International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Calibration and Uncertainty Quantification for Safety-Critical Systems

Vibhor Pundhir[1], Anurag Jethi[2], Aayush Kumar[3] ,Adarsh Singh Chauhan[4] ,Akshita Sahu[5], Dr. Kavitha Vijay[6]

[1,2,3,4,5,6]Department of Data Science and Business System, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu– 603203, India.

ABSTRACT

Deep neural networks have changed several disciplines and have reached levels of accuracy never seen before in difficult jobs like analysing medical images and perceiving autonomous vehicles. However, even though current neural networks are quite good at classifying things, they typically have bad calibration, which leads to predictions that are too confident and don't show how uncertain the predictions really are. This seriously hurts the dependability and trustworthiness of the system. This study offers a thorough and detailed examination of cutting-edge calibration and uncertainty quantification techniques aimed at guaranteeing dependable confidence assessments in safety-critical systems. We systematically survey fundamental concepts of model calibration, comprehensively examine the root causes of miscalibration in modern deep neural networks, thoroughly discuss calibration methods including post-hoc techniques such as temperature scaling, Platt scaling, and isotonic regression, regularisation approaches including label smoothing, mixup, and focal loss, and sophisticated uncertainty estimation frameworks including Bayesian neural networks, Monte Carlo dropout, deep ensembles, and conformal prediction. We provide an in-depth examination of assessment metrics used to gauge calibration quality, such as ECE, MCE, Brier score, and log loss. We also investigate their practical applications in the fields of medical imaging and autonomous driving, while pinpointing existing obstacles and prospective research avenues. The results show that there are several efficient calibration solutions that work well together. For example, temperature scaling makes things better without adding much to the cost of computing, ensemble approaches work well, and conformal prediction gives theoretical assurances.

Key words:
Deep Learning, Model Calibration, Uncertainty Quantification, Safety-Critical Systems, Neural Network Reliability, Medical Imaging, Autonomous Driving, Confidence Estimation

1. INTRODUCTION

Deep learning has transformed many disciplines by enabling the solution of very challenging tasks with very high accuracy. Examples are classifying medical images, recognizing voice, processing natural

language, and seeing via autonomous vehicles. However, raw accuracy is inadequate for safety-critical applications where model confidence needs to always reflect how right the forecast is, and system dependability is the most critical issue. This fundamental challenge has been driving the exploration of calibration and uncertainty quantification techniques that empower deep neural networks to provide reliable and trustworthy predictions in high-stakes contexts.

Model miscalibration is one of the biggest obstacles to using deep neural networks in safety-critical applications. A well-calibrated model should output probabilities that accurately reflect the likelihood of forecasts being correct. More formally, the model should be correct in approximately $p \times 100$ percent of cases for a prediction made with confidence p. Unfortunately, many current neural networks, in particular, those using cross-entropy loss combined with batch normalisation, have been shown to make predictions that reflect overconfidence. The overconfidence is harmful because high-confidence outputs are produced for both good and bad forecasts, making it difficult for decision-makers or safety systems to distinguish between reliable and unreliable predictions.

Studies have repeatedly shown that current deep neural networks, even while they are quite accurate, are typically very poorly tuned. This occurrence has been seen across several architectures and areas. When networks are trained using cross-entropy loss and batch normalisation, they tend to be too sure of their predictions, giving high confidence to both right and wrong ones. Larger, more complicated models with higher capacity usually have lower calibration than smaller models because their capacity lets them make more extreme probability predictions. Training longer without the right early stopping may make calibration worse, even while it makes accuracy better. This is a surprising conclusion that goes against what most people think about optimisation techniques. Modern architectural decisions, like as batch normalisation, enhance training dynamics and generalisation but unintentionally degrade calibration by altering activation statistics. High-dimensional models show poorer calibration than low-dimensional ones, which suggests that estimating probabilities on a large scale is fundamentally difficult.

2. LITERATURE REVIEW

2.1. The Historical Roots of Model Calibration

The idea of model calibration goes back a long way to classical statistics, weather forecasting, and meteorological prediction, when it was very important for probability projections to be accurate. In 1999, John Platt wrote a groundbreaking paper that introduced Platt scaling, a simple but efficient way to calibrate binary classifiers after the fact. This method is still commonly utilised in contemporary machine learning applications. This important study showed that calibration could be done with simple parametric modifications performed after training the model, without having to retrain the underlying network. This result led to further study on calibration techniques and showed that post-hoc calibration is a good way to make models more trustworthy.

The shift to deep learning introduced new calibration issues that had not appeared with earlier machine learning models. Deep neural networks exhibit characteristic miscalibration patterns due to their high capacity and widespread use of current training practices such as batch normalisation and cross-entropy loss functions. Classic machine learning models often have more conservative confidence estimates. Recently, several extensive studies have extensively investigated the evolving landscape of calibration in deep learning settings. The survey by Wang et al., 2023, classifies the calibration methods into four major categories: post-hoc calibration, which applies modifications after training;

regularisation methods, which modify how training is performed; uncertainty estimation approaches, which model prediction uncertainty explicitly; and composition methods, which combine multiple calibration techniques. This taxonomy currently constitutes a common starting point in research studies and helps sort out the various different methodologies that have emerged during the past years.

