibration error (ECE) across training domains and naive calibration are two approaches to implementing multi-domain calibration. However, it might happen that neither approach can learn a genuinely well-calibrated model in the multi-domain setting. Hence, this paper intends to evaluate how naive calibration and model selection with average ECE perform in the OOD generalization problem for binary classifiers. We generated many synthetic datasets and set up three experiments to answer this question. Finally, the conclusions based on empirical results are obtained: 1) Although naive calibration can improve the average accuracy across unseen domains (OOD accuracy) and the average area under the ROC Curve across unseen domains (OOD AUROC) for some binary classifiers, it does not work for all binary classifiers. However, at least it does not make the model worse for OOD generalization. 2) On the synthetic datasets we generated, if the number of training domains increases, most binary classifiers' OOD accuracy will also increase. 3) Average ECE is a reasonable metric for selecting a model in the OOD generalization problem and is better than validation accuracy. This is because a strong linear relationship exists between OOD accuracy and the average ECE across the training domains. This linear relationship is stronger than the linear relationship between OOD accuracy and validation accuracy.

1 Introduction

Under the assumption that the training and test data come from similar distributions, existing machine learning algorithms have had much success [1]. However, if the distributions between the training dataset and the testing dataset are different, a machine learning model may fail [1–3]. This frequently occurs in the real world because there is no guarantee that the training and deployment environments are always the same. This is known as the out-of-domain (OOD) generalization problem, which refers to one trying to learn a predictive model that generalizes across different environments (known as domain).

Several methods have been proposed to address this problem. Anchor regression, a regularisation approach for fitting linear models, was introduced in 2018. Still, it only applies to linear models and relies on the assumption of linearity [4]. Arjovsky et al. [5] proposed Invariant Risk Minimization (IRM), a learning method for discovering nonlinear invariant predictors across multiple environments. IRMv1 was established due to the difficulty of optimizing the original IRM objective. However, it has been shown that it does not capture the entire set of invariances [6]. Recently, multi-domain calibration was proposed by Wald et al. [7], which is a more straightforward form of invariance than the approaches above.

The following paragraph explains how multi-domain calibration resolves the OOD generalization problem. Popular approaches to solve the OOD generalization problem are based on the hypothesis that such a model is an invariant predictor that captures the mechanism that remains the same across domains [8]. More specifically, a classifier is invariant across different domains if its prediction is independent of the spurious features and only depends on the invariant features. In this context, spurious features are the features that can help predict the label in some environments, but these correlations do not hold in all domains. On the contrary, invariant features are the features that can help predict the label across

all domains, and these correlations remain. Wald et al. [7] indicated that if a classifier is calibrated on all training environments, it is also invariant w.r.t the training environments. However, many contemporary machine learning models are not designed for determining precise probabilities. Calibration is a technique for obtaining accurate probability estimation for models in classification tasks [9]. In the multi-domain setting, calibration refers to learning a model calibrated in different domains, which means that the model's output confidence should equal the actual probability of the corresponding label occurring in every domain. Lastly, Wald et al. [7] proved that calibration could generalize across domains if the number of domains is roughly equal to the number of spurious features in the linear-Gaussian models and population (as opposed to finite-sample) setting. The relationship between multi-domain calibration and the OOD generalization problem is thus established.

Three approaches were proposed by Wald et al. [7] that can achieve multi-domain calibration: (i) model selection with the average expected calibration error (ECE), (ii) post-processing tools, including naive calibration and robust calibration, and (iii) learning with Multi-Domain Calibration Error. However, the first two approaches have limitations in their ability to learn a properly calibrated model across multiple domains [7]. Hence, it is interesting to evaluate the performance of these two approaches.

Our main research question is how well naive calibration and model selection with average ECE perform in the OOD generalization problem for binary classifiers? The sub-questions are listed below to be more specific.

- Does naive calibration improve average prediction performance, as measured in the accuracy or area under the ROC Curve (AUROC), across unseen domains?
- Does OOD Accuracy improve as the number of training domains grows?
- Is model selection with average ECE a reasonable model selection strategy in the OOD generalization problem?

Multiple synthetic datasets for the OOD generalization problem were created to solve these questions. Three experiments were designed to address these sub-questions. More details in Section 4 and Section 5.

The contributions are: we implemented the naive calibration for the linear/non-linear models and investigated the performance of the naive calibration for binary classifiers. We discovered that increasing the number of training domains improves OOD accuracy. Lastly, we found the correlation between OOD accuracy and the average ECE across training domains. It implies that model selection based on average ECE is a reasonable selection strategy in the OOD generalization problem, which is explained in Section 5.4 and Section 7.

The structure of this paper is as follows: Section 2 will talk about the related work. Section 3 will describe the details of naive calibration and model selection with average ECE. Section 4 will present the research methodology. Based on Section 4, Section 5 will illustrate the experiments and corresponding results. Section 6 will talk about responsible research. The discussion will be shown in Section 7. Section 8 will present the conclusions and future work.

