Understanding Calibration of Deep Neural Networks for Medical Image Classification

Abhishek Singh Sambyal[a], Usma Niyaz[a], Narayanan C. Krishnan[b,*], Deepti R. Bathula[a]

[a]Department of Computer Science and Engineering, Indian Institute of Technology Ropar, Rupnagar, 140001, Punjab, India
[b]Department of Data Science, Indian Institute of Technology Palakkad, Palakkad, 678532, Kerala, India

Abstract

Background and Objective – In the field of medical image analysis, achieving high accuracy is not enough; ensuring well-calibrated predictions is also crucial. Confidence scores of a deep neural network play a pivotal role in explainability by providing insights into the model's certainty, identifying cases that require attention, and establishing trust in its predictions. Consequently, the significance of a well-calibrated model becomes paramount in the medical imaging domain, where accurate and reliable predictions are of utmost importance. While there has been a significant effort towards training modern deep neural networks to achieve high accuracy on medical imaging tasks, model calibration and factors that affect it remain under-explored.

Methods – To address this, we conducted a comprehensive empirical study that explores model performance and calibration under different training regimes. We considered fully supervised training, which is the prevailing approach in the community, as well as rotation-based self-supervised method with and without transfer learning, across various datasets and architecture sizes. Multiple calibration metrics were employed to gain a holistic understanding of model calibration.

Results – Our study reveals that factors such as weight distributions and the similarity of learned representations correlate with the calibration trends observed in the models. Notably, models trained using rotation-based self-supervised pretrained regime exhibit significantly better calibration while achieving comparable or even superior performance compared to fully supervised models across different medical imaging datasets.

Conclusion – These findings shed light on the importance of model calibration in medical image analysis and highlight the benefits of incorporating self-supervised learning approach to improve both performance and calibration.

Keywords: Calibration, deep neural network, fully-supervised, self-supervised, transfer learning, medical imaging.

1. Introduction

Recent advances in deep neural networks have shown remarkable improvement in performance for many computer vision tasks like classification, segmentation, and object detection (Krizhevsky et al., 2012; He et al., 2017). However, it is essential that model predictions are not only accurate but also well calibrated (Guo et al., 2017). Model calibration refers to the accurate estimation of the probability of correctness or uncertainty of its predictions. As calibration directly relates to the trustworthiness of a model's predictions, it is an essential factor for evaluating models in safety-critical applications like medical image analysis (Jiang et al., 2012; Kompa et al., 2021; Ma et al., 2022; Tomani and Buettner, 2019).

Probabilities derived from deep learning models are often used as the basis for interpretation because they provide a measure of confidence or certainty associated with the predictions. When a deep learning model assigns a high probability to a particular class, it indicates a stronger belief in that prediction. For example, in medical diagnosis, a high probability assigned to a certain disease can indicate a higher likelihood of its presence based on the observed input data. However, it is important to note that the reliability of interpretation based on probabilities depends on the calibration of the model (Murphy and Winkler, 1977; Guo et al., 2017; Caruana et al., 2015). Calibration ensures that the assigned probabilities reflect the true likelihood of events, allowing for accurate interpretation. Without proper calibration, the interpretation based solely on probabilities may be misleading or unreliable.

Apart from directly interpreting the probabilities as confidence for decision process, several explainability methods (van der Velden et al., 2022) have been proposed that depend on the information extracted from the model predictions like weighting random masks (Petsiuk et al., 2018), perturbation (Fong and Vedaldi, 2017; Uzunova et al., 2019), prediction difference analysis (Zintgraf et al., 2017), contribution scores (Shrikumar et al., 2017). The contribution of calibration to the model's explainability lies in providing reliable probability estimates, which aid in understanding the model's decision-making process and associated uncertainties. It is observed that the improved calibration has a positive impact on the saliency maps obtained as interpretations, also improving their quality in terms of faithfulness and are more human-friendly (Scafarto et al., 2023). This interplay

---

between explainability and calibrated predictions emerges as a pivotal factor in establishing a trustworthy model for medical decision support systems.

In healthcare, even minor errors in model prediction can carry life-threatening consequences. Therefore, incorporating uncertainty assessment into model predictions can lead to more principled decision-making that safeguards patient well-being. For example, human expertise can be sought in cases with high uncertainty. A model's predictive uncertainty is influenced by noise in data, incomplete coverage of the domain, and imperfect models. Effectively estimating or minimizing these uncertainties can markedly enhance the overall quality and reliability of the results (Jungo et al., 2020; Jungo and Reyes, 2019). Considerable endeavors have been dedicated to mitigating both data and model uncertainty through strategies like data augmentation (Singh Sambyal et al., 2022; Wang et al., 2019), Bayesian inference (Blundell et al., 2015; Gal and Ghahramani, 2016; Jena and Awate, 2019), and ensembling (Mehrtash et al., 2020; Lakshminarayanan et al., 2017)

Modern neural networks are known to be miscalibrated (Guo et al., 2017) (overconfident, i.e., high confidence but low accuracy, or underconfident, i.e., low confidence but high accuracy). Hence, model calibration has drawn significant attention in recent years. Approaches to improve the calibration of deep neural networks include post-hoc strategies (Platt, 1999; Guo et al., 2017), data augmentation (Zhang et al., 2018; Thulasidasan et al., 2019; Hendrycks et al., 2020) and ensembling (Lakshminarayanan et al., 2017). Similar strategies have also been utilized in the domain of medical image analysis to explore calibration with the primary goal of alleviating miscalibration (Frenkel and Goldberger, 2022; Larrazabal et al., 2021; Murugesan et al., 2023; Stolte et al., 2022). Furthermore, recent research has also investigated the impact of different training approaches on the model's performance and calibration. These include the use of focal loss (Mukhoti et al., 2020), self-supervised learning (Hendrycks et al., 2019c), and fully-supervised networks with pretraining (Hendrycks et al., 2019a). However, the scope of these studies has been limited to exploring calibration in the context of generic computer vision datasets like CIFAR10, CIFAR100, and ImageNet (Ericsson et al., 2021; Wang et al., 2023). Moreover, the majority of these studies have only utilized Expected Calibration Error (ECE) as the calibration metric. Unfortunately, ECE has several drawbacks, rendering it unfit for tasks like multi-class classification and inefficient due to bias-variance trade-off (Nixon et al., 2019). Nevertheless, as reliable and accurate estimation of predictive uncertainty is important, measuring calibration is an ongoing active research area resulting in many new metrics (Nixon et al., 2019; Singh et al., 2021; Thulasidasan et al., 2019; Guo et al., 2017; Nguyen and O'Connor, 2015).