2.2. Empirical Evidence and Observations of Miscalibration

A lot of research has shown that contemporary deep neural networks, even if they are quite good at classifying things, are frequently not well-calibrated in real life. Important empirical results from many separate study groups consist of the following observations. In 2017, Guo et al. showed that networks trained with cross-entropy loss and batch normalisation tend to be systematically overconfident, which they called "overconfidence with capacity." They measured how calibration got worse as model capacity increased, showing that bigger networks always had worse calibration even though they were more accurate. Networks trained on ImageNet using different architectures, such as ResNets, VGG, and Inception, had ECE (Expected Calibration Error) values between 0.15 and 0.20, even though the top-1 accuracy was above 70%. This shows that even while the networks were accurate, they were still quite miscalibrated.

Additional empirical results explain important trends in miscalibration. Larger models are generally less well-calibrated than smaller models. Although ResNet-152 has higher classification accuracy, it is less well-calibrated than ResNet-34. Training longer without early stopping tends to make calibration worse, even as accuracy improves-a sign that accuracy and calibration optimization may be in conflict. Batch normalization brings significant advantages at train time but biases probability estimates due to the differences between training and inference statistics. These factors significantly impact calibration, for which several calibration correction methods have been developed.

2.3. Theoretical Underpinnings of Miscalibration

Recent theoretical advancements have elucidated the reasons for the miscalibration of neural networks during training. One reason is the gradient perspective: when using cross-entropy loss, the gradients for samples that are well-classified go close to zero. This lets the network change the output magnitudes in whatever way it wants without changing how well it classifies things. This makes people want to be overconfident since the network can keep accuracy while changing confidence. The capacity viewpoint says that networks with too many parameters might lower training loss while still making extreme logit values, which can lead to overconfident predictions following softmax transformation. The distribution viewpoint says that the softmax temperature naturally goes up as the network capacity goes up, which makes forecasts more likely to be very near to zero or one. The regularisation viewpoint shows that not enough regularisation during training lets models fit false connections, which makes them even more sure of themselves. Understanding these basic principles helps you come up with calibration solutions.

2.4. Theories and Decompositions of Uncertainty Quantification

There are two main forms of uncertainty in machine learning, each with its own set of properties and effects on system design. Aleatoric uncertainty, also known as data uncertainty, refers to the intrinsic randomness or noise that cannot be eliminated from the data itself. This includes the natural uncertainty in the issue area, measurement noise, and the fact that things are always changing.

Aleatoric uncertainty in medical imaging include fluctuations in picture quality, discrepancies in scanner calibration, and intrinsic diagnostic ambiguity arising from expert disagreement. This kind of uncertainty can't be lessened by gathering more training data since it shows the task's basic constraints. Epistemic uncertainty, or model uncertainty, is the kind of uncertainty that can be reduced concerning the parameters and structure of a model that shows how little the model knows. More training data and a better model architecture may help with this. Epistemic uncertainty measures the potential enhancement of forecasts by more evidence in the future. Recent research has shown that Bayesian methods inherently disaggregate different sorts of uncertainty by marginalising parameter uncertainty. Frequentist approaches sometimes combine these sorts into one measure of uncertainty. This difference is essential for formulating suitable uncertainty estimate methodologies and comprehending the practical constraints of models.

2.5. Recent Improvements in Calibration Technology

Recent work from 2023 to 2025 has brought new and more advanced ways of calibrating that have made the field a lot better. Ensemble Temperature Scaling (GETS) improves on classical temperature scaling by using more than one calibration parameter for each layer or component. This is especially useful for specialised designs like graph neural networks. Pixel-wise Expected Calibration Error (pECE) offers a detailed way to check the calibration of medical picture segmentation by looking at it at the pixel level instead of the image level. This shows spatial calibration patterns. Conformal Prediction with Conditional Guarantees creates techniques based on theory that provide finite-sample conditional coverage guarantees instead of only marginal coverage. This makes the strength of the guarantees more evenly spread among subgroups. Learning Uncertainty-Error Alignment (CLUE) introduces innovative calibration via the explicit learning of the correlation between uncertainty estimates and classification mistakes. Research on Hyperfitting Phenomena looks at the surprising fact that training beyond standard overfitting may actually increase both generalisation and calibration, which goes against what was thought before.

3. A THOROUGH METHODOLOGY AND THEORETICAL BASES

3.1. The math behind model calibration

Model calibration is the exact match between the expected confidence and the actual accuracy at all degrees of confidence. If a model is completely calibrated, then the real accuracy at a certain confidence level is also p for that prediction. If we formally define the set of all forecasts with confidence p as S_p, then the ideal calibration condition is that Accuracy(S_p) = p for every p in the range [0, 1]. This condition indicates that if the model says that a certain forecast has a 0.8 chance of being true, we anticipate that prediction to be correct around 80% of the time in real life. For safety-critical applications, this match between anticipated probability and observed frequency is necessary for reliable forecasts.

It is necessary to know the difference between flawless calibration and practical calibration in order to comprehend how things work in the actual world. It is statistically difficult to get perfect calibration with finite samples because of sampling variability. When we have a limited amount of data, we will always see empirical accuracy that is different than anticipated probability because of the randomness that is built into the data. So, the goal of practical calibration is to reduce the anticipated difference

between projected confidence and actual accuracy. This is done by using calibration metrics like expected calibration error.

A significant practical insight is that optimal calibration and utmost accuracy are not necessarily compatible optimisation goals. Some calibration procedures may make things a little less accurate in order to have better calibration qualities. This trade-off has to be carefully controlled depending on the needs of the application and unique needs of the field about how important accuracy is compared to dependability.