2 Related work

There are several research fields closely related to naive calibration and model selection with average ECE, including but not limited to: calibration methods and model selection strate-

gies for domain generalization tasks. These are summarized in the following subsections.

2.1 Calibration method

Several notable works in the field of calibration have also been completed. Platt scaling is a method for converting the outputs of a classification model into a probability distribution across classes by fitting a logistic regression model to a classifier's outputs [10]. Although Platt scaling was initially designed to scale the outputs of SVMs, it has also been shown to have a good performance on the calibration problem for boosted models, and naive Bayes classifiers [11]. Isotonic Regression Scaling is a widely used calibration approach for binary classifiers in a single domain [12, 13]. Instead of logistic regression, Isotonic Regression Scaling fits a piecewise-constant non-decreasing function to a classifier's outputs. Naive calibration and robust calibration are proposed by Wald et al. [7] to extend the isotonic regression scaling in the multi-domain setting.

2.2 Model selection strategies

Gulrajani and Lopez-Paz [14] claimed that model selection strategies are essential in domain generalization tasks and compared different model selection strategies. Instead of using model selection strategies proposed by Gulrajani and Lopez-Paz [14], Wald et al. [7] suggested that selecting the model with the lowest ECE from those that achieve In-Domain (ID) validation accuracy above a user-defined threshold. ECE measures the difference between expected accuracy and expected confidence [9]. A few enhancements to current ECE estimators were introduced by Posocco and Bonnefoy [15]. However, the reliability diagram demonstrated the traditional method for calculating ECE, which is used in this study [16, 17].

3 Multi-domain calibration

Naive calibration and model selection with average ECE, which do not necessitate training a new model, are simple to use and help improve multi-domain calibration. These methods are briefly outlined in this section.

3.1 Naive calibration

Isotonic Regression Scaling is the most used technique to calibrate the model, requiring that the mapping function is isotonic. The pool adjacent violators (PAV) algorithm is a frequently employed method for calculating isotonic regression. The PAV algorithm-based isotonic calibration method can be thought of as a binning process, with the position of boundaries and bin sizes being determined by how accurately the classifier ranks the instances in the training data [9]. It was designed for the single domain problem, but Wald et al. [7] extended it to the multi-domain setting. The following are the detailed steps of naive calibration:

1. Split the data from all the training domains into the training, validation, and test data.

2. Train the binary classifier f on the training data and evaluate the In-Domain (ID) accuracy on the test data.

3. Take predictions of f on validation data and fit an isotonic regression z^*.

4. Report its performance of $z^* \circ f$ on the data from testing domains.

3.2 Model selection with average ECE

Model selection in domain generalization is more complicated than supervised learning since we do not have access to a validation set distributed similarly to the test data [14]. Some model selection strategies for OOD generalization have been suggested, such as selecting the model based on IRMv1 penalties over a validation set from each training domain, validation accuracy, the average ECE across training domains, and leave-one-domain-out cross-validation [7, 14]. However, Wald et al. [7] has already shown that the average ECE across training domains is better than IRMv1 penalties as a metric of model selection in the OOD generalization problem. [14] also has proven that model selection based on validation accuracy performed better than based on leave-one-domain-out cross-validation. Hence, in this paper, we mainly focus on the following strategies and compare them:

1. **Model selection based on validation accuracy:** We pool all data from training domains and divide it into training, validation, and test set. The model with the highest accuracy on the validation set will then be picked [14].

2. **Model selection based on the average ECE across training domains:** To balance In-Domain (ID) validation error with the average ECE across training domains as the measurable proxy for the robustness of a model to distribution shifts between domains, we select the model with the lowest average ECE from those that achieve ID validation accuracy above a user-defined threshold [7].

4 Methodology

In this section, we describe the methods to answer three research subquestions. The datasets we used in this research were synthetic, and the data generation process is presented in Section 5.1.

4.1 Evaluate the performance of the naive calibration

Since naive calibration is a post-processing tool, it is challenging to learn a model that can achieve good calibration in the multi-domain setting [7]. Therefore, it may not perform well in the OOD generalization problem. To evaluate the performance of naive calibration, we can compare the OOD accuracy of the binary classifier before and after naive calibration and calculate the difference in OOD accuracy before and after naive calibration. It is also worth looking at the difference in OOD AUROC before and after calibration. The difference here always means the performance metric of the model after naive calibration minus the same metric before naive calibration. So, if the difference is bigger than zero, it means an improvement.

