Gaussian Latent Representations for Uncertainty Estimation using Mahalanobis Distance in Deep Classifiers

Aishwarya Venkataramanan, Assia Benbihi, Martin Laviale, Cedric Pradalier

Self-Supervised Gaussian Regularization of Deep Classifiers for Mahalanobis-Distance-Based Uncertainty Estimation

Aishwarya Venkataramanan[1,2,3], Assia Benbihi , Martin Laviale[1,3], and Cédric Pradalier[2,3]

[1]Université de Lorraine, CNRS, LIEC, France
[2]Georgia Tech Lorraine, GT-CNRS IRL2958, France
[3]LTSER- "Zone Atelier Moselle", France

Abstract

Recent works show that the data distribution in a network's latent space is useful for estimating classification uncertainty and detecting Out-Of-Distribution (OOD) samples. To obtain a well-regularized latent space that is conducive for uncertainty estimation, existing methods bring in significant changes to model architectures and training procedures. In this paper, we present a lightweight, fast, and high-performance regularization method for Mahalanobis distance (MD)-based uncertainty prediction, and that requires minimal changes to the network's architecture. To derive Gaussian latent representation favourable for MD calculation, we introduce a self-supervised representation learning method that separates in-class representations into multiple Gaussians. Classes with non-Gaussian representations are automatically identified and dynamically clustered into multiple new classes that are approximately Gaussian. Evaluation on standard OOD benchmarks shows that our method achieves state-of-the-art results on OOD detection with minimal inference time, and is very competitive on predictive probability calibration. Finally, we show the applicability of our method to a real-life computer vision use case on microorganism classification.

1 Introduction

Current deep learning classification networks achieve superior performance and find widespread applications in various industrial domains such as biology and robotics [1–3]. While they achieve state-of-the-art accuracy, there remain two main challenges that hinder the deployment of deep classifiers in critical situations: the derivation of calibrated classification and a measure of the classification uncertainty. Without those, a network exposed to Out-of-Distribution (OOD) data makes incorrect predictions with high confidence [4] and no human-in-the-loop can catch such errors. It is thus necessary to obtain calibrated probabilities [4] *i.e.,* predict probabilities that represent true likelihood, and to estimate the uncertainty in the network's predictions to allow users to make informed decisions.

Among deep uncertainty estimation approaches [5–8] are Bayesian Neural Networks [9], MC-Dropout [10] and Deep Ensemble [11]. These stochastic methods require multiple forward-passes so they are not scalable to large systems. Aware of the scalability requirements, current research focuses on estimating uncertainty from deterministic single-forward-pass networks [12–17]. Distance-based methods belong to this category and are an attractive alternative for their excellent performance in OOD detection [18, 19].

Distance-based methods rely on the distance between the test samples and the In-Distribution (ID) samples in a network's latent space to determine if the test samples are OOD. A relevant distance is the Mahalanobis distance (MD) [20] for its superior performance over Euclidean Distance (ED) [21–23]. One key MD assumption though is that the in-distribution samples in the latent space should follow class-conditional Gaussian distributions. In practice, though, there is nothing in the classification training that constrains the latent space to fulfil such an assumption [24]. Instead, research on representation learning shows that each

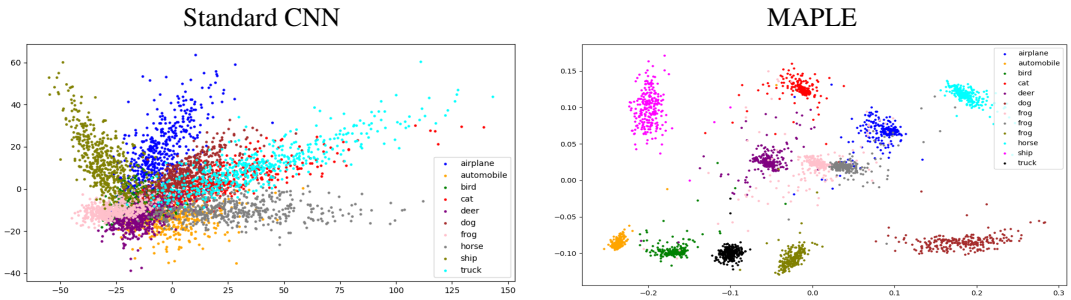

Figure 1: **Self-supervised latent space regularization with MAPLE** for uncertainty estimation and OOD detection. MAPLE improves class separation as illustrated by the PCA visualization of a CNN's latent space trained on CIFAR10 without regularization (left) and with MAPLE regularization (right). Our method constrains the latent representations to be approximately Gaussian to enable efficient distance-based uncertainty estimation.

class is usually composed of several clusters of visually similar images [25–27]. This can be due to intra-class variance of images taken from different view-points, the presence of additional objects in the image, and variations in object shapes. In the network's latent space, these variations appear as distinct distributions or deviate from a Gaussian distribution. This breaks the MD assumption, which could lead to incorrect or imprecise uncertainty estimation.

In this paper, we introduce MAPLE, a self-supervised representation learning method that regularizes a classification network's latent space to exhibit multivariate Gaussian distributions. MAPLE generates a latent space where class representations are Gaussian, making it compliant with the MD assumption and allows fast and high-performance MD-based OOD detection, uncertainty estimation, and calibrated classification. The effect of MAPLE is illustrated in Fig. 1 with the 2D projection of the latent space of a Convolutional Neural Network (CNN) trained on CIFAR10.

MAPLE stands for MAhalanobis distance based uncertainty Prediction for reLiablE classification, and is illustrated in Fig. 2. MAPLE relies on two components: i) a self-supervised intra-class label refinement through clustering in the latent space; ii) a deep metric learning loss that improves the class separation. During training, the representations associated to a class that deviate from a Gaussian distribution are divided into several clusters that are approximately Gaussian. The cluster assignments become the new labels of the representations, and the training goes on. Since each cluster gathers samples that exhibit similar intra-class variations, the clustering step is akin to automatic fine-grained annotation. The metric-learning then reinforces the fined-grained class separation by pushing apart the new classes. The combination of in-class clustering and metric learning results in classification representations that are well-clustered and approximately Gaussian, which makes them suitable for MD-based uncertainty estimation.

We evaluate MAPLE against existing uncertainty quantification methods on the three standard benchmarks: CIFAR10 [28] *vs.* SVHN [29]/CIFAR100 [28], FashionMNIST [30] *vs.* MNIST [31] for OOD detection and predictive probability calibration. Results show that MAPLE achieves the best compromise between performance and run time efficiency while being the most lightweight integration-wise. It achieves very competitive performance with the state-of-the-art and has the best inference time. Also, it introduces minor architectural changes and does not require additional fine-tuning to OOD datasets.

We summarize the paper's contributions as follows. **i)** We develop a self-supervised representation learning method that constrains a classification network's latent space to be approximately Gaussian. **ii)** We show that such representations allow for reliable OOD detection and probability calibration using MD. **iii)** We design the method such that it has a minimal impact on the network's original architecture, has low computational cost during inference, and achieves results competitive with the state-of-the-art on OOD detection.

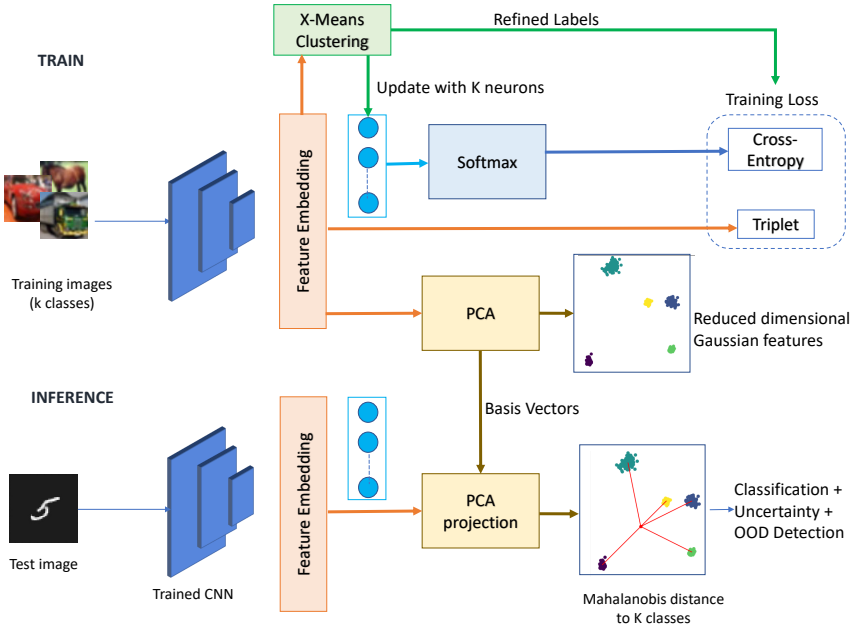

Figure 2: **Representation regularization with MAPLE for uncertainty estimation.** Our approach trains a classification network to learn representations that are approximately Gaussian for each class. During inference, the Mahalanobis distance between a test sample and the class centroids is used for classification, uncertainty estimation and OOD detection.