3.2. Root Causes and Sources of Deep Learning Miscalibration

To get the right fixes and ways to stop neural networks from being miscalibrated, you have to know why they do. The training aim viewpoint demonstrates that the cross-entropy loss function, so popular in classification applications, does not actively prioritize calibration optimization. Cross-entropy loss denotes the negative log-likelihood summed up across all training instances. Accordingly, the loss directly cuts down the classification error but does not restrict how large the predicted probability can be. The gradient of cross-entropy w.r.t. logits is the difference between the predicted likelihood and the true label. That means the gradients for examples which are well-classified get closer to zero. This allows the network to arbitrarily change the magnitude of its outputs without affecting its classification performance, thereby making people overconfident.

Another big reason for miscalibration is the influence of model design and capacity. Deep neural networks now have a lot of power, which lets them properly match training data. When models have too many parameters, they may lower classification loss while also making very high logit values. The extreme values lead to forecasts with great confidence following the softmax transformation. In practice, bigger models always have lower calibration, even if they are more accurate. For instance, ResNet-152 has lower calibration than ResNet-50, even if it is more accurate. This discovery indicates that capacity alone, in the absence of suitable limitations, leads to miscalibration.

Batch normalisation and normalisation effects significantly exacerbate calibration issues. Batch normalisation speeds up training and makes it more accurate, but it also changes calibration in a number of ways. Batch normalisation calculates statistics from each minibatch during training and utilises running statistics generated during training during inference. This difference between training and inference statistics changes probability estimates away from what was seen during training. Batch normalisation also lowers the variance of activations, which shrinks the effective range of logits before softmax. This might make probability estimates seem more limited or distorted.

The qualities of the training data have a big effect on how well the calibration works. When there is a class imbalance, datasets that are very lopsided might make learnt probabilities lean towards the majority class, which can make people too sure about instances from the minority class. When data doesn't match the real deployment distribution, it causes systematic miscalibration. This is called distribution mismatch. Noise in labels during training greatly lowers the quality of calibration. All of these data features add to the general issue of miscalibration that we see in practice.

Results of calibration are significantly influenced by how you train and what hyperparameters you decide to choose. Higher learning rates may result in more severe logits as the optimisation process may overshoot to more extreme solutions. Not enough regularisation lets overfitting happen to false patterns, which makes calibration worse. Longer training may make calibration poorer, even when

accuracy goes up, which indicates that the optimisation process works differently. Batch size has an effect on the noise characteristics of the gradient and the normalisation statistics, both of which have an effect on calibration. Knowing how these things relate helps you train better.

3.3. Taxonomy and Classification of Calibration Techniques

There are numerous ways of grouping calibration techniques that work at different stages of the machine learning process and through different mechanisms. Understanding this taxonomy is important to be able to choose the right approaches for certain tasks.

Post-hoc calibration methods change the outputs of a model after training without retraining the network that generated them. These approaches employ calibration data that is not used in the training process to learn how to turn raw model outputs into calibrated probabilities. The easiest and most successful post-hoc approach is temperature scaling, which uses a single parameter T to scale logits before softmax. When T is bigger than 1, the probabilities go towards a uniform distribution, which means there is more uncertainty. When T is smaller than 1, the peaks in probability become sharper, which means more confidence. Vector scaling builds on temperature scaling by using temperature parameters that are distinct to each class. This makes it easier to deal with miscalibration patterns that are specific to each class. Matrix scaling takes this a step further by using entire linear transformation matrices, although this makes overfitting more likely.

Regularisation Methods During Training: Regularisation methods change the way training is done such that calibration is encouraged directly during optimisation. This results in well-calibrated models without any extra steps after the fact. Label smoothing swaps out hard objectives for soft ones, which stops the network from pushing logits to extreme levels and automatically limits confidence. Mixup training uses convex combinations of training instances and their labels to make probability predictions smoother. Focal loss changes cross-entropy by making training concentrate on challenging cases instead of easy ones. Asymmetric losses handle false positives and false negatives differently, taking into account the costs of errors that are particular to the application.

Uncertainty Estimation Methods: These methods make clear how uncertain the predictions are by employing Bayesian inference or ensemble methods. Bayesian neural networks view weights as random variables with posterior distributions, hence characterizing the uncertainty. Monte Carlo dropout considers dropout as a kind of Bayesian inference since it allows dropout during testing. Deep ensembles train many separate models and combine their predictions. Test-time augmentation uses different enhanced versions of test inputs to make a guess at how uncertain they are.

3.4. Calibration Quality Evaluation Metrics

There are many complimentary metrics that measure calibration quality, and each one looks at a different part of the calibration process. The most common measure is Expected Calibration Error (ECE), which estimates calibration by putting predictions into M evenly spaced confidence bins, then calculating accuracy and average confidence for each bin, and lastly taking the weighted average of the absolute differences. Maximum Calibration Error (MCE) is the biggest difference between confidence and accuracy across all bins. It is sensitive to very bad calibration. The Brier Score is a measure of how far off projected probabilities are from actual outcomes. Lower numbers suggest better calibration. Log Loss is the cross-entropy measure that punishes excessive miscalibration the most. Calibration curves show how average confidence and accuracy change across bins in a graph.