To rule out flukes, we can create 200 synthetic datasets instead of only one dataset to train and calibrate the binary classifiers. The binary classifiers used in this research are logistic regression, linear support vector machine, decision tree, random forest, neural network, AdaBoost, and Naive Bayes. In this way, for each classifier, we can get the 200 values of difference in OOD accuracy before and after naive calibration, 200 values of difference in

OOD AUROC before and after naive calibration, the average difference in OOD accuracy before and after naive calibration across 200 datasets, and the average difference in OOD AUROC before and after naive calibration across 200 datasets. By checking the metrics mentioned above, we are able to determine if the naive calibration improves the model's OOD accuracy, OOD AUROC, and the extent of improvement. Hypothesis tests can be used to make the conclusion more convincing. We will use the bootstrapping hypothesis test due to the unknown asymptotic distribution of the given data points [18]. In such a manner, we can verify if OOD accuracy and OOD AUROC have a statistically significant improvement after naive calibration.

4.2 Relationship between OOD accuracy and the number of training domains

Wald et al. [7] additionally provided a mathematical proof that calibration can generalize across domains if the number of domains is roughly equal to the number of spurious features in the setting of linear-Gaussian models. Therefore, it is also worth looking into whether OOD accuracy improves after calibration as the number of training domains rises. To answer this question, we intend to determine whether a positive linear correlation exists between the number of training domains and the OOD accuracy. Pearson correlation coefficient (PCC) is a metric to measure the linear correlation between two sets of data. This metric can be used to measure the linear correlation between OOD accuracy and the number of training domains.

However, increasing the number of training domains means the training data also increases. Therefore, we want to test if there is a linear relationship between OOD accuracy and the number of training data from the same training domain. Thus, PCC can be used here to measure the linear correlation between OOD accuracy and the number of training data that are from the same training domain. Lastly, the partial correlation between OOD accuracy and the number of training domains can be obtained, with the effect of the amount of training data removed. In this way, we are able to determine if there is a positive linear relationship between OOD accuracy and the number of training domains without the influence of the amount of training data.

To rule out flukes, we can crate ten synthetic datasets instead of one dataset to train and calibrate the binary classifiers. The classifiers used here are mentioned in Section 4.1. In this way, for each classifier, we get PCC between the average OOD accuracy across ten datasets and the number of training domains, PCC between the average OOD accuracy across ten datasets and the number of training data that come from the same training domain, and the partial correlation between the average OOD accuracy and the number of training domains with the effect of the amount of training data removed. By checking the metrics mentioned above, we can decide if there is a positive correlation between OOD accuracy and the number of training domains.

4.3 Evaluate the model selection strategies for the OOD generalization problem

To determine if the model selection with average ECE is a reasonable model selection strategy, we want to see if a linear relationship exists between the average ECE across training domains and OOD accuracy, and how strong this relationship is.

We can train hundreds of neural networks on the synthetic dataset. For the fairness of comparing the models, the hyperparameters of each model should be the same. However, training multiple networks with the same hyperparameters can still output different models due to random initialization. Hence, each model may have different values of the average ECE across training domains, validation accuracy, and OOD accuracy.

Similar to section 4.2, PCC can be used to measure the linear correlation between the average ECE across training domains and OOD accuracy. However, we also want to test whether the average ECE across training domains is a more reasonable metric for selecting a model in the OOD generalization problem than validation accuracy. Therefore, we calculate the PCC between OOD accuracy and validation accuracy. By comparing these two PCC's, we can figure out which metric has a stronger linear relationship with OOD accuracy and thus conclude which is the better model selection strategy. Lastly, the partial correlation between OOD accuracy and average ECE is computed while adjusting for validation accuracy. By looking at PCC between the average ECE across training domains and OOD accuracy, and the partial correlation, we can determine if there is a linear correlation between the average ECE across training domains and OOD accuracy and how robust the relationship is, thereby determining if the model selection with average ECE is a reasonable model selection strategy.

5 Experiments and Results

This section explains how to create synthetic data, describes three experiments to answer the research question, and delivers the empirical results.

5.1 Data generation

The generation of the synthetic dataset used for the experiments is based on Figure 1. The set of environments E (also known as domains) is parameterized with real vectors expressing expectations and positive definite matrices expressing covariances:

$$E = \{(\mu, \sigma) | \mu \in \mathbb{R}^d, \sigma \in \mathbb{S}^d_{++}\}$$

Y is the label, which can be 0 or 1 for binary classification problems. X_{ac-ns} are the anti-causal-non-spurious features, also known as the invariant features. X_{ac-sp} are the anti-causal-spurious features; their distribution may change depending on the domain. Figure 1 means that all features are anti-causal, some are spurious while others are invariant, and the spurious features come from the environment.