Model calibration is tied to the training process that is inherently challenging for medical image analysis applications. The scarcity of labeled training datasets is a major cause for concern (Langlotz et al., 2019; Rahaman and thiery, 2021). Gathering labeled data for the medical domain is a daunting task due to the complex and intricate annotating process requiring domain expertise. *Transfer learning* is a popular learning paradigm to circumvent the labeled training data scarcity (Mei et al., 2022; Ma et al., 2022). Although transfer learning improves model accuracy, especially for smaller datasets, it also improves the quality of various complementary model components like adversarial robustness, and uncertainty (Hendrycks et al., 2019a). Remarkably, the literature suggests that the advantages of popular methods such as transfer learning on classical computer vision datasets do not extend to medical imaging applications (Raghu et al., 2019). *Self-supervised learning (SSL)* is another promising training regime when learning from scarce labeled data in classical computer vision applications (Tendle and Hasan, 2021; Doersch et al., 2015). Though fully-supervised (pretrained) and self-supervised approaches seem to improve various model performance measures like accuracy, robustness, and uncertainty (Hendrycks et al., 2019c; Navarro et al., 2021), the impact of the training regime(s) on model calibration is under-explored.

Our current work addresses these crucial gaps in the literature – understanding the calibration of deep neural networks for medical image analysis in the context of different training regimes and several calibration metrics. Accordingly, our main contributions are:

1. We study the effect of different training regimes on the performance and calibration of models used for medical image analysis. Specifically, we compare three different training paradigms: Fully-Supervised with random initialization (FS_r), Fully-Supervised with pretraining (FS_p), and Rotation-based Self-Supervision with pretraining (SSL_p).
2. We leverage several complementary calibration metrics to provide an accurate, unbiased, and comprehensive evaluation of the predictive uncertainty of models.
3. We assess the influence of varying dataset sizes, architecture capacities, and task complexity on the performance and calibration of the models.
4. We identified some of the potential factors that are correlated with the observed changes in the calibration of models. These include layer-wise learned representations as well as the weight distribution of the model parameters.

In general, we observe that the rotation-based self-supervised pretrained training approach provides better calibration for medical image analysis tasks than its fully supervised counterpart, with on-par or better performance. Additionally, our findings contradict recent literature (Raghu et al., 2019) that remarked *"transfer offers little benefit to performance"* for medical datasets. Furthermore, both the weight distribution and the learned representation analysis indicate that self-supervised training provides implicit regularization that in-turn achieves better calibration.

2. Methods

2.1. Training Regimes

2.1.1. Fully-Supervised and Transfer Learning

In a fully-supervised training regime, we use the given input data and the corresponding target value to learn the task. We can train models using two different ways, learning from scratch, i.e., initializing model weights randomly, or pretraining, i.e.,

Figure 1: Self-Supervised Learning Framework

transferring knowledge from one task to another by using the learned weights. In the transfer learning approach, a model is first pretrained using supervised learning on a large labeled dataset (Krizhevsky et al., 2012; Donahue et al., 2014). Then the learned generic representations are fine-tuned on the in-domain medical data (Raghu et al., 2019; Wen et al., 2021). Generally, fine-tuning a pretrained model achieves better generalized performance and faster convergence than training a fully-supervised network from scratch (Azizi et al., 2021; Girshick et al., 2014). We have considered FS_r as a baseline in our experiments where the model is trained from scratch. ImageNet pretraining is used as the default pretraining approach, which has shown remarkable performance on medical imaging datasets (Wen et al., 2021).

2.1.2. Self-Supervised Learning

In self-supervised training regime (Hendrycks et al., 2019c; Gidaris et al., 2018), Figure 1, we train a classifier network with a separate auxiliary head to predict the induced rotation in the image. The output of the penultimate layer is given to both the classifier and the auxiliary module. The classifier predicts a k-way softmax output vector based on the chosen task/dataset, whereas the auxiliary module predicts a 4-way softmax output vector indicating the rotation degree (0°, 90°, 180° and 270°). Given a dataset $\mathcal{D}$, of N training examples, $\mathcal{D} = \{x_i, y_i\}_{i=1}^N$, the goal is to learn representations using a self-supervised regime. The overall loss during training is the weighted sum of vanilla classification and the auxiliary task loss

$$\mathcal{L}(\theta) = \mathcal{L}(y, p(y|R_r(x)); \theta) + \lambda \mathcal{L}_{aux}(r, p(r|R_r(x)); \theta) \quad (1)$$

where, $R_r(x)$ is a rotation transformation on input image x and $r \in \{0°, 90°, 180°, 270°\}$ is the ground truth label for the auxiliary task. Note that the auxiliary component does not require ground truth training label y as input. $\mathcal{L}_{aux}$ is the cross-entropy between r and the predicted rotation.

2.2. Calibration Metrics

Perfect Calibration: In a multi-class classification problem, let the input be X and the label $Y \in \{1, 2, \cdots, K\}$ and f the learned model. The model's output is $f(X) = (\hat{Y}, \hat{P})$ where $\hat{Y}$ is a class prediction and $\hat{P}$ is its associated confidence. If $\hat{P}$ is always the

true probability, then we call the model perfectly calibrated as defined in (2).

$$\mathbb{P}\left(\hat{Y} = Y \mid \hat{P} = p\right) = p, \quad \forall p \in [0, 1] \quad (2)$$

The difference between the true confidence (accuracy) and the predicted confidence (output probability), $|\mathbb{P}\left(\hat{Y} = Y \mid \hat{P} = p\right) - p|$ for a given p is known as calibration error or miscalibration. Note that $\hat{P}$ is a continuous random variable, the probability in (2) cannot be computed using finitely many samples resulting in different approximations for the calibration error as discussed below.

2.2.1. Expected Calibration Error (ECE)

The most common miscalibration measure is the ECE (Naeini et al., 2015; Guo et al., 2017), which computes the difference in the expectation between confidence and accuracy. It is a scalar summary statistic of calibration.

$$\mathbb{E}_{\hat{P}}\left[\left|\mathbb{P}\left(\hat{Y} = Y \mid \hat{P} = p\right) - p\right|\right] \quad (3)$$

In practice, we cannot estimate ECE without quantization; therefore, the confidence scores for the predicted class are divided into m equally spaced bins. For each bin, the average confidence (conf) and accuracy (acc) are computed. The difference between the average confidence and accuracy weighted by the number of samples summed over the bins gives us the ECE measure. Formally,

$$\text{ECE} = \sum_{m=1}^{M} \frac{n_m}{N} |\operatorname{acc}(m) - \operatorname{conf}(m)| \quad (4)$$

where n_m is the number of predictions in bin m. While ECE is used extensively to measure calibration, it has some major drawbacks (Nixon et al., 2019):

(i) Structured around binary classification, ECE only considers the class with maximum predicted probability. As a result, it discounts the accuracy with which the model predicts other class probabilities in a multi-class classification setting.

(ii) Deep neural network predictions are typically overconfident, causing skewness in the output probabilities. Consequently, equal-interval binning metrics like ECE is impacted by only a few bins.