2 Related Work

Multi-forward-pass Uncertainty Estimation. Traditional uncertainty quantification methods rely on Bayesian Neural Networks [32, 33] to learn a distribution over the network weights. To extract predictive probability variance, sampling [34] or variational methods [9] are used. The application of these methods is limited, as they increase the number of parameters by a factor of two and hinder convergence. As a lighter alternative, MC Dropout [10] enables dropout at test time and averages the network's output over several forward passes. While MC Dropout paves the way towards faster and lighter uncertainty estimation, it has been shown to produce over-confident predictions [11] and underestimate uncertainty [35]. To improve uncertainty estimation, Deep Ensembles [11] average the predictions from an ensemble of trained models and achieve state-of-the-art performance on several classification tasks. It remains computationally expensive due to the training of multiple models and the several forward passes during inference. By deriving uncertainty from a single forward pass, MAPLE achieves significantly faster inference time without sacrificing performance.

Single-forward-pass Uncertainty Estimation. One line of work relies on the distribution of data samples in the network's latent space. A test sample is considered ID if it lies within the training data manifold, otherwise it is labelled as OOD. Methods differ in the way they regularize the representation space and the way they derive distances. DUQ [18] uses a Radial Basis Function (RBF) kernel in the representation space to measure distances between test samples and the centroids of various classes. Additionally, they use gradient penalty to obtain a regularized space, which improves the prediction's quality. SNGP [36] uses Spectral Normalization on the network's weights to satisfy the bi-Lipchitz condition, which is a more gradient-friendly regularization than DUQ. This condition preserves semantically meaningful distance changes in the representation space with respect to input changes. The prediction's uncertainty is then given by a Gaussian Process layer on the output. To improve the scalability of the Gaussian Process estimation, [14] proposes Deep Kernel Learning to process the input images with a distance-preserving network and fit a Gaussian on inducing

points only. Contrary to these methods, MAPLE avoids the Gaussian Process estimation and gradient regularization during training and instead relies on simple metric learning. Similarly, VMDLS [24] simplifies the Gaussian enforcement by training the network with a KL-divergence loss so that each class representations follow an isotropic Gaussian distribution in the latent space. However, this ignores the possible intra-class variation within each class and requires the Gaussian variance to be tuned manually. Instead, MAPLE uses a simpler self-supervised clustering that automatically fits the data. Also, MAPLE makes the latent space not only suitable for OOD detection but also for calibrated probability prediction.

Mahalanobis-Distance for OOD detection. MD is a common distance in the OOD detection literature. Early work by Lee et al. [19] derives confidence values as a function of MD to predict the likelihood of a sample being ID. To obtain competitive performance, the method requires several tweaks such as adding noise to input samples, combining confidence values from multiple feature layers, and fine-tuning on OOD datasets. [23] proposes two light improvements: Partial MD and Marginal MD. In Partial MD, the MD is computed on lower dimensional representations with PCA. Marginal MD uses all training representations to fit a single Gaussian to calculate the MD. While both perform well on Far-OOD datasets *i.e.,* where ID and OOD samples are significantly distinct, their results are limited on Near-OOD [37], where the OOD samples are semantically similar to the ID ones. Relative MD (RMD) [22] improves the MD performance on Near-OOD by computing a global MD between the test sample and the samples of all classes combined, and then subtracting this value from the per-class MDs. All these methods exhibit satisfying performance, but their main limitation is their strong assumption that the image representations follow a Gaussian distribution, even though standard classification training does not enforce such a constraint. MAPLE addresses this limitation with a self-supervised regularization. By doing so, the features better fit the theoretical framework of MD-based OOD detection, thereby improving the performance.

3 Method

In this section, we describe MAPLE, a self-supervised regularization method for MD-based OOD detection, uncertainty estimation, and calibrated classification. It augments a standard CNN classifier with a self-supervised regularization to output both class probabilities and MD-based uncertainty. To enable MD for OOD detection, the representations of the training samples are dynamically clustered into multiple Gaussians using X-Means [38] during training. The samples are assigned new pseudo-class labels defined by their cluster assignment. The network is then optimized with the cross-entropy loss and the triplet loss. With periodic validation, the clusters are updated and the total number of classes change with every validation. At inference time, the MD between a test sample and each cluster's centroid is used to estimate the classification uncertainty and the probability of the point being OOD. Note that the only modification to the original network architecture is in the final layer, where the number of output neurons change according to the number of clusters identified. This makes MAPLE easy to integrate to any classification network. An algorithmic description is provided in Appendix C.

Self-Supervised Dynamic Relabelling. During training, MAPLE updates the training labels to make them representative of the features' separation in the latent space. Every p epochs, the network is evaluated on $\boldsymbol{\mathcal{D}_{val}}$ and the classes with a false negative ratio higher than a threshold t are updated. This is representative of the scenarios where the samples of a given class are misclassified, which is typical of classes with high intra-class variations. For every class to update, the training representations belonging to such a class are extracted and clustered using X-Means [38]. The resulting clusters form well-separated groups and we use the cluster assigmnent as new pseudo-labels for the train samples. If k' additional clusters are introduced by X-Means, each of them are considered as independent classes. Thus, the number of classes becomes $K = k + k'$, and the final layer of the model is updated to have K neurons. Then, the network training continues with the new labels. During inference, the pseudo labels are remapped to the original set of k labels to identify their original class.

Fig. 3 illustrates the benefits of jointly using X-Means and the triplet loss on the representations: X-Means splits classes with high intra-class variations into separated classes that are semantically more representative

Figure 3: **Visualizing intra-class label refinement and feature optimization.** The original data is not perfectly Gaussian due to intra-class variations. X-Means refines the labelling by dividing the samples into multiple clusters that are approximately Gaussian. The clusters are considered as separate classes during training. Triplet loss optimizes the representations by bringing the in-class samples together and separating them from other classes.

of the data, and the triplet loss reinforces this separation.

The method introduces three hyperparameters: false negative ratio threshold t, frequency of validation epochs p and the maximum number of clusters (`max_num_cluster`), which is a parameter needed for X-Means. More details on the hyperparameters are provided in Appendix. B.4.

Clustering. The motivation for using X-Means over other commonly used clustering methods such as K-Means [39], DB-SCAN [40] and Gaussian Mixture Models (GMMs) are two-folds: (1) X-Means is scalable and automatically identifies the number of clusters based on the Bayesian Information Criterion (BIC); (2) BIC uses a maximum likelihood estimation of the variance under the spherical Gaussian assumption, which means that the samples are approximately spherical Gaussian in each cluster.

3.1 Representation Distance

This section describes the MD derivation over the latent representations. To avoid matrix singularities, the latent representations are first reduced using PCA.

Dimensionality reduction. Representations extracted from large neural networks usually have a high dimension and redundant dimensions. The MD requires calculating the inverse covariance matrix of these features, but the presence of redundancy causes the covariance matrix to be singular. Furthermore, [22] shows that the presence of non-informative dimensions could be detrimental to MD performance. This motivates the use of dimensionality reduction.

A common dimensionality reduction method is t-SNE [41], widely used for latent space's visualization. While t-SNE maintains the local distribution of points, it fails to represent global distributions accurately, which is undesirable in distance-based uncertainty predictions. Instead, we use Principal Component Analysis (PCA) for dimensionality reduction. The principal components are constructed from the covariance matrix of the standardized training representations. The eigen vectors of the covariance matrix are the principal components and the eigen values account for the amount of original information (variance) present in these components. We automatically estimate the number of principal components by the number of eigen values in decreasing order, required to explain 95% of the original data variance. This transformation is denoted by $g : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$, where d' is the dimension of the reduced features. With $\boldsymbol{x'_{train}} = f^{\theta}(\boldsymbol{x_{train}})$ the full dimensions training features, we denote $\boldsymbol{z_{train}} = g(\boldsymbol{x'_{train}})$ the reduced features.

Mahalanobis Distance. The MD is a generalized version of Euclidean distance that takes into account the data correlation to measure the distance. Hence, the MD is more accurate when predicting the distance between a point and a distribution of points. Here, MD is calculated on the PCA-reduced representations as follows. Let $\{z_i\}$ be the set of training representations after dimensionality reduction, μ_c be the class centroids with $c = 1, 2, ..., K$, and $\boldsymbol{\Sigma}$ be the shared covariance for all training samples, given by

$$\mu_c = \frac{1}{N_c} \sum_{i:y_i=c} z_i$$
$$\mathbf{\Sigma} = \frac{1}{N} \sum_{c} \sum_{i:y_i=K} (z_i - \mu_c)(z_i - \mu_c)^T \tag{1}$$

The following Eq. 2 gives the Mahalanobis distance between the centroid μ_c of class c and a test sample $\tilde{x}$ with reduced representation $\tilde{z} = g(f^\theta(\tilde{x}))$

$$MD_c(\tilde{x}) = \sqrt{(\tilde{z} - \mu_c)^T \mathbf{\Sigma}^{-1} (\tilde{z} - \mu_c)} \tag{2}$$

3.2 Classification and Uncertainty Estimation

We now show how to use the MD distance calculated in Eq. 2 for three purposes: classification, predictive probability, and uncertainty prediction.