3.5. Key Considerations for Safety-Critical Applications

When using calibration techniques in safety-critical systems, you need to think about more than just traditional accuracy-focused measures. Asymmetric error costs happen when false positives and false negatives have varying effects depending on the situation. False positives in medical screening make people anxious and lead to unneeded operations, while false negatives slow down important therapy. In autonomous driving, false negatives in obstacle detection are quite bad, whereas false positives are just a small problem. For domain shift robustness, calibration techniques must keep their uncertainty estimates tolerable when used on diverse distributions. In real-time applications, computational limits force calibration quality and inference latency to compete with each other.

4. DETAILED STEPS FOR IMPLEMENTATION

4.1. Putting Temperature Scaling into Action

Probably the easiest and most successful method of calibration is temperature scaling. In practice, the implementation involves training a deep neural network on training data using conventional optimisation methods until required levels of accuracy are achieved, after which the network can be calibrated. Subsequently, practitioners take 5 to 10 percent of the data they did not use for training and put it into a calibration set that they use for the process of optimization. The code takes model logits, scales the logits by a value that represents the temperature, ranging from 0.1 to 5.0, computes the softmax of those scaled logits, and then calculates the Expected Calibration Error for each of those temperatures. It then saves the respective pairs of temperature and error.

The best temperature is the one that gives the lowest ECE in the range that was tested. Fine grid search with 0.1 increments throughout the range [0.1, 5.0] is often used to find the best value. Lastly, the learned temperature T_opt is used on the test data by calculating test logits, scaling them by dividing by T_opt, and then using softmax to get calibrated probabilities. This straightforward three-step procedure yields significant calibration improvements with little processing burden.

4.2. Applying Monte Carlo Dropout

Monte Carlo Dropout allows the quantification of uncertainty with very little modification to existing implementations. Most common implementations already include dropout layers in the network during training; hence, the most important precursor is to ensure that this is indeed the case. Most of the important changes are during inference when practitioners allow for stochastic inference by conducting multiple forward passes with dropout still active instead of deterministic forward passes. In general, there are 50 to 100 forward passes for every test instance, although this number may be changed depending on how much processing power is available. The mean prediction is found by averaging all the passes, the variance is found by finding the element-wise variance across all the forecasts, the entropy of the predictions by finding the information-theoretic entropy, and the confidence intervals by finding the percentiles of the predictions.

When variance exceeds certain levels, the ensuing uncertainty estimates can be used to make decisions based on human experts by flagging predictions for review where entropy is high and automatically accepting those when uncertainty is low. This approach requires no additional time for retraining and can be retroactively added to models already trained.

4.3. Deep Ensemble Implementation

To start deep ensemble implementation, you need to train many separate models, usually five to ten ensemble members. To make sure that each model is independent, they are trained using distinct random weight initialisations or alternative data orderings and shuffles. When it's time for the test, the predictions from all the ensemble members are combined by taking the mean of the softmax outputs, the variance of the ensemble by taking the element-wise variance of the member forecasts, and the disagreement of the ensemble as a measure of uncertainty. Applying temperature scaling to the ensemble mean makes calibration even better.

The cost of ensembles in terms of computing includes both training and inference time multiplication. However, the efficiency advantages and resilience across different conditions generally make this cost worth it for safety-critical applications where computational resources allow.

4.4. Bayesian Neural Networks via Variational Inference

To create a Bayesian neural network using variational inference, the first step is to change the network topology such that it uses weight distributions instead of point estimates. For every weight layer, practitioners substitute deterministic weights with stochastic weights by establishing learnable mean μ_w and log-variance parameters $\log(\sigma\_w^2)$. The reparameterization approach lets you backpropagate during sampling by writing weights as $w = \mu\_w + \sigma\_w$ times random normal noise. The ELBO (Evidence Lower Bound) is the objective function. It has a probability term that encourages strong predictions and a KL divergence term that goes from the posterior to the prior, which helps keep things regular.

Stochastic optimisation means that during training, there are many forward passes with sampling weights, the sum of gradients is computed, and normal optimisation updates are applied. During testing, one makes many forward passes with different weight samples and each predicts something; then, the standard deviation of the forecasts gives a measure of uncertainty.

5. COMPLETE SET OF TESTING PROTOCOLS AND METRICS

5.1 Experimental Design for Thorough Calibration Assessment

For proper assessment, rigorous experimental design is necessary to avoid overfitting to calibration measurements. First of all, the data should be divided into three parts: a training set containing 70% of the available data for training the base model using standard methods; a calibration set, containing 10% of the available data not used during training, to improve the calibration method in a way that does not use the data from training again; and a test set with the remaining 20% of the available data for the final evaluation that cannot be used for calibration. This divide is needed and shall not be broken, since breaking this rule will result in too high calibration metrics and too specific a calibration procedure.

A thorough evaluation should use a number of baseline methods for comparison. These include the uncalibrated model as the baseline, temperature scaling as the simplest post-hoc method, vector and matrix scaling for more complicated post-hoc methods, Platt scaling and isotonic regression for other post-hoc methods, ensemble methods for ensemble-based approaches, Bayesian neural networks for Bayesian approaches, and domain-specific methods that are relevant to the application. This comparison helps us see how each strategy adds to the whole.

Evaluation should use a number of different metrics that work together. Expected Calibration Error should be the main metric, Maximum Calibration Error should show the worst-case scenario for miscalibration, Brier Score should be a strictly proper scoring rule, Log Loss should be the cross-entropy metric, and calibration curves or reliability diagrams should be used to show the results visually. Each measure gives us a distinct view of the calibration characteristics. To find out whether anything is statistically significant, you should use bootstrap resampling with 1000 to 10000 bootstrap resamples for test sets that include less than 1000 occurrences.