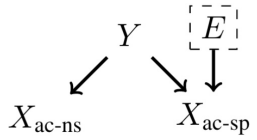


Figure 1: The causal diagram of the synthetic data [7]

There could be multiple domains in one synthetic dataset; for each domain i in the synthetic dataset, the features are $X_i = [X_{ac-ns}, X_{i-ac-sp}]$, the label is 0 or 1. For each

domain i, the distribution of X_{ac-ns} is the same, but the distribution of $X_{i-ac-sp}$ might be different. By changing the μ_i and σ_i in the equation (1), we can create X_{ac-sp} from different distributions for different domains. To simplify the data generation process, we use the `sklearn.datasets.make_classification` method to create the X_{ac-ns} for all domains. The data generating process for X_{ac-sp} features of domain i is given by:

$$X_{ac-sp}|Y = y \sim \mathcal{N}((y - 1/2)\mu_i, \sigma_i) \tag{1}$$

5.2 Experiment A: Performance of naive calibration

This experiment is mainly focused on evaluating the performance of naive calibration. Here are the details of the experiment.

First, we constructed a synthetic dataset containing four training domains and four unseen testing domains. Each domain has 6000 data points. Each data point is composed of four spurious features and four non-spurious features. The mean vector and covariance matrix of the spurious features in each domain were generated randomly. The details of the data generation process were described in Section 5.1. Then we trained seven binary classifiers mentioned in Section 4.1 on this dataset. After that, we computed OOD accuracy and OOD AUROC before calibration for each model. Next, naive calibration was applied to these models. We implemented the naive calibration using the `CalibratedClassifierCV` method in the sklearn package. After that, we computed OOD accuracy after calibration and calculated the difference in OOD accuracy before and after calibration for each model. The same calculation was applied to OOD AUROC. Then we obtained the difference in OOD AUROC before and after calibration.

By repeating the above process 200 times, We obtained 200 values of difference in OOD accuracy before and after naive calibration, 200 values of difference in OOD AUROC before and after naive calibration, and their corresponding averages.

Finally, two bootstrapping hypothesis tests were used here. The first hypothesis test determines whether the difference in OOD accuracy before and after calibration is statistically significantly different from 0. The null hypothesis is formulated as the difference in OOD accuracy before and after calibration is zero. The alternative hypothesis is formulated as the difference in OOD accuracy before and after calibration significantly deviates from zero. The second hypothesis test was constructed similarly, but its purpose is to determine if the difference in OOD AUROC before and after calibration is statistically significantly different from 0. The significance level here is 0.05, and the confidence level is 95%.

The results are summarized in Table 1 and Table 2. Logistic Regression, linear support vector machine, Neural Network, and AdaBoost are the four models with a positive Avg Diff OOD ACC and Avg Diff OOD AUROC and their corresponding p-values are smaller than 0.05. In addition, the lower values of the confidence intervals of their means are also bigger than zero in these two tables. From this, we can conclude that the null hypotheses for both bootstrapping hypothesis tests are rejected. This means there are statistically significant changes in OOD accuracy and OOD AUROC after naive calibration for these binary classifiers. More specifically, these models have a statistically significant improvement in OOD accuracy and OOD AUROC after naive calibration. For Decision Tree, Random Forest, and Naive Bayes, their p-values in two tables are bigger than 0.05. This means no statistically significant change in OOD accuracy and OOD AUROC after naive calibration.

	Avg Diff OOD ACC	P-value	Confidence interval of the mean
Logistic Regression	0.032	0.0	(0.024, 0.041)
Linear SVM	0.021	0.0	(0.014, 0.031)
Decision Tree	0.009	0.056	(-0.001, 0.021)
Random Forest	0.010	0.086	(-0.004, 0.023)
Neural Network	0.015	0.0	(0.011, 0.019)
AdaBoost	0.005	0.0008	(0.0033, 0.010)
Naive Bayes	0.001	0.371	(-0.004, 0.005)

Table 1: ACC results of Experiment A. Avg Diff OOD ACC: the average difference in OOD Accuracy of the model before and after calibration across 200 datasets. The models that have a statistically significant improvement in OOD accuracy are in **bold**.

	Avg Diff OOD AUROC	P-value	Confidence interval of the mean
Logistic Regression	0.038	0.0	(0.030, 0.050)
Linear SVM	0.024	0.0	(0.016, 0.033)
Decision Tree	0.010	0.070	(-0.003, 0.024)
Random Forest	0.007	0.208	(-0.010, 0.023)
Neural Network	0.019	0.0	(0.015, 0.025)
AdaBoost	0.003	0.0096	(0.0013, 0.007)
Naive Bayes	-0.0003	0.57	(-0.005, 0.004)

Table 2: AUROC results of Experiment A. Avg Diff OOD AUROC: the average difference in OOD AUROC of the model before and after calibration across 200 datasets. The models that have a statistically significant improvement in OOD AUROC are in **bold**.

5.3 Experiment B: Increase the number of training domains

This experiment tests if OOD accuracy improves as the number of training domains grows. Details of the experiment are listed below, and the results are summarized in Table 5.3.