MD-based Classification. The predicted class is the one whose centroid c^* is closest to the test sample $\tilde{x}$:

$$c^* = \underset{c}{\text{argmin}}(MD_c(\tilde{x})) \tag{3}$$

Note that this classification is inferred in addition to the usual classification done by the network by taking the maximum of the output logits.

Predictive Probability. We convert the MD into a calibrated classification probability using the following property: the squared MD on representations with dimension d' follows a chi-squared distribution $\chi^2_{d'}$ with d' degrees of freedom. The MD is converted as follows:

$$P^c_{MD} = 1 - \text{cdf}(\chi^2_{d'})(MD_c(\tilde{x})^2) \tag{4}$$

where cdf(.) is the cumulative distribution function. P^c_{MD} represents the probability that a test sample belongs to class c. When the test point belongs to a particular class, the MD to that class is low and the corresponding P^c_{MD} is high. The predictive probability is the one associated with the class c^* obtained in Eq. 3:

$$P^{c^*}_{MD} = \max_c (P^c_{MD}) \tag{5}$$

Note that contrary to a CNN softmax 'probabilities', this classification probability is calibrated and can be interpreted as a confidence in the classification output. This means P^c_{MD} represents the actual probability that a sample belongs the class c.

Uncertainty Prediction. We define the predictive uncertainty, which is the uncertainty in the network prediction as

$$u_{c^*} = 1 - P^{c^*}_{MD} \tag{6}$$

For small values of MD, u_{c^*} is around 0 and goes to 1 as the MD increases.

4 Experiments

We compare MAPLE with the following related works: two multi forward-pass methods MC-Dropout [10] (10 dropout samples) and Deep ensemble [11] (10 models), four single forward-pass methods: DUQ [18], SNGP [36], DUE [14] and VMDLS [24]. Following the standard evaluation on OOD detection, we evaluate the methods on classification, predictive probability calibration, and OOD detection on the three benchmark datasets: FashionMNIST [30] *vs.* MNIST [31], CIFAR10 [28] *vs.* SVHN [29], CIFAR10 *vs.* CIFAR100 [28].

We also compare MAPLE with MD-based methods on OOD detection, namely, the approach by Lee et al. [19], Marginal MD [23] and RMD [22]. We used the near-OOD CIFAR10 *vs.* CIFAR100 for the comparison, which is notably challenging for OOD detection.

4.1 Evaluation Metrics

We report the standard evaluation metrics [18, 36] namely, the classification accuracy, the Expected Calibration Error (ECE), the Negative Log-Likelihood (NLL), the Area Under the Receiver Operating Characteristics (AUROC) and the Area Under the Precision-Recall curve (AUPR). For qualitative analysis, we use calibration plots and uncertainty histograms (Appendix. D.1). As mentioned previously, MAPLE produces two classification outputs so we report the accuracies obtained from both the traditional softmax probability and the MD-based classification (Sec. 3.2). The ECE and the NLL are calculated from the predictive probability $P^{c^*}_{MD}$. AUROC and AUPR are calculated from the uncertainty u_{c^*}. The definition of these standard metrics are recalled in Appendix A.

4.2 Implementation Details

As in [18], the network architecture used for training FashionMNIST is a three layer CNN. The CIFAR10 training follows [14, 36] and uses a Wide ResNet 28-10 [42] for the classification backbone. The hyperparameters for the trainings are $p = 10, t = 0.3$ and `max_num_cluster`$= 5$. Additional details on the network architecture, dataset splits, hyperparameter search, and the hardware used for training are provided in the Appendices B.1, B.2 and B.4.

4.3 Results

We report the results on FashionMNIST and CIFAR10 in Table (Tab.) 1 and 2 respectively.

Method	ID metrics			OOD metrics		Latency↓
	Accuracy ↑	ECE ↓	NLL ↓	AUROC ↑	AUPR ↑	(ms/sample)
MC Dropout [10]	0.923	0.069	**0.213**	0.912	0.895	15.46
Deep ensemble [11]	**0.939**	0.018	0.238	0.874	0.866	23.87
DUQ [18]	0.923	0.045	0.276	0.941	0.945	2.61
SNGP [36]	0.924	**0.009**	0.259	0.981	0.978	2.54
DUE [14]	0.923	0.028	0.284	0.954	0.948	2.57
VMDLS [24]	0.920	-	-	0.963	0.970	2.60
MAPLE	0.925/0.924	0.020	0.262	**0.995**	**0.994**	**2.48**

Table 1: **FashionMNIST (ID) vs MNIST (OOD).** MAPLE achieves the best performance on OOD detection and has the best inference time. It is very competitive with other single-pass methods on the classification task. Blue: Classification based on prediction from softmax probability Orange: MD-based classification.

OOD Detection Results. MAPLE outperforms all baseline methods by upto 12% on the AUROC and AUPR scores, and achieves so with the least computation time[1]. Note that competitive approaches, such as SNGP and DUE, derive their performance from spectral normalization and Gaussian process layer, which are invasive training add-ons. In contrast, MAPLE relies only on the layers of a standard CNN architecture to achieve superior performance.

When it comes to inference speed, MC Dropout and Deep Ensemble perform the worst, which is expected since they require multiple forward passes during inference. In contrast, most single-forward-pass methods achieve scores comparable to MC Dropout and Deep Ensemble while being faster, with a factor close to 8 times faster when comparing MAPLE and Deep Ensemble. This reinforces MAPLE's motivation: the distribution of feature points in a network's latent space holds reliable information for fast prediction of a network's uncertainty and detection of OOD samples.

[1]Latency value for MAPLE includes time for inference+post-processing with MD. Latency for MC Dropout and Deep Ensemble are when the inferences are performed serially.

Method	ID metrics			OOD AUROC ↑		OOD AUPR ↑		Latency↓
	Accuracy ↑	ECE ↓	NLL ↓	SVHN	CIFAR100	SVHN	CIFAR100	(ms/sample)
MC Dropout [10]	0.960	0.048	0.293	0.932	0.835	0.965	0.829	27.10
Deep Ensemble [11]	**0.964**	0.014	**0.134**	0.934	0.864	0.935	0.885	38.10
DUQ [18]	0.945	0.023	0.222	0.927	0.872	0.973	0.833	8.68
SNGP [36]	0.957	0.016	0.153	0.991	0.911	0.994	0.907	6.25
DUE [14]	0.956	0.015	0.179	0.936	0.852	0.967	0.850	6.94
VMDLS [24]	0.951	-	-	0.932	0.868	0.953	0.864	5.61
MAPLE	0.956/0.954	**0.012**	0.142	**0.996**	**0.926**	**0.997**	**0.918**	**4.96**

Table 2: **CIFAR10 (ID) vs SVHN / CIFAR100 (OOD).** MAPLE outperforms all single and multi pass methods on OOD detection, and results in significantly faster derivation. Classification with MAPLE is very competitive with the state-of-the-art and the predicted probabilities are better calibrated. Blue: classification based on prediction from softmax probability. Orange: MD-based classification.

Classification Results. MAPLE achieves results competitive to state-of-the-art, only 1% below the top method Deep ensemble [11] whose score comes at the cost of training and inference on several models. Note that both MAPLE accuracies, the softmax probability and the MD-based one are close. A finer analysis of the accuracy shows that the slight difference in accuracy with the MD-based classification occurs on samples the network is uncertain about: MAPLE achieves top accuracy on high-confidence predictions (above 80% and 90% confidence) and the accuracy slightly decreases for lower-confidence predictions. See Appendix. D.2 for an extended analysis.

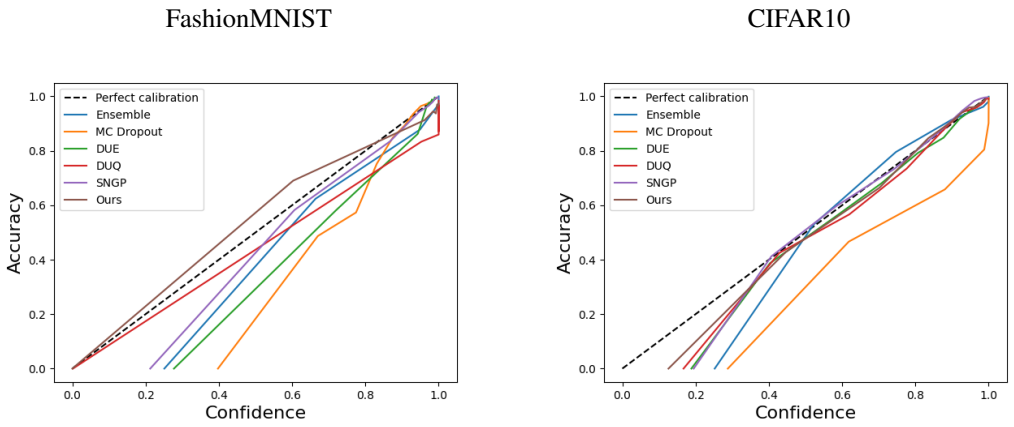

Figure 4: **Calibration plots.** A perfectly calibrated plot is when the predicted confidence equals the true likelihood *i.e.,* the accuracy. This is shown by the linear dotted line in the plots. MAPLE is closer to optimal calibration than existing methods, especially for low-accuracy samples.