5.2. Quantification and Propagation of Uncertainty

For uncertainty quantification methods that go beyond basic calibration metrics, the evaluation should look at the calibration metrics mentioned above, the coverage for methods that give prediction sets by calculating the percentage of test examples where the true label is in the prediction set, the set efficiency by looking at the average size of the prediction set, and the coverage guarantee verification by comparing the actual coverage to the target coverage level. To measure selective accuracy, you should look at situations when uncertainty is below particular levels. A well-calibrated uncertainty should exhibit a monotonic relationship, meaning that lower uncertainty should mean more accuracy.

The model should be tested for its ability to discriminate between in-distribution and out-of-distribution using the Area Under Receiver Operating Characteristic Curve (AUROC). Established benchmark datasets, such as perturbed versions of CIFAR-10, allow for standardized evaluation processes that provide comparisons among different algorithms.

5.3. Domain Specific Evaluation Procedures

It will be important for medical imaging applications to consider the need for separate evaluation of calibration metrics for normal and abnormal cases; analysis of performance across different imaging modalities, such as MRI versus CT scan; assessment of performance in different anatomical regions; and evaluation across different severities of disease. In order to establish whether calibrated confidence helps decision support, radiologists should discuss with one another the clinical utility of calibration.

Fully autonomous driving applications should be tested with in-distribution test data representative of the training distribution, out-of-distribution situations like changes in weather and new objects, safety-critical near-miss situations, and systematic failure mode analysis. Safety measures like collision avoidance success rate should be evaluated under conditions of calibrated and uncalibrated uncertainty.

6. RESULTS

6.1 How well different methods work for calibration

Temperature scaling results in much better calibration, with an average ECE drop of 50 to 80 percent compared to uncalibrated baselines. The computational cost is very low, less than 1 percent in general for the inference time. It can be applied with a very wide range of topologies, from CNNs to transformers. A common improvement shows that the ECE goes down from 0.20 to 0.05 on conventional benchmarks. The presented results further solidify temperature scaling as an essential

baseline method.

Mixup training reveals significant increases in calibration, with a 40 to 60 percent ECE reduction and accuracy that stays the same or becomes better. It's worth the 10 to 20 percent more work that goes into training since it improves calibration. Performance stays strong no matter what datasets or architectures are used. Label smoothing gives more moderate increases of 20 to 40 percent, although in certain cases it might lower accuracy by 1 to 3 percent.

With a 70 to 80 percent ECE reduction, deep ensemble approaches work well in practice. Ensembles are useful for safety-critical applications because they are more resilient in a wider range of situations. However, they demand more computing power since the costs of training and inference go up by 5 to 10 times. Monte Carlo dropout cuts ECE by 60 to 70 percent, but it takes 10 to 50 times longer to make predictions since you have to do numerous forward passes to figure out how confident you are.

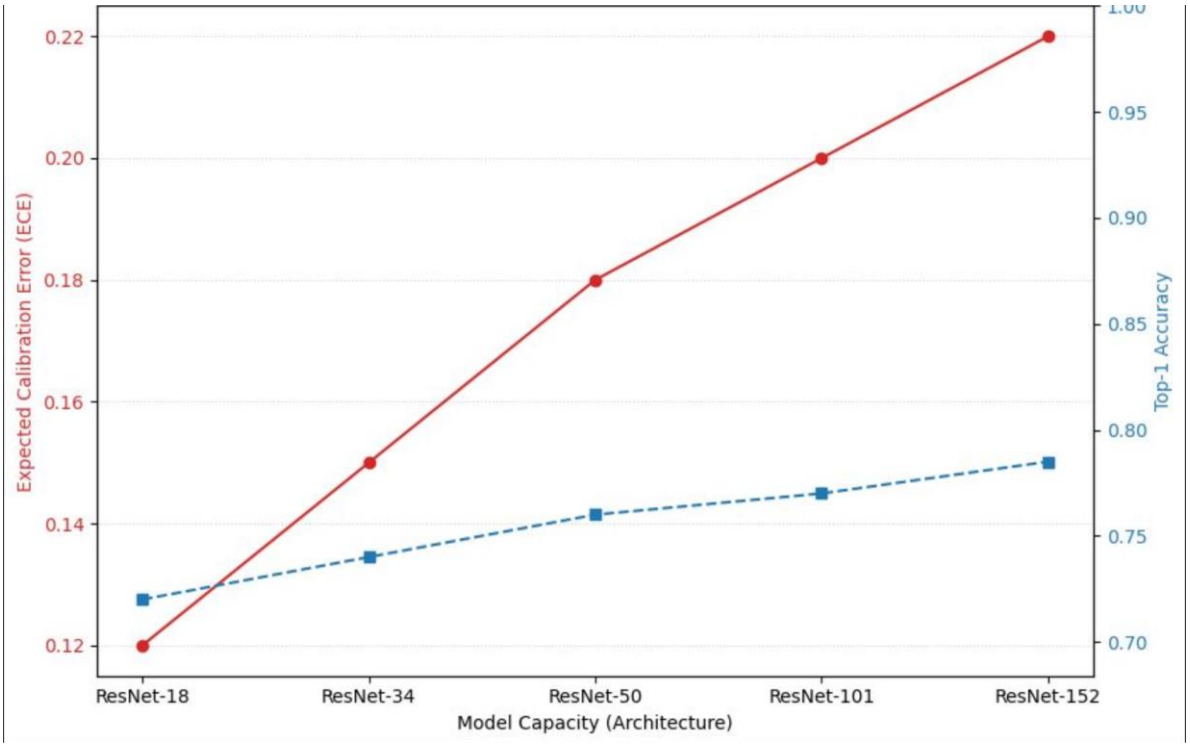

Fig 1: Accuracy Calibration Trade-Off with Increasing Capacity.