1. A collection of ten synthetic datasets are generated. Each dataset contains fifteen different domains, ten training domains, and five unseen domains. Each domain has 5000 data points. Each data point is composed of four spurious features and three non-spurious features. The mean vector and covariance matrix of the spurious features in each domain were generated randomly. More details of the data generation process are described in Section 5.1.

2. At the beginning, for each dataset, the amount of current training data is zero, and all the training domains are unused.

3. For each dataset, add the data from one unused training domain to the current training data and mark this training domain as a used domain.

4. Then, for each dataset, train and calibrate the binary classifiers mentioned in Section 4.1 on the current training data. For each classifier, calculate OOD accuracies on these ten datasets and the average accuracy across these ten datasets.

5. At the same time, for each dataset, train and calibrate the same kinds of classifiers on the training data that follows the same distribution of the first training domain. For each dataset, the amount of the training data here matches the total amount of current training data in the previous step. We can generate these data based on the data generation process mentioned in Section 5.1. Similarly, for each classifier, calculate OOD accuracies on these ten datasets and the average accuracy across these ten datasets.

6. For each dataset, if all training domains are marked as used, then go to step 7; otherwise, go to step 3.

7. Then, for each model, we obtain the average OOD accuracy against the different number of training domains and the average OOD accuracy against the different number of training data.

8. Next, calculate the PCC between OOD accuracy and the number of training domains and the PCC between OOD accuracy and the number of training data.

9. Lastly, calculate the partial correlation between OOD accuracy and the number of training domains, with the effect of the amount of training data removed.

There is a positive linear correlation between the number of training domains and OOD accuracy for all binary classifiers. As shown in Table 5.3, PCC between the number of training domains and OOD ACC for other classifiers are equal to or more than 0.85, which can be considered highly correlated; except for the AdaBoost, whose PCC between the number of training and OOD ACC is only 0.37.

There is no significant pattern of PCC between OOD accuracy and the number of training data with the same distribution of one of the training domains. For the logistic regression and neural network, the values of PCC between the number of training data and OOD ACC are -0.94 and -0.88, which means adding more training data from the exact domains makes the classifiers worse for OOD generalization. However, adding more training data from the single domain seems to help improve OOD accuracy for linear SVM and random forest. This is also true for decision trees and Naive Bayes, but their values of PCC are smaller compared to those of linear SVMs and random forests. For the AdaBoost, there is no significant linear correlation between the number of training data and OOD accuracy (its PCC is only 0.04).

Looking at the partial correlation in Table 5.3, without the influence of the amount of the training data, there still is a positive correlation between the number of training domains and OOD accuracy for each classifier. Except for the AdaBoost, partial correlation values for other classifiers are high (>0.8). Given these findings, we conclude that adding training data from more training domains can improve the OOD accuracy of the binary classifiers. However, adding more training data from single domains can not. It should be noted that this experiment has its limitations and will be discussed in Section 7.

	PCC between the number of training domains and OOD ACC	PCC between the number of training data and OOD ACC	the Partial Correlation
Logistic Regression	0.85	-0.94	0.81
Linear SVM	0.88	0.49	0.85
Decision Tree	0.92	0.17	0.92
Random Forest	0.90	0.31	0.89
Neural Network	0.86	-0.88	0.81
AdaBoost	0.37	0.04	0.37
Naive Bayes	0.90	0.21	0.90

Table 3: Results of Experiment B. PCC between the number of training domains and OOD ACC: PCC between the number of training domains and OOD accuracy. PCC between the number of training data and OOD ACC: PCC between the number of training data from the same domain and OOD accuracy. The Partial Correlation: the partial correlation between OOD accuracy and the number of domains, with the effect of the number of training data removed.

5.4 Experiment C: Lowest average ECE as a selection strategy

This experiment determines whether a linear relationship exists between the average ECE across training domains and OOD accuracy. Three synthetic datasets were generated based on the data generation process mentioned in Section 5.1. Datasets A and B contain seven training domains and eight unseen domains. Dataset C has ten training domains and ten unseen domains. All datasets have 5000 points per domain. Each data point in Dataset A is composed of four spurious and four non-spurious features. However, each data point in Dataset B comprises five spurious and non-spurious features. Each data point in dataset C consists of seven spurious and seven non-spurious features. The mean vector and covariance matrix of the spurious features in each domain were generated randomly.

400 Neural Network models with identical hyperparameters were trained on each synthetic dataset. The input layer is followed by the batch normalization layer and the dropout layer for each neural network. The fraction of the input units to drop is 0.2. The dropout layer is followed by two hidden layers. The first hidden layer has two neurons. The second hidden layer has four or six neurons, according to the number of features of the dataset. The activation function of each hidden layer is ReLU. The output layer has one neuron, and the activation function is sigmoid. The optimizer is RMSprop, and the loss function is binary cross-entropy. Because the number of features in three datasets is small, we did not use many hidden layers and neurons to avoid overfitting. After training, three metrics mentioned in Section 4.3 for each dataset were calculated.