(iii) The number of bins, as a hyperparameter, plays a crucial role in the quality of calibration estimation. However, determining the optimal number of bins is challenging due to the bias-variance tradeoff.

(iv) In a static binning scheme like ECE, overconfident and underconfident predictions occurring in the same bin result in a reduction of calibration error. In such cases, it is difficult to infer the true cause of improvement in model calibration.

These issues have resulted in the development of novel calibration metrics discussed in the following subsections.

3

2.2.2. Adaptive Calibration Error (ACE)

As *ECE* suffers from skewness in the output predictions, ACE mainly focuses on the regions where the predictions are made. It uses an adaptive binning scheme to ensure an equal number of predictions in each bin (Nguyen and O'Connor, 2015; Nixon et al., 2019). Formally,

$$\text{ACE} = \frac{1}{KR} \sum_{k=1}^{K} \sum_{r=1}^{R} |\text{acc}(r, k) - \text{conf}(r, k)| \quad (5)$$

where, $\text{acc}(r, k)$ and $\text{conf}(r, k)$ represent the accuracy and confidence for the adaptive calibration range or bin r and class label k, respectively. Due to adaptive binning, the bin spacing can be unequal; wide in the areas where the number of data points is less, and narrow otherwise.

2.2.3. Maximum Calibration Error (MCE)

It refers to the upper-bound estimate of miscalibration useful in safety-critical applications. MCE (Naeini et al., 2015; Guo et al., 2017) captures the worst-case deviation between confidence and accuracy by measuring the maximum difference across all bins m, as shown below:

$$\text{MCE} = \max_{m \in \{1, \dots, M\}} |\text{acc}(m) - \text{conf}(m)| \quad (6)$$

2.2.4. Overconfidence Error (OE)

Modern deep neural networks provide high confident outputs despite being inaccurate. Thus a metric that captures the model's overconfidence provides better model insights. OE (Thulasi-dasan et al., 2019) captures the overconfidence in the model prediction by penalizing the confidence score only when the model confidence is greater than the accuracy.

$$\text{OE} = \sum_{m=1}^{M} \frac{n_m}{N} \left[\text{conf}(m) \times \max \left(\text{conf}(m) - \text{acc}(m), 0 \right) \right] \quad (7)$$

2.2.5. Brier or Quadratic Score

It is a strictly proper scoring rule that measures the accuracy of the probabilistic predictions (Brier, 1950; Gneiting and Raftery, 2007; Kruppa et al., 2014). It is the mean squared difference between one-hot encoded true label and predicted probability. Formally,

$$\text{Brier} = \sum_{k=1}^{K} (\mathbb{1}_{[Y=k]} - \hat{P}(Y = k \mid X))^2 \quad (8)$$

2.2.6. Negative Log Likelihood (NLL)

For safety-critical applications, using a probabilistic classifier that predicts the correct class and gives the probability distribution of the target classes is encouraged. Using NLL, we can evaluate models with the best predictive uncertainty by measuring the quality of the probabilistic predictions (Vaicenavicius et al., 2019; Kull and Flach, 2015; Quiñonero-Candela et al., 2006). Formally,

$$\text{NLL} = - \sum_{k=1}^{K} \mathbb{1}_{[Y=k]} \log[\hat{P}(Y = k \mid X)] \quad (9)$$

Additionally, *Root Mean Square Calibration Error (RMSCE)* (Nguyen and O'Connor, 2015; Hendrycks et al., 2019a,b) measures the square root of the expected squared difference between confidence and accuracy. As it defines the magnitude of miscalibration, it is highly correlated to ECE. Similar to ACE, *Static Calibration Error (SCE)* (Nixon et al., 2019), extends ECE by measuring calibration over all classes in each bin for a multiclass setting but does not use an adaptive binning approach. As a result, we exclude these metrics from our experimental analysis. It can be observed from the above definitions that none of the individual metrics takes a holistic approach. Hence, it is important to recognize that individual metrics are limited in their ability to provide accurate estimates of calibration. Consequently, a collective evaluation of these metrics is necessary for a better or unbiased understanding of calibration performance.

2.3. Experimental Setup

2.3.1. Datasets

We used three different datasets to investigate the classification performance and calibration of models trained under different regimes. The datasets have varying characteristics such as different imaging modalities, and sizes.

- The Diabetic Retinopathy (DR) dataset contains 35K high-resolution ($\sim 5000 \times 3000$) retinal fundus scans (EyePACS, Diabetic Retinopathy Detection). Each image is rated for the severity of diabetic retinopathy on a scale of 0-4, which makes it a five-class classification problem. The images are captured under varying imaging conditions, like different models and camera types.

- The Histopathologic Cancer dataset contains 220K images (patches of size 96×96) extracted from larger digital pathological scans (Ehteshami Bejnordi et al., 2017; Histopathologic Cancer Detection: Modified version of the PatchCamelyon (PCam) Benchmark Dataset; Veeling et al., 2018). Each image is annotated with a binary label indicating the presence of tumor tissue in the histopathologic scans of lymph node sections.

- The COVID-19 is a small dataset consisting of 317 high-resolution ($\sim 4000 \times 3000$) chest X-rays images (Covid-19 Image Dataset; Cohen et al., 2020a,b). This dataset corresponds to a three-class classification problem.

Both DR and Histopathology cancer datasets are segregated into four training datasets of sizes: 500, 1000, 5000, and 10000; and a common test dataset of 2000 images. The *Covid-19* dataset is partitioned into 60/20 train/validation split and a separate 20% test set for evaluation. The images in all the datasets are resized to 224×224, which is the standard input resolution for ResNet architectures.

2.3.2. Implementation Details

Architectures – Due to the popularity of ResNet architectures in medical imaging for classification tasks (Azizi et al., 2021; Wen et al., 2021; Mei et al., 2022), we choose the standard ResNet18, ResNet50 (He et al., 2016), and WideResNet (Zagoruyko and

Komodakis, 2016) architectures as the network backbone to simulate small, medium, and large architecture sizes, respectively. For the training regimes relying on a pretrained model, we initialize the backbone architectures using ImageNet-pretrained weights, and the classifier and self-supervised modules using the Kaiming uniform initialization (He et al., 2015) variant.

Evaluation Metrics – We use two performance metrics - *Accuracy* and *Area under the Receiver Operating Characteristic curve (ROC AUC)*; and six calibration metrics - *ECE, MCE, ACE, OE, Brier* and *NLL*. The architecture details and hyperparameter settings are presented in the supplementary material Section 5.1.

3. Results

3.1. Effect of Training Regimes on Calibration

In this study, we investigate the performance and calibration of three different architectures - *ResNet18, ResNet50 & WideResNet* using three different training regimes - *Fully-Supervised with random initialization* (FS_r), *Fully-Supervised with pretraining* (FS_p) and *Rotation-based Self-Supervision with pretraining* (SSL_p).