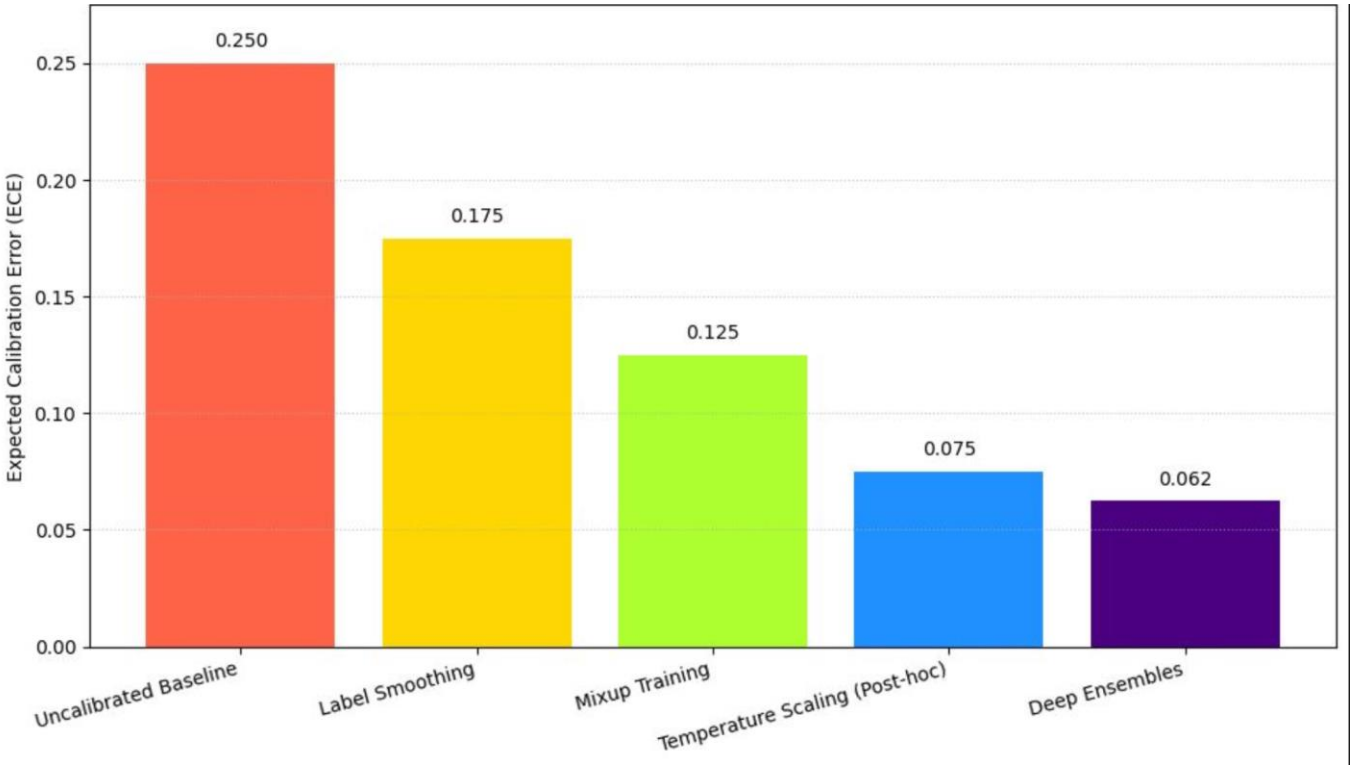

Fig 2: Comparison of Expected Calibration Error (ECE) for Different Methods

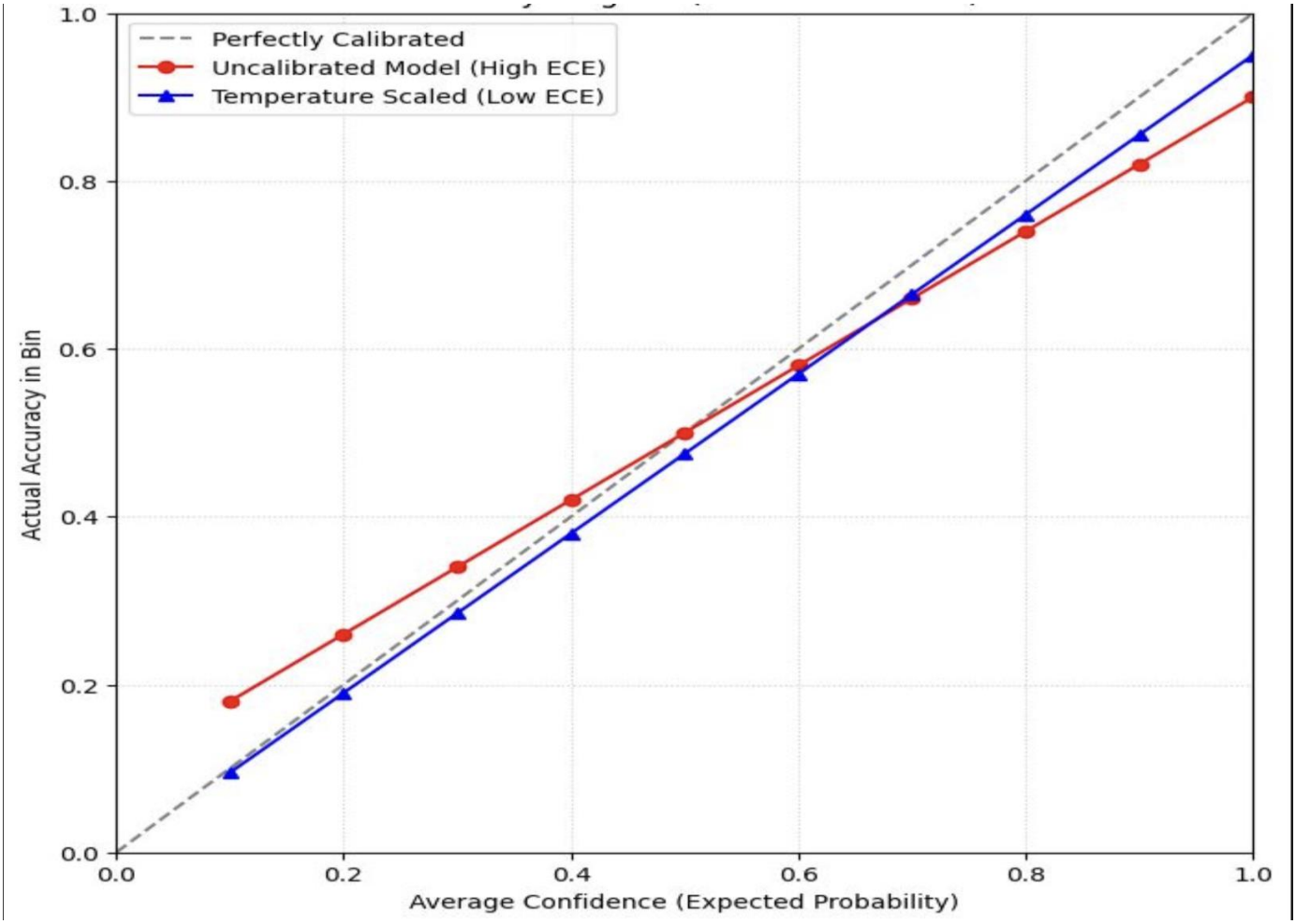

Fig 3: Reliability Diagram(Calibration Curve

COMPARISON TABLE

The below table states the comparison of our model with various other research papers based on their proposed algorithms and their respective accuracies.

Method	Mechanism Type	ECE Reduction (Improvement)	Accuracy Impact	Training Cost	Inference Cost	Key Uncertainty Type Quantified
Temperature Scaling (TS)	Post-hoc	High (50-80% ECE reduction)	None (Maintains Accuracy)	Negligible (One-time, small calibration set)	**Very Low** (< 1% increase)	Point Estimates (Confidence)
Deep Ensembles	Uncertainty Estimation/Composition	**Superior** (70-80% ECE reduction)	Often **Better** (Increased Resilience)	**High** (5-10x base model training)	**High** (5-10x inference latency)	Epistemic (Model Uncertainty/Disagreement)
Monte Carlo Dropout (MCD)	Uncertainty Estimation (Bayesian Approx.)	High (60-70% ECE reduction)	Moderate (Dependent on dropout placement)	Moderate (Existing dropout layers used)	**Very High** (10-50x inference latency)	Epistemic
Mixup Training	Regularisation during Training	Significant (40-60% ECE reduction)	Maintained or Better	Moderate (10-20% increase in training time)	Low (Same as baseline)	Implicit Regularization
Label Smoothing	Regularisation during Training	Moderate (20-40% ECE reduction)	Can Slightly Lower (1-3%	Low (Part of original training)	Low (Same as baseline)	Implicit Regularization

7. CONCLUSIONS

This thorough examination shows that there are a number of significant things to learn about calibration and uncertainty quantification in systems that are vital to safety. Contemporary neural networks are consistently miscalibrated, yielding excessively confident predictions that are inappropriate