Calibration Results. MAPLE is competitive with state-of-the-art SNGP [36] and Deep Ensembles. When training on FashionMNIST, one source of ECE error is MAPLE's under-confidence on the accuracy range below 80%. This is visible in the calibration plot (Fig. 4) where the curve goes above the ideal calibration: the confidence is lower than the accuracy. This is typical of the scenario where the inter-class representations are widely spread out. Even though a sample falls closest to its ground-truth centroid, their inter-distance remains high, which decreases the confidence. The sample is then correctly classified, but with a low confidence. Note that while optimal calibration is the gold-standard, MAPLE's under-confidence still makes it more compliant with hazardous applications than other methods that make over-confident predictions, which can be disastrous. On CIFAR10, all methods are well-calibrated, except for the overconfident MC-Dropout, which explains its high ECE score. When the accuracy is below 0.4, baseline methods become

overconfident whereas MAPLE is closer to optimal calibration and achieves the best ECE score.

4.4 Comparison with other MD methods

Setup. MAPLE is compared against MD-based OOD detectors [19, 22, 23]. These methods are tailored for OOD detection, so we report the metric relevant to this task only for the sake of fairness. We report the AUROC score on the challenging near-OOD dataset CIFAR10 *vs.* CIFAR100. The experiments are done with a Wide ResNet 28-10 [42].

Method	AUROC ↑
Lee et al. [19]	0.893
Marginal MD [23]	0.838
RMD [22]	0.897
MAPLE	**0.926**

Table 3: **Comparison with MD-based OOD detection.** MAPLE performs significantly better in OOD detection than existing MD-based methods on the CIFAR10 *vs.* CIFAR100 setup. By enforcing the learned representations to follow a Gaussian distribution, MAPLE allows for distance derivations that are more semantically meaningful.

Method	ID metrics			OOD metrics - SVHN		OOD metrics—CIFAR100		#Eig
	Softmax Accuracy ↑	MD-based Accuracy↑	ECE ↓	AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑	
DNN+MD (1)	0.950	0.943	0.086	0.752	0.762	0.583	0.564	-
DNN+PCA+MD (2)	0.950	0.946	0.053	0.855	0.839	0.813	0.859	12
DNN+PCA+ED (3)	0.950	0.943	0.105	0.829	0.804	0.734	0.765	12
DNN+Triplet+PCA+MD (4)	0.954	0.953	0.013	0.945	0.948	0.912	0.894	11
DNN+Clustering+PCA+MD (5)	0.947	0.945	0.032	0.922	0.908	0.811	0.815	12
MAPLE (6)	**0.956**	**0.954**	**0.012**	**0.996**	**0.997**	**0.926**	**0.930**	12

Table 4: **Ablation study.** We evaluate the influence of several MAPLE components. **PCA** (1 vs 2) results in a significant improvement of the OOD detection by discarding non-informative dimensions. The distances derived on these reduced features are better representative of the similarity between the input samples. The **MD** (2 vs 3) is better suited than ED for calibrated classification and OOD detection, which reiterates conclusions already found in previous works. The **triplet loss** (2 vs 4) improves both the accuracy and the OOD metrics by increasing the class separation. **Clustering** alone (2 vs 5) also contributes to a better separation of the classes, but the results are not as significant. The joint use of **triplet loss and clustering**, as done in MAPLE (6) achieves the best results on both classification and OOD detection. Note: #Eig refers to the number of principal components, whenever applicable.

OOD Detection Results. MAPLE achieves top-performance on Near-OOD detection (Tab. 3), which supports MAPLE's representation regularization. Note that the primary difference between MAPLE and the baselines is their lack of constraints on the latent representation. In contrast, we force the samples of every class to be Gaussian before calculating MD. Non-Gaussian samples lead to incorrect mean and covariance calculations, resulting in incorrect distance values. The error is more pronounced when the samples deviate from the Gaussian distribution by a large factor. This explains why the MD-based approaches under-perform compared to MAPLE on Near-OOD.

4.5 Ablation analysis

In this study, we assess how the different components of MAPLE impact its performance. We train a wide ResNet 28-10 [42] network on CIFAR10 and use SVHN and CIFAR100 as OOD datasets.

Dimensionality Reduction. We consider two scenarios: **(1) DNN+MD** - A baseline where a standard Deep Neural Network (DNN) is trained with the cross-entropy loss and with no feature regularization. The MD is computed on the raw features, and we add a value of $1e^{-20}$ to the diagonal elements [22] to avoid a singular covariance matrix. **(2) DNN+PCA+MD** - It follows (1) except that the MD is derived on PCA-reduced features.

Results: Dimensionality reduction (2) drastically improves the network's performance, as shown in the first line of Tab. 4. The improvement amounts to 7-30% on the OOD metrics and 3% on the ID metrics. One possible explanation is that the reduced dimensions are the ones that contribute to distinguishing ID samples from OOD ones, as previously observed by [22]. When including all the feature dimensions in the MD, the dimensions that do not contribute to discriminating ID and OOD samples add up and dominate the final MD score.

Distance Definition. We compare Mahalanobis distance and Euclidean distance (ED) in the network's latent space. We compare **(2) DNN+PCA+MD** with the new experiment **(3) DNN+PCA+ED** - It follows (2) except that the MD is replaced with ED. As for MD, the $\chi^2_{d'}$ distribution is used to obtain the probability values from ED (Sec 3.2).

Results: The results show that MD boosts the performance in terms of ID and OOD metrics. The improvement is ECE score is by 5%, and the OOD metrics improved by 3-9% when using MD. This is because MD takes into account the data correlation, which gives a better estimate of the probability and uncertainty values.

Representation training. To study the influence of the training on the representations, we consider three experiments: **(4) DNN+Triplet+PCA+MD** - We train the DNN using both cross-entropy and triplet loss. **(5) DNN+Clustering+PCA+MD** - We train using the cross-entropy loss only and periodically cluster the feature points using X-Means. **(6) MAPLE** - This is our proposed method that fuses (4) and (5). For all experiments, the MD is derived on the reduced features.

Results: Using the triplet loss (4) improves the performance considerably compared to training with the cross-entropy loss only (2). An explanation is that the triplet loss pulls in-class feature embeddings together, and pushes the other class features apart. This encourages the representations to be well separated and makes it easier to distinguish OOD features. Choosing the triplet loss for metric learning is empirically motivated: experiments using contrastive loss showed that triplet loss has a slightly better performance.

Periodic clustering (5) improves the ECE score by 2%, and the AUROC and AUPR scores on SVHN by about 7% compared to (2). However, there is a slight drop in accuracy by 0.3% and OOD metric by 4% on CIFAR100. One explanation is that clustering increases the chances of new classes to overlap. This phenomenon is illustrated in the centre plot of Fig. 3. The class overlap is particularly hindering when the new domain is close to the training one: with clustering (5), the SHVN scores are better but the near-OOD CIFAR100 performs better without clustering (2).

MAPLE uses clustering together with triplet loss and achieves top-performance. The triplet loss reduces the overlap introduced with the clustering by pulling apart the newly created classes. With MAPLE, the latent representations are approximately Gaussian and well-clustered resulting in better MD estimates and superior performance in both ID and OOD metrics. Compared to experiment (2), the calibration error drops by 4%and the OOD scores improved by 4-11%.

False Negative Ratio t. We evaluate the influence of the clustering trigger *i.e.,* the False Negatives Ratio. We train MAPLE with a range of t values on CIFAR10 (Tab. 5).

Results: A low value of t results in overclustering, where multiple clusters contain similar images. This further increases the chances of misclassifications, leading to decrease in the metric values. On the other hand, high t values result in underclustering. Note that for $t > 0.3$, there are no additional clusters generated. This is because, the classes have false negative ratios that are below this threshold and so, they are not clustered. For CIFAR10, a t value of 0.3 yields the best results. An extended ablation analysis on the influence of classification backbones, clustering methods, and hyperparameters is provided in Appendix. E.

False Negative Ratio (t)	#Classes	Accuracy↑	ECE↓	SVHN AUROC↑	CIFAR100 AUROC↑
0.0	23	0.9449	0.014	0.922	0.888
0.1	18	0.9534	0.013	0.964	0.918
0.2	14	**0.9544**	0.012	0.991	0.925
0.3	12	0.9541	**0.012**	**0.996**	**0.926**
0.4	10	0.9535	0.013	0.961	0.921
0.5	10	0.9535	0.012	0.955	0.915

Table 5: **Metrics for different values of False Negative Ratio evaluated on CIFAR10** #Classes refers to the total number of output classes obtained after clustering. A low value of t results in overclustering, whereas a high t fails to detect classes with high variance.

5 Discussion

With the periodical clustering and the dynamic re-labeling, a natural question that arises is *'Is there a drop in performance when the ground truth labels change during training?'*. Experimentally, we observe a drop in training accuracy by 2-3% in the following epoch after every clustering phase. However, the network makes up for the drop within 4-5 epochs of training.