For medical image analysis, both the accuracy and reliability of the models are crucial. In this context, there are two key scenarios we need to consider:

1. *High accuracy and high calibration error* – When a model has high accuracy but is miscalibrated, the model's predictions may not be trustworthy. Both incorrect predictions with high confidence and correct predictions with low confidence are detrimental in healthcare applications. Reliance on accuracy alone is hazardous.

2. *High accuracy and low calibration error* – This is the ideal scenario, where a model has high accuracy and well-calibrated confidence scores. Predictions from such a model can be trusted in the decision-making process.

3.1.1. Effect of Architecture and Dataset Size

In this section, we present the findings of our analysis of the DR dataset. The performance and calibration scores of various architectures, as well as the effects of increasing training dataset size,

Figure 2: Joint evaluation for performance and calibration across different dataset sizes (x-axis) using WideResNet architecture on Histopathology dataset. The shaded region corresponds to $\mu \pm \sigma$, estimated over 3 trials. ↑: higher is better, ↓: lower is better.

are depicted in Figure 3 for the three different training regimes. Similar results and analysis of WideResNet architecture on the Histopathology dataset is presented in Figure 2 and rest can be found in the supplementary material (Figure 11). Owing to the difficulty of the task, the performance of all training regimes across all the models is not very high ($\leq 75\%$). However, we do see a clear improvement in performance as the training dataset size increases across all architectures and regimes. Additionally, we observe that initializing models with pretrained weights (with SSL_p having an edge over FS_p) offer a significant advantage over random initialization, which contradicts existing assumptions that transfer learning from ImageNet models is not beneficial. Both FS_p and SSL_p result in similar performance when using larger models (Raghu et al., 2019).

Comparing the effect of FS_p and SSL_p training regimes on calibration, we see that SSL_p significantly improves calibration across all metrics for all architectures and training dataset sizes as illustrated in Figures 3(c)-(h). The gap in the calibration metrics for SSL_p and FS_p is highest when using the largest architecture (WideResNet). While a randomly initialized model (FS_r) results in marginally better calibration (sometimes even better than SSL_p), the performance is significantly poor. Overall, we observe that models trained using self-supervision with pretrained weights show better or similar performance with a significant improvement in calibration error compared to fully-supervised pretraining. These results suggest that self-supervised training can help improve both performance and calibration, leading to more robust and reliable models for medical image analysis.

We discuss the results on the Covid-19 dataset separately owing to its small size. Figure 4 depicts that all the models result in high performance on this dataset indicating the ease of learning the task. The superior performance of FS_p and SSL_p indicate a definite advantage of transfer through pretrained over random initialization, contradicting the recent findings (Raghu et al., 2019). It is also evident that larger models result in better performance than shallow models. The negative impact of training from a random initialization (FS_r) for over-parameterized models is also evident from the drop in the performance and calibration with the increase in architecture size. While we observe a significant difference in the performance, there is only a marginal change in the calibration metrics. There is no definite trend in the calibration across the three training regimes. Thus, while transfer seems to have a positive impact on performance, calibration does not enjoy a commensurate impact.

3.1.2. Issues with using Single Calibration Metric

In this section, we discuss the importance of collective evaluation of calibration metrics. For this purpose, let's consider the question - *Does transfer learning improve calibration?* In the context of DR dataset, we analyze the results in Figure 3. Comparing FS_r and FS_p using only *Brier* for all architectures and dataset sizes, the general trend we observe is that transfer learning improves calibration. However, this observation fails when we chose *ECE* metric, which gives us mixed results. Similarly, incorrect conclusions could be drawn when using individual metrics like *NLL* and *ACE*.

5

Figure 3: Joint evaluation for performance and calibration across different dataset sizes (x-axis) and architectures for DR dataset. The shaded region corresponds to $\mu \pm \sigma$, estimated over 3 trials. The underline shows the statistical significance between FS_p and SSL_p. Black and Pink color signifies $p < 0.05$ and $0.05 < p < 0.1$ level of significance, respectively. ↑: higher is better, ↓: lower is better.

Figure 4: Comparing performance and calibration across different architectures and training regimes for *Covid-19* dataset. The error bars correspond to $\mu \pm \sigma$, estimated over 3 trials. Relying on a single calibration error metric, such as ECE or ACE, can lead to conflicting conclusions when it comes to model selection. By considering a combination of metrics, we gain a more comprehensive understanding of the model's calibration performance. ↑: higher is better, ↓: lower is better.

Likewise, we consider the effect of architecture on performance and calibration in the context of the small Covid-19 dataset. From Figure 4, we observe that FS_p and SSL_p have comparable performances with nominal improvement with increasing architecture size. In this case, using only *ECE* as the calibration metric would lead us to infer that FS_p provides better calibration than SSL_p for large capacity models. In contrast, *ACE* suggests the opposite. However, these two training regimes are quite similar across most other metrics.

These examples further highlight that in scenarios where models provide mixed calibration results, selecting the best model is non-trivial/subjective. In section 4, we discuss some potential model selection criteria to address this issue.

3.2. Factors affecting Performance and Calibration

In this section, we explore two potential factors linked to the enhanced calibration of the self-supervised training regime. Firstly, we examine the standard deviation of weight distributions and calibration metrics across different training regimes. Secondly, we investigate the similarity of learned representations in the activations.

3.2.1. Weight Distribution

The weight distribution of a neural network can provide useful insights into the model's performance. Regularization schemes like $\mathcal{L}_1$, $\mathcal{L}_2$, dropout (Ng, 2004; Srivastava et al., 2014) are often employed to find optimal parameters of a model with low generalization error. By adding a parameter norm penalty term

to the objective function, the $\mathcal{L}_1$ and $\mathcal{L}_2$ norms encourage sparse weights with many zero values and small weight values respectively. Weighting the contribution of the penalty term controls the regularization effect. For instance, with $\mathcal{L}_2$ norm, the histogram of weights tends to a zero-mean normal distribution with a high penalty that causes the model to underestimate the weights and hence leads to underfitting. In contrast, a low penalty yields a flatter histogram that causes the model to overfit the training data. To strike the right balance, careful hyperparameter tuning is needed to determine the data-dependent optimal penalty term contribution for better generalization. Based on this intuition, we attempt to interpret the performance and calibration of networks trained using different regimes using weight distribution analysis. To the best of our knowledge, the calibration of a model has not been explained in the context of the weight distribution of a network, especially for medical image analysis.

The comparison of weight distributions between the models trained using FS_r, FS_p, and, SSL_p for the DR dataset in Figure 5a-(1),(2) reveals some interesting observations. The weight distribution of the model trained with FS_p exhibits a higher peak than SSL_p, indicating that most of the weights are small. Conversely, the FS_r model exhibits the highest standard deviation, resembling a uniform distribution. Now, the question arises: which distribution is preferable, and which scenario leads to better generalization with improved calibration? To address this, we analyze the impact of weight distribution on the performance and calibration of FS_p and SSL_p models using Figure 3 and Figure 5. We observe that both models show similar AUC performance,

with SSL_p displaying a smaller peak in the weight distribution. This difference in weight distribution influences the calibration metrics, where SSL_p demonstrates significantly lower calibration error across most metrics. In other words, the predicted probabilities align more closely with the true probabilities using the SSL_p model.