The results are summarized in Table 4 and visualized in Figure 2 to Figure 4. In datasets A, B, and C, the absolute value of PCC between the average ECE across training domains and OOD accuracy is greater than the absolute value of PCC between validation accuracy and OOD accuracy. This indicates a stronger linear relationship between OOD accuracy and average ECE than between OOD accuracy and validation accuracy. After removing the interference of validation accuracy, there is still a strong linear correlation between the average ECE and OOD accuracy (the partial correlations are -0.82, -0.56, and -0.71), as shown in Table 4. From the preceding arguments, we conclude that model selection based

on the average ECE across training domains is appropriate for the OOD generalization problem and is preferable to model selection based on validation accuracy. Additionally, the partial correlations reveal a negative linear correlation between average ECE and OOD accuracy, which means the lower the average ECE and the higher the OOD accuracy.

	PCC between ECE and OOD accuracy	PCC between validation accuracy and OOD accuracy	the Partial Correlation
Dataset A	-0.84	0.37	-0.82
Dataset B	-0.64	0.37	-0.56
Dataset C	-0.70	0.31	-0.71

Table 4: Results of Experiment C. PCC between average ECE and OOD accuracy: PCC between the average ECE across training domains and OOD accuracy. PCC between validation accuracy and OOD accuracy: PCC between the accuracy of the validation set pooled from all training domains and OOD accuracy. The Partial Correlation: the partial correlation between OOD accuracy and the average ECE across the training domains of the model while adjusting for the validation accuracy.

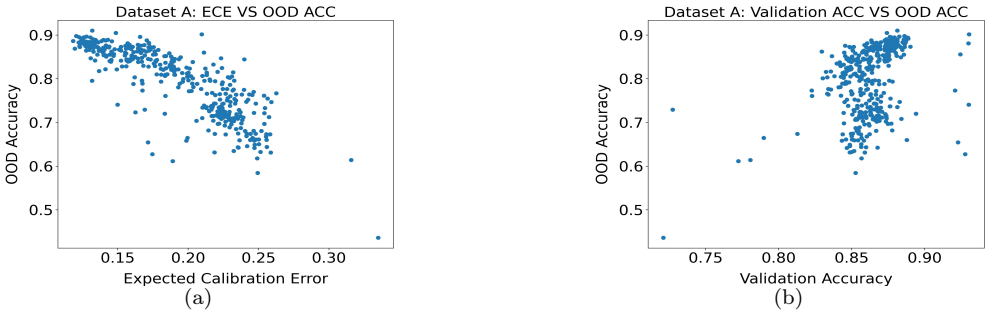
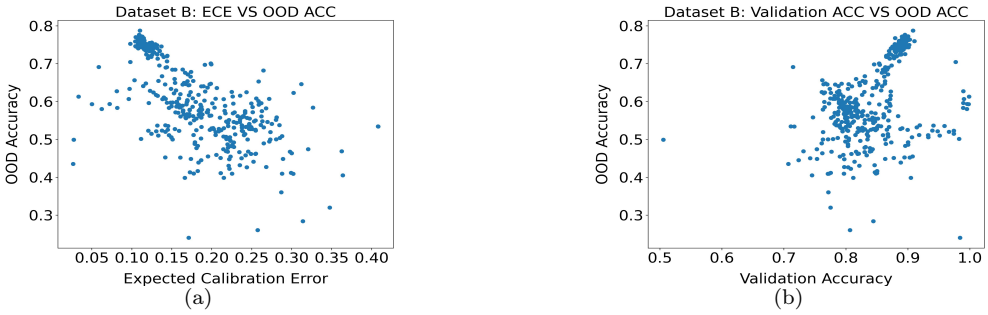
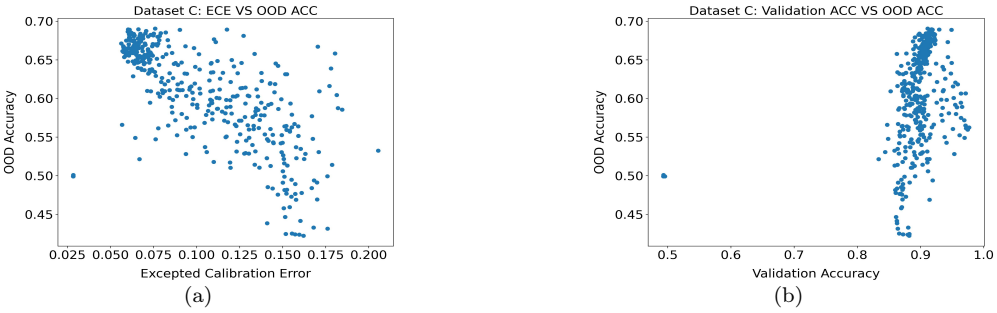


Figure 2: Dataset A. Each point corresponds to a model trained. (a) Correspondence between the average ECE across the training domains and OOD accuracy in dataset A. (b) Correspondence between the validation accuracy and OOD accuracy in dataset A.