It can happen that the clusters contain very few samples, which introduces label imbalance when classifying. This is exacerbated when the samples are over-clustered. To mitigate this, we restrict X-Means to only cluster the classes that get misclassified. These are the classes with a false negative ratio higher than the threshold t. Automatic clustering regularization [43–45] is left for future work.

6 Use Case: Microorganism Classification

We consider the real-life computer vision use-case of image-based diatom identification [46]. Diatoms are microorganisms present in the water. The distribution of diatoms in the water is a useful indicator for predicting the water quality. Diatoms consist of several species or 'taxa', each corresponding to a different class with a different appearance. Typical in several biology applications, the image dataset includes a lot of intra-class variance (Fig. 5). In this study, we evaluate the performance of different approaches when encountering taxa that were not previously trained on.

Method	Accuracy ↑	ECE ↓	AUROC ↑	AUPR ↑	Latency (ms/sample)↓
MC-Dropout [10]	0.936	0.039	0.548	0.589	129.7
Deep Ensemble [11]	**0.969**	**0.025**	0.589	0.570	146.81
SNGP [36]	0.954	0.196	0.798	0.826	26.25
MAPLE	0.963	0.036	**0.864**	**0.865**	**17.38**

Table 6: **Real Case Application: microorganism classification.** With its top performance and state-of-the-art speed, MAPLE makes for a particularly applicable method for classification and OOD detection on real case datasets.

We train a Wide ResNet 28-10 on 130 taxa and use 36 taxa as OOD. The dataset is particularly challenging since it is fine-grained and Near-OOD. Additional details on the dataset and experimental setup are provided

in Appendix B.3. As shown in Tab. 6, MAPLE outperforms all baselines on OOD detection. While Deep Ensemble has a slightly better classification accuracy and ECE score, MAPLE significantly outperforms it in OOD with a 30% score boost and a runtime 8 times faster.

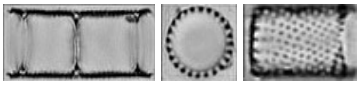

Figure 5: **Micro-organisms belonging to the same class.** These images of **one** diatom class show wide appearance changes due to different viewpoints during the acquisition. These translate into separate distributions in the latent space, deviating from Gaussian distribution. MAPLE's regularization makes the latent space Gaussian, hence suitable for MD calculation.

7 Conclusion

This paper presents MAPLE, a self-supervised regularization method for uncertainty estimation and out-of-distribution detection on CNN classifiers. The uncertainty is derived from the Mahalanobis Distance (MD) between an image representation and the class representations in the network's latent space. MAPLE derives meaningful MD distances by introducing a regularizer based on self-supervised label refinement and metric learning. Thus, MAPLE learns well-clustered representations that are approximately Gaussian for each class, which complies with the theoretical requirements of MD-based uncertainty estimation. Experimental results show that MAPLE achieves state-of-the-art results on out-of-distribution detection with the shortest inference time, and is very competitive with existing methods on predictive probability calibration. MAPLE also has the significant advantage of introducing the least architectural changes. Finally, we demonstrate a real-life use-case of our method on microorganism classification for the automatic assessment of water quality in natural ecosystems.

References

[1] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017. 1

[2] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," *Journal of Field Robotics*, vol. 37, no. 3, pp. 362–386, 2020. 1

[3] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. Nelson, A. Bridgland, *et al.*, "Improved protein structure prediction using potentials from deep learning," *Nature*, vol. 577, no. 7792, pp. 706–710, 2020. 1

[4] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*, pp. 1321–1330, PMLR, 2017. 1

[5] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods," *Machine Learning*, vol. 110, no. 3, pp. 457–506, 2021. 1

[6] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, *et al.*, "A survey of uncertainty in deep neural networks," *arXiv preprint arXiv:2107.03342*, 2021. 1

[7] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion*, vol. 76, pp. 243–297, 2021. 1

[8] Y. Gal *et al.*, "Uncertainty in deep learning," 2016. 1

[9] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *International conference on machine learning*, pp. 1613–1622, PMLR, 2015. 1, 3

[10] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, pp. 1050–1059, PMLR, 2016. 1, 3, 6, 7, 8, 11, 18, 20, 21

[11] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in neural information processing systems*, vol. 30, 2017. 1, 3, 6, 7, 8, 11, 18, 20, 21

[12] J. Postels, H. Blum, C. Cadena, R. Siegwart, L. Van Gool, and F. Tombari, "Quantifying aleatoric and epistemic uncertainty using density estimation in latent space," *arXiv preprint arXiv:2012.03082*, 2020. 1

[13] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21464–21475, 2020. 1

[14] J. van Amersfoort, L. Smith, A. Jesson, O. Key, and Y. Gal, "On feature collapse and deep kernel learning for single forward pass uncertainty," *arXiv preprint arXiv:2102.11409*, 2021. 1, 3, 6, 7, 8, 18, 20, 21

[15] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," *Advances in neural information processing systems*, vol. 31, 2018. 1

[16] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136*, 2016. 1

[17] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," *Advances in neural information processing systems*, vol. 31, 2018. 1

[18] J. Van Amersfoort, L. Smith, Y. W. Teh, and Y. Gal, "Uncertainty estimation using a single deep deterministic neural network," in *International conference on machine learning*, pp. 9690–9700, PMLR, 2020. 1, 3, 6, 7, 8, 16, 18, 20, 21

[19] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," *Advances in neural information processing systems*, vol. 31, 2018. 1, 4, 6, 9

[20] P. C. Mahalanobis, "On test and measures of group divergence," *Journal of Asiatic Society of Bengal*, vol. 26, pp. 541–588, 1930. 1

[21] G. Vareldzhan, K. Yurkov, and K. Ushenin, "Anomaly detection in image datasets using convolutional neural networks, center loss, and mahalanobis distance," in *2021 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT)*, pp. 0387–0390, IEEE, 2021. 1

[22] J. Ren, S. Fort, J. Liu, A. G. Roy, S. Padhy, and B. Lakshminarayanan, "A simple fix to mahalanobis distance for improving near-ood detection," *arXiv preprint arXiv:2106.09022*, 2021. 1, 4, 5, 6, 9, 10

[23] R. Kamoi and K. Kobayashi, "Why is the mahalanobis distance effective for anomaly detection?," *arXiv preprint arXiv:2003.00402*, 2020. 1, 4, 6, 9

[24] O. Dinari and O. Freifeld, "Variational-and metric-based deep latent space for out-of-distribution detection," in *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022. 1, 4, 6, 7, 8

[25] S. Carbonnelle and C. De Vleeschouwer, "Intraclass clustering: An implicit learning ability that regularizes dnns," in *International Conference on Learning Representations*, 2020. 2

[26] A. Venkataramanan, M. Laviale, C. Figus, P. Usseglio-Polatera, and C. Pradalier, "Tackling inter-class similarity and intra-class variance for microscopic image-based classification," in *International Conference on Computer Vision Systems*, pp. 93–103, Springer, 2021. 2

[27] Y. Em, F. Gag, Y. Lou, S. Wang, T. Huang, and L.-Y. Duan, "Incorporating intra-class variance to fine-grained visual recognition," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1452–1457, IEEE, 2017. 2

[28] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," 2009. 2, 6, 17, 20

[29] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," 2011. 2, 6

[30] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017. 2, 6, 16

[31] Y. LeCun, C. Cortes, and C. Burges, "Mnist handwritten digit database," *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, vol. 2, 2010. 2, 6, 16, 20

[32] E. Goan and C. Fookes, "Bayesian neural networks: An introduction and survey," in *Case Studies in Applied Bayesian Data Science*, pp. 45–87, Springer, 2020. 3

[33] L. V. Jospin, H. Laga, F. Boussaid, W. Buntine, and M. Bennamoun, "Hands-on bayesian neural networks—a tutorial for deep learning users," *IEEE Computational Intelligence Magazine*, vol. 17, no. 2, pp. 29–48, 2022. 3

[34] T. Chen, E. Fox, and C. Guestrin, "Stochastic gradient hamiltonian monte carlo," in *International conference on machine learning*, pp. 1683–1691, PMLR, 2014. 3

[35] L. Smith and Y. Gal, "Understanding measures of uncertainty for adversarial example detection," *arXiv preprint arXiv:1803.08533*, 2018. 3

[36] J. Liu, Z. Lin, S. Padhy, D. Tran, T. Bedrax Weiss, and B. Lakshminarayanan, "Simple and principled uncertainty estimation with deterministic deep learning via distance awareness," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7498–7512, 2020. 3, 6, 7, 8, 11, 18, 20, 21

[37] S. Fort, J. Ren, and B. Lakshminarayanan, "Exploring the limits of out-of-distribution detection," *Advances in Neural Information Processing Systems*, vol. 34, pp. 7068–7081, 2021. 4

[38] D. Pelleg, A. W. Moore, *et al.*, "X-means: Extending k-means with efficient estimation of the number of clusters.," in *Icml*, vol. 1, pp. 727–734, 2000. 4, 22

[39] S. Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982. 5

[40] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *kdd*, vol. 96, pp. 226–231, 1996. 5

[41] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008. 5

[42] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016. 7, 9, 17, 21, 22

[43] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 132–149, 2018. 11

[44] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with gpus," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019. 11

[45] M. Larsson, E. Stenborg, C. Toft, L. Hammarstrand, T. Sattler, and F. Kahl, "Fine-grained segmentation networks: Self-supervised segmentation for improved long-term visual localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 31–41, 2019. 11

[46] H. Du Buf, M. Bayer, S. Droop, R. Head, S. Juggins, S. Fischer, H. Bunke, M. Wilkinson, J. Roerdink, J. Pech-Pacheco, *et al.*, "Diatom identification: a double challenge called adiac," in *Proceedings 10th International Conference on Image Analysis and Processing*, pp. 734–739, IEEE, 1999. 11

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. 21, 22

[48] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, pp. 6105–6114, PMLR, 2019. 21, 22

[49] Z. Zhao, S. Guo, Q. Xu, and T. Ban, "G-means: a clustering algorithm for intrusion detection," in *Advances in Neuro-Information Processing: 15th International Conference, ICONIP 2008, Auckland, New Zealand, November 25-28, 2008, Revised Selected Papers, Part I 15*, pp. 563–570, Springer, 2009. 22

A Metrics Definitions

In this section, we provide the definitions and formulas of metrics used for evaluation in this paper. Let the samples be represented by $[(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)]$, where N is the total number of samples. x_i is the input and y_i is the corresponding label, having values between 1 and K.