For Histopathology dataset, the weight distribution of the SSL_p model is similar to that of the FS_p, as seen in Figure 5b-(1),(2). This similarity in weight distribution could be attributed to an easier task, leading to higher test performance. However, despite the similarity in weight distribution, the SSL_p model still provides better-calibrated outputs compared to the FS_p, but the difference in calibration error between these training regimes is now smaller. Considering the standard deviation of the weight distributions, it is suggested that a balance in the spread of weights is important for achieving good performance and calibration. It is important to note that the FS_r model has the highest standard deviation and comparable calibration error, it exhibits low AUC performance, making it inconsequential among other training regimes.

In Figure 5-(3),(4), we analyze the layer-wise standard deviation and Frobenius norm of the weights. In Figure 5a, we observe SSL_p influence on the standard deviation and weight magnitudes in every layer of the network. Additionally, we notice that the standard deviation tends to be higher in the initial layers and decreases as we move towards higher layers of the network. In Figure 5b, the standard deviation and magnitude of weights are similar for both SSL_p and FS_p training regimes. This suggests that the features extracted by each layer of the network are similar, which could be attributed to the high performance achieved by both training regimes. Despite the similarity, the SSL_p training regime still produces a better-calibrated model than the FS_p, indicating the additional benefits of self-supervised training. For a more comprehensive analysis, Figure 6 further consolidates the trends between performance, the standard deviation of the weights, and model calibration. The figure highlights that achieving good performance and calibration in a model necessitates finding a balance in the spread of weights, a balance which the SSL_p training regime was able to achieve successfully. Due to the different scales of the calibration metrics, we plot them on multiple axes. The weight values and their standard deviation are very small; therefore, we scaled them by 10^2. In Figure 6a, FS_r (top left, orange) has the highest standard deviation (wide distribution) and gives us the best calibration error (x-axis) but the worst performance compared to other training regimes. The standard deviation for FS_p (bottom right, red) is the lowest, but the calibration error is still high, which is not ideal. On the other hand, SSL_p has a low standard deviation but yields the best performance and calibration. So, when we encounter the gap in the standard deviation of weights between different training regimes (SSL_p and FS_p), we observe the calibration error metrics are well separated (Figure 6a). Alternatively, when the gap is negligible, the calibration error metrics overlap (Figure 6b). In summary, we observe that the $SSLp$ training regime consistently provides better calibration than the FS_p regime for both datasets. The magnitude of improvement or change in calibration is directly related to the

Figure 5: Comparing different aspects of WideResNet learned weights for dataset size 10000 on DR-(a) and Histopathology Cancer-(b) datasets. (1) and (2) the normalized histogram of weights of three training regimes. (3) Layer-wise comparison of standard deviation (SD) between FS_p and SSL_p. (4) Layer-wise comparison of Frobenius norm between FS_p and SSL_p.

differences in weight distributions.

3.2.2. Learned Representation

In addition to the diversity of the whole weight space, we explore the impact of layer-wise, learned neural representations on performance and calibration. Towards this end, we use the widely popular Centered Kernel Alignment (CKA) (Kornblith et al., 2019) metric that measures the similarity between the activations of hidden layers in a neural network. Literature suggests that high representational similarity across layers indicates redundancy in learned representations of a network. Furthermore, redundant representations impact the generalizability due to the influence of regularized training (Doimo et al., 2022), which in turn improves the model calibration (Guo et al., 2017).

CKA analysis of WideResNet's layer representations for different training regimes on the DR dataset is shown in Figure 7. The CKA plots for FS_p and SSL_p depict comparatively similar patterns. However, the higher layers of FS_p show a significant decrease in representational similarity (darker region shown in blue box) with increasing dataset size. The relatively high CKA values of the deeper layers of SSL_p depict redundancy of learned representations lighter regions) that provides implicit regularization. This in turn explains the reduced calibration error of SSL_p compared to FS_p as seen in Figure 3. A similar

Figure 6: Comparing calibration metrics (x-axis) vs. standard deviation (SD, y-axis) of WideResNet architecture for dataset size 10000 on DR and Histopathology cancer datasets. Colors represent training regimes (orange for FS_r, blue for SSL_p, and red for FS_p), and markers are the lowercase initials of each calibration metric; e – ECE, o – OE, a – ACE, m – MCE, b – Brier, n – NLL. Alongside each calibration error cluster, the performance is also reported. Ideally, the metrics should be at the bottom left with comparable performance. **(a)** SSL_p has less calibration error with on-par performance than FS_p training regime, indicating it to be a suitable choice. Calibration error metrics clusters of SSL_p and FS_p are noticeably well separated, correlating with the gap in their SD. **(b)** Here, SSL_p seems to be the best in calibration and performance compared to other training regimes. The noticeable difference we observed here is that the calibration error metrics clusters of SSL_p and FS_p are close (somewhat overlapping) when the SD of their weight distributions are similar.

pattern is observed for ResNet18 and ResNet50 architectures as depicted in Figure 10 of the supplementary material. For the Histopathology dataset, the CKA plots (shown in Figure 12 of the supplementary material) for FS_p and SSL_p depict very similar patterns that explain comparable performance and calibration afforded by these training regimes.

To facilitate quantitative comparison, we present the mean CKA value as a summary statistic to represent the CKA plots of individual networks in Table 2 of the Supplementary Material, Section 5.5. While not very significant, these findings align with the trends observed in Figure 7. Furthermore, the difference in the mean CKA values of SSL_p and FS_p fairly correlates with the difference in the magnitude of the calibration metrics of these regimes.

4. Discussion

For safety-critical applications like medical image analysis, it is imperative to choose models with high accuracy and low calibration errors. In this study, we investigate the performance and calibration of three different architectures using three different training regimes on medical imaging datasets of varying sizes and task complexities. Furthermore, we use six complementary calibration metrics that collectively provide a comprehensive evaluation of the predictive uncertainty of the models.

Model selection with mixed calibration results – While using multiple calibration metrics provides a more comprehensive evaluation, deciding on the best model can still be challenging as observed in Section 3.1.2. There are a few strategies that can be employed to aid in the decision-making process. One approach is to use a voting-based scheme, where each model is assigned a vote based on its performance across the calibration metrics. The model with the maximum number of votes is then selected as the best choice. This approach treats all metrics equally and can be useful when there is no significant variation in the importance of different metrics.