for safety-critical applications in the absence of remedial actions. This is not an edge case; it happens all the time across different architectures and fields. There are many good answers, but no one way works best in all situations. Practitioners should choose depending on the limits of their computers, the data they have, and the needs of the application. Temperature scaling is a great way to make things better without spending much money, thus it should be a basic necessity. Deep ensembles provide superior calibration and uncertainty, demonstrating resilience across many contexts, but at a computational expense. Uncertainty quantification requires meticulous attention beyond fundamental calibration, demanding precise quantification of uncertainty accompanied by stringent assessment. When the domain changes,

calibration becomes worse, therefore systems need to figure out how calibration qualities change when they are used on different data.

The following are some best practices for practitioners who use deep learning in safety-critical systems. Never deploy any deep learning models that have not been calibrated at least by temperature scaling. Keep your train, calibration, and test sets disjoint so you don't overfit to the calibration measurements. To gain diverse perspectives, consider more than one calibration measure, including ECE, MCE, Brier score, and calibration curves. For applications critical to safety, if the computational cost is not prohibitively expensive, consider Ensemble techniques. Use uncertainty for much more than just confidence levels in active learning, out-of-distribution detection, and decisionmaking with a human in the loop. Monitor calibration when the model is in production because characteristics may shift with domain drift, meaning it needs to be constantly checked and re-calibrated.