(a) (b)

Figure 3: Dataset B. Each point corresponds to a model trained. (a) Correspondence between the average ECE across the training domains and OOD accuracy in dataset B. (b) Correspondence between the validation accuracy and OOD accuracy in dataset B.



(a) (b)

Figure 4: Dataset C. Each point corresponds to a model trained. (a) Correspondence between the average ECE across the training domains and OOD accuracy in dataset C. (b) Correspondence between the validation accuracy and OOD accuracy in dataset C.

6 Responsible Research

This section discusses the responsibilities associated with this research. Experiment reproducibility and the integrity of research are given special consideration.

The reproducibility of the experiments is considered in this study. The data generation process was explained in detail in Section 5.1. We described every step clearly for each experiment in Section 5. In Section 3, we also described the naive calibration and the model selection strategies in detail. The code we created is publicly available on GitHub [19]. It is anticipated that these remarks will assist other researchers in reproducing the experiments.

We also guarantee the integrity of the research. Firstly, all resources used in this research were referenced correctly. Secondly, all datasets and results generated by the experiments during this research can be found on GitHub [19]. As some implementations are based on the open-source code, we also cited the source in the Github README file.

As for the ethical part of the research, all datasets used in this research are synthetic, meaning there is no privacy problem associated with this research.

7 Discussion

The following empirical results are obtained based on the experiments in Section 5.

Although naive calibration can improve OOD accuracy and OOD AUROC for some binary classifiers, it does not improve OOD accuracy and OOD AUROC for all the binary classifiers mentioned in this study. However, at least it does not make the model worse for OOD generalization. For logistic regression, linear SVM, neural network, and AdaBoost, naive calibration can improve OOD accuracy and OOD AUROC. Specifically, OOD accuracy and OOD AUROC increase by at least 2% after naive calibration for logistic regression and linear SVM. After naive calibration, the average OOD accuracy of AdaBoost increases by only 0.5%, while the average OOD AUROC increases by only 0.3%, which can be regarded as minor. In addition, there are no statistically significant improvements in OOD accuracy and OOD AUROC after naive calibration for the decision tree, random forest, and Naive Bayes. As described in Section 3.1, the data used for the naive calibration is the validation data pooled from all training domains. In this case, naive calibration may learn a model calibrated on the pooled data, but uncalibrated on individual domains [7]. However, multi-domain calibration requires calibrating the model simultaneously across domains. This can explain why naive calibration may not learn a good-calibrated model across multiple domains and performs poorly sometimes [7]. Robust calibration and learning with Calibration Loss Over Environments (CLOvE) could be other choices to calibrate the model in the multi-domain setting. Both were proven to be better than naive calibration in four WILDS datasets [7].

Wald et al. [7] has proven that calibration can generalize across domains if the number of domains is roughly equal to the number of spurious features in the linear-Gaussian models and population (as opposed to finite-sample) setting. In Section 5.3, we concluded that training the binary classifiers on data from more training domains leads to higher OOD accuracy when the data is generated based on the process mentioned in 5.1. However, there is a limitation to this experiment. As shown in Section 5.3, the data from one more unused training domain will be added to the existing training data for each repetition. However, this chosen domain may be too excellent or terrible for OOD generalization. Therefore, experiment B might output an excellent or terrible OOD accuracy. For example, if the distribution of the data from this selected domain is very similar to the distribution of the data from the unseen domain, this domain can be considered too excellent. Suppose the data distribution from the chosen domain is very similar to that from one of the used domains. In that case, this domain can be regarded as too terrible, which might not provide extra helpful information to the model. So, it is better to try all the combinations of the training domains in each repetition and compute the average OOD accuracy to eliminate this contingency.

Section 5.4 indicates that the average ECE across training domains is a reasonable metric for selecting a model and is better than validation accuracy in the OOD generalization problem. In the Colored MNIST datasets, it displayed a strong linear correlation between OOD accuracy and average ECE (around -0.92) as well [7]. The logic behind this is that the average ECE across training domains can be seen as an indicator of how well the model is calibrated in the multi-domain setting [7]. Although model selection with high accuracy on validation set shows a good performance for the OOD generalization problem in some datasets, it does not mean it is the right strategy to select the model or tune hyperparameters [14]. Because this selection strategy is based on the assumption that the validation set and testing set come from the same distribution.