Accuracy. This gives the fraction of samples that were correctly identified by the network.

$$acc = \frac{1}{N}\sum_{n=1}^{N} 1[\text{argmax}(p(y_n|x_n)) = y_n]$$

where, $p(y_n|x_n)$ is the predicted probability that the sample x_n belongs to the class y_n. A higher accuracy indicates better performance.

Expected Calibration Error. ECE is a measure of predictive probability calibration error. The output probability is divided into a histogram of B equally spaced bins. The expected calibration error gives the difference between the *observed relative frequency* (accuracy) and the *average predicted frequency* (confidence).

$$ECE = \sum_{b=1}^{B} \frac{n_b}{N} |acc(b) - conf(b)|$$

where n_b is the number of samples in bin b, N is the total number of samples, $acc(b)$ and $conf(b)$ are the accuracy and confidence of bin b. A lower ECE score means that the accuracy and confidence are aligned, indicating better calibration.

Negative Log Likelihood. NLL calculates the negative log-likelihood for the predicted class probability. While it is generally used for optimization using cross-entropy loss, it is also commonly used to evaluate the prediction uncertainty. A lower NLL score is preferred.

$$NLL = \frac{-1}{N}\sum_{n=1}^{N} log(p(y_n|x_n))$$

Area Under Receiver Operating Characteristic Curve. AUROC indicates the ability to separate ID and OOD samples. To calculate this metric, the predicted uncertainty is used to determine if a sample is ID or OOD. This can be considered as a binary classification problem. The area under the plot between the true positive rate and the false positive rate gives the AUROC value. Higher AUROC value means better separation between ID and OOD.

Area Under Precision-Recall Curve. AUPR, like AUROC measures the ability to separate ID and OOD samples. Considering ID and OOD separation as a binary classification problem, the area under the plot between precision and recall values give the AUPR score.

B Experimental details

B.1 FashionMNIST

FashionMNIST [30] consists of 10 classes. We split the original training set consisting of 60000 samples into train and validation set, in the ratio of 80:20. The validation set was used for hyperparameter tuning. The test set consists of 10,000 samples, which we used for inference and calculating the metrics. For analyses on OOD dataset, we use the test set of MNIST [31], containing 10,000 instances. For realistic evaluation, the normalization of MNIST is done the same way as FashionMNIST.

We use the network backbone of [18]. The CNN consists of three layers of convolution with 64, 128 and 128 3×3 filters, a dense layer for feature extraction and an output layer with softmax activation. Each convolutional layer is accompanied by a batch normalization and a 2×2 max pooling. The feature embedding's dimension is 256. The dimension of the final layer is equal to the total number of classes obtained after

clustering, which was 14 for MAPLE . We trained the network for 50 epochs. For training, we used an SGD optimizer with a learning rate of 0.05, momentum 0.9, weight decay of $1e^{-4}$. The training was performed on a 12Gb NVIDIA GeForce GTX 1080Ti with a batch size of 128. The reduced dimensional feature after PCA had a dimension of 5.

B.2 CIFAR10

CIFAR10 [28] consists of 10 classes. We split the original training set consisting of 50000 samples into train and validation set, in the ratio of 80:20. The validation set was used for hyperparameter tuning. The test set consists of 10,000 samples, used for inference. For OOD analyses, we use the test set of SVHN and CIFAR100, which consists of 26,032 and 10,000 samples respectively. The OOD images are normalized the same way as train images during inference.

The network architecture is Wide ResNet 28-10 [42]. The feature embedding layer has a dimension of 640. After training MAPLE , the number of classes were 12, and hence, the final layer has a dimension of 12, followed by softmax. We trained the model for 200 epochs. We used an SGD optimizer with a learning rate of 0.05. The momentum was set to 0.9 and weight decay of $1e^{-4}$. The training was performed on a 12Gb NVIDIA GeForce GTX 1080Ti with a batch size of 64. The dimension of the reduced features from PCA is 12.

B.3 Diatoms

The diatom dataset consists of 9895 individual RGB images of size 256×256, belonging to 166 classes. We divide it into ID dataset consisting of 130 classes (7874 images) and the remaining 36 classes as OOD (2021 images). 70% of the ID images were used for training, 10% for validation and 20% for testing. While training, horizontal and vertical flips were used for data augmentation.

The network architecture is Wide ResNet 28-10 [42]. The feature embedding layer has a dimension of 640. After training, there were a total of 158 classes, hence the output layer consists of 158 neuron with a softmax activation. We trained the model for 100 epochs with an Adam optimizer. The learning rate was $2e^{-4}$ and batch size 4. The training was performed on a 12Gb NVIDIA GeForce 1080Ti. The dimension of the features after PCA reduction was 31.

B.4 Hyperparameter Tuning

Our training depends on the following hyperparameters: (1) **Frequency of epochs** p - After every p epochs, validation is performed to obtain the new cluster assignments using X-Means. (2) **False negative ratio threshold** t - t is a threshold used to decide the class features to be clustered. From the normalized confusion matrix obtained during the validation step, the classes having false negative greater than t are clustered using X-Means. (3) **Maximum number of clusters** - This is a parameter of X-Means, that specifies the upper bound to the number of clusters that X-Means can generate for each class.

To find the optimal value of these parameters, a grid search was performed. For the grid search, the values of hyperparameters used were: False negative ratio threshold $t \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$, frequency of validation epochs $p \in \{5, 10, 15, 20\}$ and maximum number of clusters that X-Means can generate $\{3, 5, 7, 10\}$.

From the grid-search analysis, the best performance was obtained when $t = 0.3$, $p = 10$ and maximum number of clusters=5. These values worked best for all the datasets that were trained on.

B.5 Loss Functions

For our training, we use the Cross-Entropy Loss and the Triplet Loss.

B.5.1 Cross-Entropy Loss

To estimate the cross-entropy loss, the final layer of the model is passed through a softmax layer to obtain probability values. Cross-entropy loss increases proportional to the difference between the predicted probability and the actual probability (typically 1) of the ground truth class. The cross-entropy loss is given by:

$$\mathcal{L}_{\text{cross-entropy}} = -\sum_{i=1}^{K} y_i \log(p_i) \tag{7}$$

where K is the total number of samples, y_i is the binary one-hot encoding value corresponding to ground truth class, which equals 1, and p_i is the probability predicted by the network.

B.5.2 Triplet Loss

To estimate the triplet loss, we use the feature embedding obtained from the penultimate layer of the classification network. Triplet loss tries to minimize the distance of intra-class data points, while maximizing the inter-class distance. Consider three input samples, which are feature embeddings extracted: anchor x'_a, positive x'_p and negative x'_n. x'_a and x'_p belong to the same class while x'_n belongs to a different class. The triplet loss is given as:

$$\mathcal{L}_{\text{triplet}} = \max\{||x'_a - x'_p|| - ||x'_a - x'_n|| + \alpha, 0\} \tag{8}$$

The final objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cross-entropy}} + \mathcal{L}_{\text{triplet}} \tag{9}$$

C Algorithm

The proposed method is summarized in Algorithm 1 and Algorithm 2. Algorithm 1 provides the steps using in training MAPLE . Algorithm 2 summarizes the procedure for estimating uncertainty from MD. At regular intervals of the training process, validation is performed, and the train feature representations are clustered using X-Means. The time complexity for X-Means is O(log K), where K is the number of clusters. The train features are reduced in dimension using PCA, which has a complexity of O(nd^2+d^3), where n is the number of train data and d is the feature dimension. Mahalanobis distance calculation requires calculating mean and the covariance matrix, which has a complexity of O(n) and O(nd'^2), where d' is the PCA reduced feature dimension.

The algorithm requires the mean and covariance matrix calculation to be performed only once, at the end of the training. During inference, only the mean and covariance matrix from the train data is used to calculate the Mahalanobis distance for all the test points.