Domain specific metric relevance – However, it is important to consider that different calibration metrics may have different objectives and importance in specific domains. For example, metrics like OE (Overconfidence Error) explicitly measure the overconfidence of the model predictions, while MCE (Maximum Calibration Error) provides an upper bound on the mistakes made by the model. In such cases, it might be necessary to assign more weightage to these important metrics during the voting process. The determination of metric importance is subjective and can vary depending on the application. Expert knowledge and domain expertise play a crucial role in assigning relative importance to different metrics. By incorporating the opinions of experts, the voting process can be tailored to reflect the specific requirements of the application.

Margin for model selection – In addition to assigning weights

9

Figure 7: CKA plots of trained WideResNet architecture using fully-supervised (pretrained, FS_p) and self-supervised (pretrained, SSL_p) regime for DR dataset. The plots represents similarity between representations of features. The range of the CKA metric is between 0 and 1, with 0 indicating two completely distinct activations (not similar) and 1 indicating two identical activations (similar).

to metrics, introducing a margin or threshold in the voting scheme can help refine the model selection process. This threshold represents the minimum difference in calibration error between two training regimes that must be surpassed for a metric to be considered in the model selection. By setting a threshold, the metrics can be filtered out that do not exhibit significant differences and focus on those that have a substantial impact on model calibration.

It is worth noting that the difficulty of choosing a model also arises when one model has higher accuracy but poorer calibration while another model has lower accuracy but better calibration. This dilemma has been discussed in the literature (Minderer et al., 2021), highlighting the need for careful consideration of calibration metrics during model selection. *Selective prediction* is one scenario where we abstain the classifier that gives us low-confident predictions based on some threshold or cost structure of the specific application (Hernández-Orallo et al., 2012). In such cases, low-confidence predictions are referred to an expert for further analysis or diagnosis. This approach allows for cautious decision-making when the model's confidence is not sufficient for reliable predictions. Overall, the selection of the best model with mixed calibration results requires a combination of objective evaluation, subjective judgment of metric importance, and consideration of domain-specific requirements.

Calibration Metrics – While we have elaborated on the drawbacks of ECE, it provides an intuitive and straightforward interpretation, is simple to implement, and captures pure calibration. Additionally, ECE is associated with the reliability diagram - a powerful tool to visualize model calibration. It's also worth noting that alternative calibration metrics have their own shortcomings. The majority of the existing metrics suffer from challenges like scale-dependent interpretation, lack of normalized range, arbitrary choice of number of bins, etc. (Matsubara et al., 2023). Moreover, composite measures like NLL and Brier blend calibration and refinement, making it challenging to isolate calibration effects. Multiclass settings introduce additional complexity due to the multitude of classes, their diverse interrelations, and the

absence of a universally accepted metric for gauging refinement. Moreover, the choice of calibration metric can also be domain or application-dependent. As there is no universally applicable or acceptable calibration metric, we proposed collective evaluation of these metrics for a better or unbiased understanding of calibration performance.

Limitations – Our current study focused on medical image classification tasks across three different benchmark datasets. However, due to limited computational resources, we selected datasets with 2D images. Extending this work to 3D datasets as well as other tasks like medical image segmentation and registration, can help broaden our understanding of calibration in the general context of medical image analysis. Additionally, our study highlights that using the rotation-based self-supervised learning (SSL) approach gives better-calibrated results compared to the usual fully-supervised learning. A comparison of other SSL techniques, such as contrastive SSL or generative SSL, would be interesting.

Conclusion – In general, for medical image classification tasks, we observe that training regimes have a varying impact on model calibration. Overall, we observe that across different architectures, training regimes, datasets, and sample sizes, (a) transfer learning through pretraining helps improve performance over random-initialized models and (b) pretrained self-supervised approach provides better calibration than its fully supervised counterpart, with on-par or better performance. While we notice a sizeable increase in performance with dataset sizes, only nominal improvement is realized with increasing model capacity.

Furthermore, we identified weight distribution and learned representations of a neural network as potential confounding factors that provide useful insights into model calibration, in particular, to explain the superiority of a rotation-based self-supervised training regime over fully supervised training.

Broader Impact – We anticipate that this analysis will offer significant insights into calibration across datasets of varying sizes and models of different complexities. This work raises a

broader question regarding the search for a unified metric that can provide a comprehensive understanding of model calibration, thereby reducing the need to evaluate models based on multiple criteria. Ensuring accurate and reliable probabilistic predictions is vital for effective risk management and decision-making. It is particularly important when relying on the outputs of probabilistic models that require trust. Additionally, developing well-calibrated models is essential for promoting the widespread acceptance of machine learning methods, especially in fields like AI-driven medical diagnosis, as it directly influences the level of trust in new technologies and improves their explainability.

Acknowledgment

References

Shekoofeh Azizi et al. Big self-supervised models advance medical image classification. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 3, 4

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *International Conference on Machine Learning (ICML)*, 2015. 2

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 1950. 4

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015. 1

Joseph Paul Cohen, Paul Morrison, and Lan Dao. Covid-19 image data collection. *arXiv 2003.11597*, 2020a. URL https://github.com/ieee8023/covid-chestxray-dataset. 4

Joseph Paul Cohen, Paul Morrison, Lan Dao, Karsten Roth, Tim Q Duong, and Marzyeh Ghassemi. Covid-19 image data collection: Prospective predictions are the future. *arXiv 2006.11988*, 2020b. URL https://github.com/ieee8023/covid-chestxray-dataset. 4

Covid-19 Image Dataset. https://www.kaggle.com/datasets/pranavraikokte/covid19-image-dataset. 4

Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 2

Diego Doimo, Aldo Glielmo, Sebastian Goldt, and Alessandro Laio. Redundant representations help generalization in wide neural networks. In *Neural Information Processing Systems (NeurIPS)*, 2022. 8

Jeff Donahue, Yangqing Jia, et al. Decaf: A deep convolutional activation feature for generic visual recognition. In *International Conference on Machine Learning (ICML)*, 2014. 3

Babak Ehteshami Bejnordi et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*, 2017. 4

Linus Ericsson, Henry Gouk, and Timothy M. Hospedales. How well do self-supervised models transfer? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

EyePACS, Diabetic Retinopathy Detection. https://www.kaggle.com/competitions/diabetic-retinopathy-detection/. 4

Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. *IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

Lior Frenkel and Jacob Goldberger. Calibration of medical imaging classification systems with weight scaling. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2022. 2

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, 2016. 2

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 2018. 3

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 3

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association (JASA)*, 2007. 4

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, 2017. 1, 2, 3, 4, 8

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *IEEE International Conference on Computer Vision (ICCV)*, 2015. 5

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning (ICML)*, 2019a. 2, 4

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations (ICLR)*, 2019b. 4

Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Neural Information Processing Systems (NeurIPS)*, 2019c. 2, 3

Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. *International Conference on Learning Representations (ICLR)*, 2020. 2