8 Conclusions and Future Work

This paper evaluated how well naive calibration and model selection with the average ECE perform in the OOD problem for binary classification. By testing naive calibration for seven different binary classifiers on synthetic datasets, we conclude that naive calibration improves average prediction performance for some binary classifiers, as measured in the accuracy or AUROC, across unseen domains. But it does not work for all binary classifiers mentioned in this study. However, naive calibration at least does not make the model perform worse in the OOD generalization problem. The following empirical results were obtained as well. First, for most binary classifiers used in this study, OOD accuracy will increase as the number of training domains increases when the data follows the setting described in Section 5.1. Second, the model selection strategy based on average ECE makes more sense than the one based on validation accuracy in the OOD generalization problem because the linear correlation between the average ECE across training domains and OOD accuracy is greater than the linear correlation between validation accuracy and OOD accuracy.

There are still several limitations to this research. First, all experiments are based on synthetic data. There could be a chance that the synthetic data generated in this research is much simpler than the real-world data. It might cause the performance of the naive calibration to be worse in real-world applications. Second, isotonic regression is the only method to implement naive calibration in this research. It might not be the best calibration method for some models in the multi-domain setting. Third, experiment B also has a limitation, as mentioned in Section 7. Fourth, the metrics used to measure the correlation between two variables are partial correlation and PCC. Both only measure linear relationships. However, there may be no linear relationship between the variables, but there is a non-linear relationship instead. Finally, the only way to calculate ECE is based on the reliability diagram. Hence, there is a chance that it does not provide a suitable calibration measurement.

Some improvements can be made to address the limitations in the future. First, these methods can be applied to the real-world dataset and evaluate their performance in the OOD generalization problem. In addition, other calibration methods can be tested and compared with the isotonic regression scaling, such as Bayesian Binning into Quantiles (BBQ). As a new nonparametric binary classifier calibration method, it has been shown that BBQ is often better at calibration than Platt's method, isotonic regression, and histogram binning on both simulated and real data [9]. Furthermore, experiment B can be improved as mentioned in Section 7. Lastly, different estimations to measure calibration error could be used and tested in model selection and naive calibration, such as Local Calibration Error and Density-based estimator [15].

References

[1] Qiao, F., & Peng, X. (2021). Uncertainty-guided Model Generalization to Unseen Domains. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6786–6796. https://doi.org/10.1109/CVPR46437.2021.00672

[2] Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B. A., Haque, I. S., Beery, S., Leskovec, J., Kundaje, A., … Liang, P. (2021). WILDS: A Benchmark of in-the-Wild Distribution Shifts. *arXiv:2012.07421 [cs]*.

[3] Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. *CVPR 2011*, 1521–1528. https://doi.org/10.1109/CVPR.2011.5995347

[4] Rosenfeld, E., Ravikumar, P., & Risteski, A. (2021). The Risks of Invariant Risk Minimization. *arXiv:2010.05761 [cs, stat]*.

[5] Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2020). Invariant Risk Minimization. *arXiv:1907.02893 [cs, stat]*.

[6] Kamath, P., Tangella, A., Sutherland, D. J., & Srebro, N. (n.d.). Does Invariant Risk Minimization Capture Invariance?, 10.

[7] Wald, Y., Feder, A., Greenfeld, D., & Shalit, U. (2022). On Calibration and Out-of-domain Generalization. *arXiv:2102.10395 [cs]*.

[8] Koyama, M., & Yamaguchi, S. (2021). When is invariance useful in an Out-of-Distribution Generalization problem ?

[9] Naeini, M. P., Cooper, G. F., & Hauskrecht, M. (n.d.). Obtaining Well Calibrated Probabilities Using Bayesian Binning, 7.

[10] Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. *Proceedings of the 22nd International Conference on Machine Learning - ICML '05*, 625–632. https://doi.org/10.1145/1102351.1102430

[11] Leathart, T., Frank, E., Holmes, G., & Pfahringer, B. (2018). Probability Calibration Trees.

[12] Zadrozny, B., & Elkan, C. (n.d.[a]). Transforming Classifier Scores into Accurate Multiclass Probability Estimates, 6.

[13] Zadrozny, B., & Elkan, C. (n.d.[b]). Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers, 9.

[14] Gulrajani, I., & Lopez-Paz, D. (2020). In Search of Lost Domain Generalization.

[15] Posocco, N., & Bonnefoy, A. (2021). Estimating Expected Calibration Errors. *arXiv:2109.03480 [cs]*.

[16] Bröcker, J. (2008). Some Remarks on the Reliability of Categorical Probability Forecasts. *Monthly Weather Review*, *136*(11), 4488–4502. https://doi.org/10.1175/2008MWR2329.1

[17] Murphy, A. H., & Winkler, R. L. (1977). Reliability of Subjective Probability Forecasts of Precipitation and Temperature. *Applied Statistics*, *26*(1), 41. https://doi.org/10.2307/2346866

[18] Martínez-Camblor, P., & Corral, N. (2012). A general bootstrap algorithm for hypothesis testing. *Journal of Statistical Planning and Inference*, *142*(2), 589–600. https://doi.org/10.1016/j.jspi.2011.09.003

[19] Liu, A. (2022). *Multi-domain calibration*. https://doi.org/10.5281/zenodo.6660575