D Additional Experiments

In this section, we provide results for additional evaluation of MAPLE .

D.1 Qualitative Evaluation using Uncertainty Histograms

Uncertainty histograms (UH) are a means to visualize the uncertainty values predicted. When provided with OOD samples, it is expected that the network makes predictions with high uncertainty. The frequency of predicted uncertainties is plotted as a histogram. A high frequency at top uncertainty ranges show that the network is uncertain when provided with OOD. Fig. 6 shows the uncertainty histograms for the different approaches (deep ensemble [11], MC-Dropout [10], DUQ [18], DUE [14], SNGP [36] and MAPLE) on the three OOD datasets of FashionMNIST *vs.* MNIST, CIFAR10 *vs.* SVHN and CIFAR10 *vs.* CIFAR100.

Algorithm 1: MAPLE training

Data: Ground truth labels $\boldsymbol{y} \in \{1, 2, ...k\}$,
Input samples $\boldsymbol{x} \in \mathbb{R}^D$,
Train input samples $\boldsymbol{x_{train}} = \{x_n\}_{n=1}^N$,
Train dataset $\boldsymbol{\mathcal{D}_{train}} = \{(x_n, y_n)\}_{n=1}^N$,
Validation dataset $\boldsymbol{\mathcal{D}_{val}} = \{(x_v, y_v)\}_{m=1}^M$
Initialize: $n_c = k, p = 10, t = 0.3, max\_clusters = 5$
Model : $f^\theta : \mathbb{R}^D \rightarrow \mathbb{R}^d$
for *epoch = 1* **to** *max-epochs* **do**
  Train f^θ with $\boldsymbol{\mathcal{D}_{train}}$ and n_c classes and loss given by $L_{total} = L_{cross-entropy} + L_{triplet}$
  if *epoch%p==0* **then**
    $\boldsymbol{x'_{train}} = f^\theta(\boldsymbol{x_{train}})$
    Get softmax predictions on $\boldsymbol{\mathcal{D}_{val}}$
    if $n_c > k$, remap pseudo-labels to original class labels
    Compute confusion matrix
    for *i=1* **to** *k* **do**
      if *false_negative_ratio(i) > t* **then**
        Cluster using X-Means. X-Means($\boldsymbol{x'_{train}(i)}$, max_clusters)
    $K \leftarrow$ total number of clusters obtained from all the classes
    $n_c = K$
    Update $\boldsymbol{\mathcal{D}_{train}}$ with pseudo-labels from clustering

Algorithm 2: MAPLE Prediction

Data: Train feature embeddings $\boldsymbol{x'_{train}}$
Input: Test sample $\tilde{\boldsymbol{x}}$
Compute the reduced dimensional train features: $\boldsymbol{z_{train}} = g(\boldsymbol{x'_{train}})$
Compute individual class means and shared covariance $\mu_c, \boldsymbol{\Sigma}$
$\mu_c = \frac{1}{N_c} \sum_{i:y_i=c} z_i$
$\boldsymbol{\Sigma} = \frac{1}{N} \sum_c \sum_{i:y_i=K} (z_i - \mu_c)(z_i - \mu_c)^T$
Get reduced dimensional feature for $\tilde{\boldsymbol{x}}$: $\tilde{\boldsymbol{z}} = g(f^\theta((\tilde{\boldsymbol{x}}))$
Compute Mahalanobis distance: $MD(\tilde{\boldsymbol{x}}) = \sqrt{(\tilde{\boldsymbol{z}} - \mu_c)^T \boldsymbol{\Sigma}^{-1} (\tilde{\boldsymbol{z}} - \mu_c)}$
Get the prediction probabilities: $P_{MD} = 1 - \text{cdf}(\chi^2_{d'})(MD^2)$
Predicted class = argmax(P_{MD})
Compute uncertainty $u = \text{cdf}(\chi^2_{d'})(MD^2)$

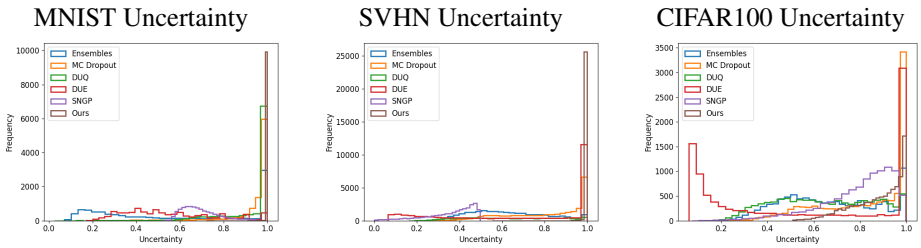

Figure 6: **Uncertainty Histograms for OOD datasets.** A high frequency of prediction at top uncertainty ranges indicate that the network is uncertain about its prediction. MAPLE outperforms the other methods and predicts OOD samples with high uncertainty. In other words, MAPLE is able to correctly identify these samples as OOD.

Results. From the plots, MAPLE assigns high uncertainty for OOD datasets. Compared to the other methods, MAPLE exhibits a higher frequency peak at an uncertainty around 1 for MNIST and SVHN. While MC-Dropout and DUE has a high frequency at uncertainty of 1 for CIFAR100, these methods also have a relatively higher peak at low uncertainties, since they make over-confident predictions on some OOD samples. Whereas, MAPLE's uncertainty values are spread across the higher end, which is desirable.

Method	acc@.50	acc@.80	acc@.90
MC Dropout [10]	0.962	0.976	0.988
Deep ensemble [11]	**0.967**	0.987	**0.995**
DUQ [18]	0.950	0.977	0.982
SNGP [36]	0.959	0.978	0.985
DUE [14]	0.962	0.974	0.979
MAPLE	0.958	**0.989**	**0.995**

Table 7: **Accuracy on CIFAR10 with different confidence levels.** MAPLE achieves top accuracy at confidence levels of 0.80 and 0.90.

D.2 Accuracy based on prediction confidence

We evaluate the accuracy of prediction when selecting samples with predictive confidence above a given threshold. In other words, classification is performed only when the network's confidence is above a threshold. This is representative of real-life applications where a network's prediction is considered only when the confidence is high. We consider three probability thresholds: 0.50, 0.80 and 0.90. For all samples with predictive probability above these values, we report the classification accuracy. Tables 7 and 8 give the results on the test set of CIFAR10 [28] and FashionMNIST [31] dataset respectively.

Results. MAPLE achieves the best accuracy at confidence values of 0.80 and 0.90 on CIFAR10. Overall, on both CIFAR10 and FashionMNIST, MAPLE has competitive accuracy with the other approaches. This shows that even though MAPLE is computationally efficient, it can achieve the same level or better performance as the other methods.

Method	acc@.50	acc@.80	acc@.90
MC Dropout [10]	0.924	0.931	0.948
Deep ensemble [11]	**0.946**	**0.975**	**0.983**
DUQ [18]	0.925	0.947	0.962
SNGP [36]	0.931	0.963	0.977
DUE [14]	0.929	0.951	0.964
MAPLE	0.930	0.972	0.974

Table 8: **Accuracy for FashionMNIST samples with different confidence levels.** MAPLE achieves competitive accuracies at different confidence values.

D.3 Gaussian test

In Section **??**, it was theoretically shown that X-Means creates clusters of feature points that are Gaussian. In this section, we empirically test this. A commonly adopted method to check for multivariate Gaussian is to use a quantile-quantile plot, where an observed quantile is compared with a theoretical one. If the samples are Gaussian, their squared MD follows a χ^2 distribution. Thus, we use $MD^2_{c^*}$ of the samples feature embeddings as our observed quantile and compare with theoretical χ^2 quantiles.

For our test, we use the reduced feature embeddings, $\boldsymbol{z_{train}}$, from a standard classifier network and MAPLE . The $MD^2_{c^*}$ of samples are calculated and plotted with χ^2 quantiles with d' degrees of freedom, where d' is the dimension of feature embeddings. We measure the error, which is the mean absolute difference between the two quantiles, to test which method generates feature embeddings that are closer to a Gaussian. In the ideal situation, this value should be zero. The larger the error, the greater is the deviation from a Gaussian distribution.

Table 9 shows the errors computed on feature embeddings from CIFAR10 and FashionMNIST dataset. From the results, MAPLE's error is reduced by over 50%, which shows that the feature representations of MAPLE are more Gaussian than when using a standard DNN classifier.

Method	CIFAR10	FashionMNIST
Standard CNN	3.540	2.564
MAPLE	**1.395**	**1.215**

Table 9: **Mean absolute error between squared MD and χ^2 distribution.** The lower the error, the more Gaussian are the samples. MAPLE's training generates sample distributions that are approximately Gaussian, fitting with the theoretical framework for MD calculation.

E Extended Ablation Analyses

E.1 MAPLE evaluated on different backbones

MAPLE is tested on three networks: Wide ResNet 28-10 [42], ResNet-18 [47] and EfficientNet-B0 [48]. Table 10 gives the quantitative metrics for evaluation on CIFAR10 *vs.* SVHN and CIFAR100. While it is expected that the accuracy depends on the architecture used, the calibration and OOD detection are also

influenced by the architecture. Wide ResNet, which has more number of parameters than the other two architectures, learns better feature representations for discriminating each class. As the model parameters decrease, there are overlapping feature points between different classes, which explains the lower accuracy and worse calibration and OOD metrics.