José Hernández-Orallo, Peter Flach, and Cèsar Ferri. A unified view of performance metrics: Translating threshold choice into expected classification loss. *Journal of Machine Learning Research*, 2012. 10

Histopathologic Cancer Detection: Modified version of the Patch-Camelyon (PCam) Benchmark Dataset. https://www.kaggle.com/competitions/histopathologic-cancer-detection/. 4

Rohit Jena and Suyash P. Awate. A bayesian neural net to segment images with uncertainty estimates and good calibration. In *Information Processing in Medical Imaging (IPMI)*, 2019. 2

Xiaoqian Jiang, Melanie Osl, Jihoon Kim, and Lucila Ohno-Machado. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association : JAMIA*, 2012. 1

Alain Jungo and Mauricio Reyes. Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, 2019. 2

Alain Jungo, Fabian Balsiger, and Mauricio Reyes. Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. *Frontiers in Neuroscience*, 14, 2020. 2

Benjamin Kompa, Jasper Snoek, and Andrew L. Beam. Second opinion needed: communicating uncertainty in medical machine learning. *npj Digital Medicine*, 2021. 1

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, 2019. 8

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Neural Information Processing Systems (NeurIPS)*, 2012. 1, 3

Jochen Kruppa, Yufeng Liu, et al. Probability estimation with machine learning methods for dichotomous and multicategory outcome: Theory. *Biometrical Journal*, 2014. 4

Meelis Kull and Peter Flach. Novel decompositions of proper scoring rules for classification: Score adjustment as precursor to calibration. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 2015. 4

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Neural*

Information Processing Systems (NeurIPS), 2017. 2

Curtis P. Langlotz et al. A roadmap for foundational research on artificial intelligence in medical imaging: From the 2018 nih/rsna/acr/the academy workshop. *Radiology*, 2019. 2

Agostina J. Larrazabal, Cesar Martinez, José Dolz, and Enzo Ferrante. Maximum entropy on erroneous predictions (meep): Improving model calibration for medical image segmentation. *ArXiv*, abs/2112.12218, 2021. 2

Kai Ma, Siyuan He, Pengcheng Xi, Ashkan Ebadi, Stéphane Tremblay, and Alexander Wong. A trustworthy framework for medical image analysis with deep learning. *arXiv preprint arXiv:2212.02764*, 2022. 1, 2

Takuo Matsubara, Niek Tax, Richard Mudd, and Ido Guy. TCE: A test-based approach to measuring calibration error. In *Uncertainty in Artificial Intelligence*, 2023. 10

Alireza Mehrtash, William M. Wells, Clare M. Tempany, Purang Abolmaesumi, and Tina Kapur. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Transactions on Medical Imaging*, 2020. 2

Xueyan Mei et al. Radimagenet: An open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 2022. 2, 4, 14

Matthias Minderer et al. Revisiting the calibration of modern neural networks. In *Neural Information Processing Systems (NeurIPS)*, 2021. 10

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep neural networks using focal loss. In *Neural Information Processing Systems (NeurIPS)*, 2020. 2

Allan H. Murphy and Robert L. Winkler. Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 1977. 1

Balamurali Murugesan, Bingyuan Liu, Adrian Galdran, Ismail Ben Ayed, and Jose Dolz. Calibrating segmentation networks with margin-based label smoothing. *Medical Image Analysis*, 2023. 2

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *AAAI Conference on Artificial Intelligence*, 2015. 3, 4

Fernando Navarro, Christopher Watanabe, et al. Evaluating the robustness of self-supervised learning in medical imaging. *ArXiv*, 2021. 2

Andrew Y. Ng. Feature selection, l1 vs. l2 regularization, and rotational invariance. In *International Conference on Machine Learning (ICML)*, 2004. 7

Khanh Nguyen and Brendan O'Connor. Posterior calibration and exploratory analysis for natural language processing models. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2015. 2, 4

Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019. 2, 3, 4

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference (BMVC)*, 2018. 1

John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 1999. 2

Joaquin Quiñonero-Candela, Carl Edward Rasmussen, Fabian Sinz, Olivier Bousquet, and Bernhard Schölkopf. Evaluating predictive uncertainty challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, 2006. 4

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. In *Neural Information Processing Systems (NeurIPS)*, 2019. 2, 3, 5

Rahul Rahaman and alexandre thiery. Uncertainty quantification and deep ensembles. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Neural Information Processing Systems (NeurIPS*, pages 20063–20075, 2021. 2

Gregory Scafarto, Nicolas Posocco, and Antoine Bonnefoy. Calibrate to interpret. In Massih-Reza Amini, Stéphane Canu, Asja Fischer, Tias Guns, Petra Kralj Novak, and Grigorios Tsoumakas, editors, *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 2023. 1

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning (ICML)*, 2017. 1

Aditya Singh, Alessandro Bay, Biswa Sengupta, and Andrea Mirabile. On

the dark side of calibration for modern neural networks. In *Workshop on Uncertainty and Robustness in Deep Learning (UDL)*, 2021. 2

Abhishek Singh Sambyal, Narayanan C Krishnan, and Deepti R Bathula. Towards reducing aleatoric uncertainty for medical imaging tasks. In *IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, 2022. 2

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 2014. 7

Skylar E. Stolte, Kyle Volle, Aprinda Indahlastari, Alejandro Albizu, Adam J. Woods, Kevin Brink, Matthew Hale, and Ruogu Fang. Domino: Domain-aware model calibration in medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2022. 2

Atharva Tendle and Mohammad Rashedul Hasan. A study of the generalizability of self-supervised representations. *Machine Learning with Applications*, 2021. 2

Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *Neural Information Processing Systems (NeurIPS)*, 2019. 2, 4

Christian Tomani and Florian Buettner. Towards trustworthy predictions from deep neural networks with fast adversarial calibration. In *AAAI Conference on Artificial Intelligence*, 2019. 1

Hristina Uzunova, Jan Ehrhardt, Timo Kepp, and Heinz Handels. Interpretable explanations of black box classifiers applied on medical images by meaningful perturbations using variational autoencoders. In *Medical Imaging 2019: Image Processing*, 2019. 1

Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. Evaluating model calibration in classification. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019. 4

Bas H.M. van der Velden, Hugo J. Kuijf, Kenneth G.A. Gilhuijs, and Max A. Viergever. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, 2022. 1

Bastiaan S. Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2018. 4

Dongdong Wang, Boqing Gong, and Liqiang Wang. On calibrating semantic segmentation models: Analyses and an algorithm. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 2019. 2

Yang Wen, Leiting Chen, Yu Deng, and Chuan Zhou. Rethinking pre-training on medical imaging. *Journal of Visual Communication and Image Representation*, 2021. 3, 4

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference (BMVC)*, 2016. 4

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018. 2

Luisa M. Zintgraf, Taco S. Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. In *International Conference on Learning Representations (ICLR)*, 2017. 1

5. Supplementary Material

The supplementary material is organized as follows: Section 5.1 contains architectures and hyperparameters Details. Section 5.2 investigates the effect of domain-specific transfer learning using RadImageNet pretraining. Section 5.3 consists of the standard deviation of weights distribution vs. calibration scores analysis and CKA plots of ResNet18 and ResNet50 architectures for Diabetic Retinopathy dataset. Section 5.4 contains performance and calibration plots (Figure 11) and CKA plots (Figure 12) for Histopathology Cancer dataset. Section 5.5 shows the quantitative results of the CKA analysis using mean CKA values. Section 5.6 compares different training regimes with the reconstruction-based self-supervised task.

5.1. Architectures and Hyperparameters Details

5.1.1. Architecture Details

We choose three architectures from the ResNet family to evaluate the performance and calibration using three training regimes, FS_r, FS_p, and SSL_p.

- **WideResNet (*WRN-d-k*:)** It is a variant of residual networks to simulate large architecture size. The depth and width of WideResNet are regulated by a deepening factor d and a widening factor k. We used *WRN-50-2* for our experiments, i.e., WideResNet with 50 convolutional layers and a widening factor of 2.

- **ResNet50 & ResNet18:** We choose the standard architectures to simulate medium and small architecture sizes, respectively.

Table 1: Overview of the models used in this study.

Model Name	Number of Layers	Parameters
ResNet18	18 layers	11M
ResNet50	50 layers	23M
WideResNet	ResNet50, 2×width	66M

5.1.2. Hyperparameter Details

We used the following parameter values for FS_r, FS_p, and SSL_p training regimes across all datasets and architecture sizes for our experiments. We set batch size=16, epochs=300, optimizer=SGD, learning rate=0.001, momentum=0.9, and weight decay=0.0005.

For pretrained setups, FS_p and SSL_p, we trained the classifier and auxiliary module for the first 30 epochs with a learning rate=0.001 and then fine-tuned the complete network with the learning rate=0.00001. In SSL_p training, $\lambda \in \{0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$ is empirically chosen for different datasets and architectures sizes based on the best validation accuracy.

5.2. RadImageNet Pretraining

To investigate the effect of domain-specific transfer learning, we conducted experiments using RadImageNet Mei et al. (2022) – a pretrained neural network (ResNet50) trained only on medical imaging datasets shown in Figure 8. Overall, we notice consistent patterns in calibration, where SSL_p either outperforms or matches FS_p, in line with our observations from other experiments. In this context, we observe that FS_p and SSL_p exhibit comparable performance in (a) and (b). However, in the MCE plot (e), SSL_p demonstrates superior calibration compared to FS_p. For the remaining metrics, SSL_p tends to show marginal improvement or comparable calibration. Taken together, these findings provide additional evidence that SSL_p consistently delivers calibration models on par with, or sometimes even superior to, those produced by FS_p.

Figure 8: Joint evaluation for performance and calibration across different dataset sizes (x-axis) of DR dataset using ResNet50 architecture with RadImageNet pretraining. The shaded region corresponds to $\mu \pm \sigma$, estimated over 3 trials. ↑: higher is better, ↓: lower is better.

5.3. Diabetic Retinopathy Dataset

Figure 9: Standard Deviation of Weights distribution vs. Calibration scores analysis. (a), (b), (c), and (d) depict the relationship between the SD of weights distribution and calibration metrics from the smallest dataset size to the largest one (500, 1000, 1000, 10000), respectively of the DR dataset. Additionally, the corresponding weight distribution plots have been overlaid for convenience of reference. Considering the four plots, we can observe the trend that the calibration metrics of different regimes are segregated when there is a difference in the spread of their distributions (as shown in plots c & d) and overlapping when there is no difference in the SD of weights distribution (as shown in plots a & b). Based on the characteristics of SSL_p (shown in blue), it can be remarked that a balance in the spread of weights is necessary to achieve both good performance and calibration.

Figure 10: CKA plots of trained ResNet18 and ResNet50 architectures using FS_r, FS_p, and SSL_p regimes for DR dataset.

15

5.4. Histopathology Cancer Dataset

Figure 11: Joint evaluation for performance and calibration across different dataset sizes (x-axis) and architectures for Histopathology Cancer dataset. The shaded region corresponds to $\mu \pm \sigma$, estimated over 3 trials. The underline shows the statistical significance between FS_p and SSL_p. Black and Pink color signifies $p < 0.05$ and $0.05 < p < 0.1$ level of significance, respectively.

5.5. Quantitative Comparison CKA

Table 2 presents the quantitative results of the CKA analysis, using mean CKA values. These findings align with the trends observed in Figure 7. In the case of the DR dataset, the mean CKA values of SSL_p rise as the dataset size increases. This supports our previous findings, where the calibration of SSL_p is superior to that of FS_p, and this distinction grows more pronounced as the dataset size becomes larger (Figure 6a). In the context of the Histopathology dataset, previous observations also indicated that SSL_p outperforms FS_p in terms of calibration, although the difference in calibration metrics values' magnitude is less (6b). Consequently, we notice that there is no significant difference in the mean CKA values between the two training approaches indicating the representations learned are quite similar.

Table 2: Mean CKA values of different training regimes across varying architectures, datasets and their sizes.

Architecture	Training Regime	Diabetic Retinopathy				Histopathology Cancer			
		500	1000	5000	10000	500	1000	5000	10000
ResNet18	FS_p	0.86	0.84	0.85	0.85	0.76	0.76	0.75	0.75
	SSL_p	0.85	0.85	0.88	0.88	0.76	0.75	0.74	0.75
ResNet50	FS_p	0.84	0.84	0.84	0.84	0.74	0.75	0.74	0.73
	SSL_p	0.85	0.85	0.86	0.87	0.75	0.74	0.74	0.73
WideResNet	FS_p	0.84	0.83	0.81	0.81	0.70	0.69	0.69	0.71
	SSL_p	0.84	0.85	0.86	0.87	0.69	0.70	0.69	0.71

Figure 12: CKA plots of trained architectures using different regimes for Histopathology Cancer dataset.

5.6. Comparison of Fully-Supervised and Reconstruction-Based Self-Supervised Task

Figure 13: Comparison of fully supervised (FS_r, random initialization), fully supervised (FS_p, pretraining), and reconstruction-based auxiliary SSL task (SSL_p, pretraining) on DR dataset. Notably, the calibration of models achieved through the auxiliary task does not precisely align with that of the rotation task. Remarkably, the plots reveal a notable contrast: very low OE (f) but high ECE (c). This discrepancy could hint at potential underconfidence, stemming from substantial regularization induced by the reconstruction-based auxiliary SSL task. However, drawing definitive conclusions is premature, as further experiments, encompassing various architectures and hyperparameter tuning, are necessary. Relying solely on the plots, we abstain from making a judgment regarding the superiority of either FS_p or reconstruction-based SSL_p.