Architecture	Accuracy ↑	ECE ↓	SVHN AUROC ↑	CIFAR100 AUROC ↑
Wide ResNet 28-10 [42]	**0.954**	**0.012**	**0.996**	**0.926**
ResNet-18 [47]	0.945	0.029	0.979	0.886
EfficientNet-B0 [48]	0.902	0.035	0.942	0.893

Table 10: **MAPLE evaluated on different architectures.** The metrics improve as the model parameters increase, suggesting that the network learns better discriminative feature representations, thereby improving the performance.

E.2 Evaluation of different clustering methods

We analyse the performance of MAPLE when clustering is performed using K-Means, G-Means [49] and X-Means [38]. The value of K in K-Means is set to 3. Tab. 11 shows the results obtained. Based on the results, X-Means yields the best performance. K-Means and G-Means causes overclustering, which leads to worser performance on OOD detection. Using X-Means, we choose the optimal number of clusters, which performs superior to the others.

Clustering method	#Classes	Accuracy↑	ECE↓	SVHN AUROC↑	CIFAR100 AUROC↑
K-Means	30	0.952	0.154	0.871	0.850
G-Means	67	0.910	0.266	0.710	0.627
X-Means	12	**0.954**	**0.012**	**0.996**	**0.926**

Table 11: **Metrics for different frequency of validation epoch** #Classes refers to the total number of output classes obtained after clustering. K-Means and G-Means lead to overclustering, whereas using X-Means, the optimal number of clusters are generated leading to better performance.

E.3 Effect of False Negative Ratio t on FashionMNIST

For different values of False Negatives Ratio (t), X-Means clustering is performed on the extracted features while training MAPLE . Tab. 12 shows the metrics obtained on the FashionMNIST datasets respectively. A low value of t results in most of the classes getting clustered. This results in overclustering, where multiple clusters contain similar looking images. This further increases the chances of misclassifications, leading to decrease in the metric values. On the other hand, high t values result in underclustering. This is because, most of the classes have false negative ratios that are below this threshold and so, they are not clustered.

False Negative Ratio (t)	#Classes	Accuracy↑	ECE↓	AUROC↑
0.0	26	0.9149	0.028	0.942
0.1	19	0.9218	0.022	0.975
0.2	14	0.9242	0.023	0.990
0.3	14	**0.9244**	**0.020**	**0.995**
0.4	12	0.9241	0.024	0.988
0.5	10	0.9238	0.021	0.974
0.6	10	0.9240	0.023	0.969
0.7	10	0.9233	0.025	0.972
0.8	10	0.9229	0.023	0.969
0.9	10	0.9231	0.024	0.970
1.0	10	0.9232	0.022	0.971

Table 12: **Metrics for different values of False Negative Ratio evaluated on Fashion MNIST** #Classes refers to the total number of output classes obtained after clustering. A low value of t results in overclustering, whereas a high t fails to detect classes with high variance.

E.4 Effect of maximum number of clusters

Tab. 13 shows the results when the maximum number of clusters that can be generated for every class by X-Means is varied, along with different values of false negative ratio t for CIFAR10. For $t > 0.5$, none of the classes are clustered, and hence we do not include them. From the results, when the maximum number of clusters are low, MAPLE fails to capture all the within-class variances, whereas higher values result in overclustering. With the maximum number of clusters as 5, MAPLE achieves the best performance.

E.5 Effect of frequency of validation epochs.

Tab. 14 summarizes the metrics for CIFAR10 when the number of epochs after which the validation and cluster refinements are performed is varied. A low value of validation epochs does not give the network enough time to learn representations for the new clusters generated. Whereas, with larger number of epochs, the number of cluster refinements are low. In both these situations, the network does not identify the optimal clusters. MAPLE gives the best results when the validation is performed every 10 epochs.

F Proof of squared Mahalanobis Distance following a χ^2 distribution

In this section, we derive the proof that the squared Mahalanobis distance follow a χ^2 distribution with d' degrees of freedom, where d' is the dimension of the feature vectors used to calculate MD. A χ^2 distribution with d' degrees of freedom is defined as the distribution of a sum of the squares of d' independent standard normal random variables.

The squared Mahalanobis distance of $\boldsymbol{Z}$ and the mean vector $\vec{\mu}$ of a Multivariate Gaussian distribution is given as

$$D^2 = (\boldsymbol{Z} - \vec{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{Z} - \vec{\mu}) \tag{10}$$

Max. number of clusters	t	#Classes	Accuracy↑	ECE↓	SVHN AUROC↑	CIFAR100 AUROC↑
3	0.1	14	0.9542	0.012	0.996	0.925
	0.3	10	0.9540	0.014	0.972	0.919
	0.5	10	0.9533	0.012	0.958	0.917
5	0.1	18	0.9534	0.013	0.964	0.918
	0.3	12	0.9541	0.012	**0.996**	**0.926**
	0.5	10	0.9535	0.012	0.955	0.915
7	0.1	18	0.9537	0.013	0.959	0.894
	0.3	13	**0.9545**	0.012	0.992	0.921
	0.5	10	0.9531	0.013	0.944	0.911
10	0.1	26	0.9519	0.014	0.909	0.863
	0.3	22	0.9521	0.013	0.918	0.886
	0.5	11	0.9534	0.012	0.952	0.908

Table 13: **Effect of maximum number of clusters per class on MAPLES's performance.** A high value of cluster numbers causes overclustering whereas a low value does not generate enough clusters. A value of 5 results in optimal number of clusters for MAPLE to learn meaningful representations.

$\boldsymbol{\Sigma}$ is the covariance matrix, which is symmetric. By property of matrices, the matrix inverse and it's square root are also symmetric. Thus,

$$\begin{aligned} D^2 &= (\boldsymbol{Z} - \vec{\mu})^T \boldsymbol{\Sigma}^{-\frac{1}{2}} \boldsymbol{\Sigma}^{-\frac{1}{2}} (\boldsymbol{Z} - \vec{\mu}) \\ &= \left(\boldsymbol{\Sigma}^{-\frac{1}{2}} (\boldsymbol{Z} - \vec{\mu})\right)^T \left(\boldsymbol{\Sigma}^{-\frac{1}{2}} (\boldsymbol{Z} - \vec{\mu})\right) \end{aligned} \tag{11}$$

Let $\boldsymbol{W} = \boldsymbol{\Sigma}^{-\frac{1}{2}}$ and $\boldsymbol{X} = (\boldsymbol{Z} - \vec{\mu})$. The whitening transform is given as $\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{X}$ and $\boldsymbol{W}$ is also called the Mahalanobis whitening matrix. Eq. 11 can be written as

$$\begin{aligned} D^2 &= \boldsymbol{Y}^T \boldsymbol{Y} \\ &= ||\boldsymbol{Y}||^2 \\ &= \sum_{i=1}^{d'} Y_i^2 \end{aligned} \tag{12}$$

$(\boldsymbol{Z} - \vec{\mu}) \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, and so $\boldsymbol{Y}$ has zero mean. The covariance of $\boldsymbol{Y}$ is given as

$$\begin{aligned} \boldsymbol{\Sigma}_{\boldsymbol{Y}} &= \boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^T \\ &= \boldsymbol{\Sigma}^{-\frac{1}{2}} \boldsymbol{\Sigma} \boldsymbol{\Sigma}^{-\frac{1}{2}T} \\ &= \boldsymbol{\Sigma}^{-\frac{1}{2}} \left(\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\Sigma}^{\frac{1}{2}}\right) (\boldsymbol{\Sigma}^{-\frac{1}{2}})^T \\ &= \left(\boldsymbol{\Sigma}^{-\frac{1}{2}} \boldsymbol{\Sigma}^{\frac{1}{2}}\right) \left(\boldsymbol{\Sigma}^{-\frac{1}{2}} \boldsymbol{\Sigma}^{\frac{1}{2}}\right) \\ &= (\boldsymbol{I})(\boldsymbol{I}) = \boldsymbol{I} \end{aligned} \tag{13}$$

Validation epochs	#Classes	Accuracy↑	ECE↓	SVHN AUROC↑	CIFAR100 AUROC↑
5	16	0.895	0.025	0.914	0.876
10	12	0.954	0.012	**0.996**	**0.926**
15	12	**0.955**	0.012	0.987	0.922
20	10	0.953	0.013	0.968	0.917

Table 14: **Metrics for different frequency of validation epoch** #Classes refers to the total number of output classes obtained after clustering. With lower validation epochs, the clustering is too frequent for the network to learn meaningful representations. At lower frequency, the number of cluster refinements are not sufficient.

The covariance of $\boldsymbol{Y}$ is an identity matrix, which means that the elements from $\boldsymbol{Y}$ are drawn from an independent standard Gaussian distribution *i.e.,* $Y_i \sim \mathcal{N}(0, 1)$.

From the definition of the χ^2 distribution, we can infer that D^2 follows a χ^2 distribution with d' degrees of freedom.