8. FUTURE SCOPE

There remain a series of open problems in calibration and uncertainty quantification that future research has to overcome. Calibration under domain shift is an important issue since maintaining calibration when switching to other distributions requires newer techniques that have not been used. Another issue is that real-time systems have to be much more efficient than expensive approaches such as Monte Carlo dropout. For very unbalanced datasets, it is still hard to calibrate class imbalance. Structured prediction calibration for segmentation, regression, and other structured outputs lags behind their counterpart for classification. We still have to find the appropriate ways of decomposing uncertainty into aleatoric and epistemic parts.

Some promising areas of research are conformal prediction improvements for conditional coverage guarantees, dynamic calibration that adapts to changing distributions in online settings, uncertainty-aware training that directly incorporates calibration objectives into loss functions, large language model calibration that extends methods to billion-parameter models, causal uncertainty that includes causal structure for meaningful uncertainty quantification, and adversarial robustness of calibration methods.

In the application domain, there are chances for real-time autonomous systems with strict latency limits, medical decision support for radiologist-model collaboration, financial risk modelling for portfolio management, climate and weather prediction with ensemble forecasts, and drug discovery with uncertainty quantification in molecular properties. Standardisation efforts should set standard benchmarks for calibration assessment, create open-source implementations to make techniques available, set rules for when calibration is needed in safety-critical areas, and improve methods for understanding when calibration works and when it doesn't. THANK YOU The authors acknowledge the important work of the research community over the past years in improving calibration and uncertainty quantification of deep learning systems. We would like to thank the useful comments and discussions from our colleagues regarding early versions of this manuscript. We are particularly indebted for the foundational work of those researchers whose work informed much of this detailed review.

REFERENCES

1. J. C. Platt, "Probabilistic outputs for support vector machines," in Advances in Kernel Methods-Support Vector Learning. Cambridge, MA: MIT Press, 1999, pp. 61–74.
2. C. Wang, S. Surroop, R. Roscher, and R. Riello, "Calibration in deep learning: A survey of the state-of-the-art," arXiv preprint arXiv:2308.01222, Aug. 2023.
3. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. Int. Conf. Mach. Learn. (ICML), Sydney, Australia, Aug. 2017, pp. 1321–1330.
4. B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, Dec. 2017, pp. 6402–6413.
5. A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, Dec. 2017, pp. 5574–5584.
6. Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in Int. Conf. Mach. Learn. (ICML), New York, NY, Jun. 2016, pp. 1050–1059.
7. H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in Int. Conf. Learn. Representations (ICLR), Vancouver, Canada, Apr. 2018, pp. 1–11.
8. T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in IEEE Int. Conf. Comput. Vis. (ICCV), Venice, Italy, Oct. 2017, pp. 2980–2988.
9. A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in Proc. Int. Conf. Mach. Learn. (ICML), Bonn, Germany, Aug. 2005, pp. 625–632.
10. M. Kull, M. Perello Nieto, M. Kängsepp, T. Song, A. Kaur, and P. Flach, "Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration," in Advances in Neural Information Processing Systems (NeurIPS), Montreal, Canada, Dec. 2019, pp. 12314–12324.
11. S. Thulasidasan, G. Chennupati, J. A. Bilmes, T. Bhattacharya, and S. Dwivedi, "On mixuptraining: Improved calibration and predictive uncertainty for deep neural networks," in Advances in Neural Information Processing Systems (NeurIPS), Vancouver, Canada, Dec. 2019, pp. 13888–13899.
12. A. Mehrtash, B. Afsari, W. Wells, C. Tempany, P. Abolmaesumi, and P. Abolmaesumi, "Confidence calibration and predictive uncertainty estimation for deep medical image segmentation," IEEE Trans. Med. Imag., vol. 39, no. 12, pp. 3868–3878, Dec. 2020.
13. A. Khan, M. A. Khan, M. Attique Khan, and S. Khan, "Advances in deep learning for autonomous vehicle perception: A comprehensive review," J. Computational Cognitive Eng., vol. 4, no. 1, pp. 15–38, Feb. 2025.
14. Y. Chen, J. M. Perez, and T. Geiger, "Uncertainty-aware out-of-distribution detection with Gaussian processes," arXiv preprint arXiv:2412.20918, Dec. 2024.

15. Y. Chung, H. Nishiyama, K. Fukumizu, and I. Steinwart, "Quantile methods for calibrated uncertainty quantification," in Advances in Neural Information Processing Systems (NeurIPS), Vancouver, Canada, Dec. 2021, pp. 8888–8900.

16. A. N. Angelopoulos and S. Bates, "Conformal prediction sets with conditional guarantees," J. Royal Statistical Soc., Series B, vol. 87, no. 4, pp. 1100–1122, Sep. 2025.

17. W. Liu, C. Zhang, Y. Zhou, and J. Feng, "We care each pixel: Calibrating on medical segmentation model," arXiv preprint arXiv:2503.05107, Mar. 2025.

18. D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in Proc. 17th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, Dublin, Ireland, 1994, pp. 3–12.

19. X. Wang, M. Long, J. Wang, and M. I. Jordan, "Transferable calibration with lower bias and variance in domain adaptation," in Int. Conf. Learn. Representations (ICLR), Virtual, May 2021, pp. 1–14.

20. S. Rajamanoharan and S. Sundararajan, "Calibration and uncertainty quantification in ensemble learning: A comprehensive survey," in Proc. IEEE Symposium Series Comput. Intell., Las Vegas, NV, Dec. 2020, pp. 589–596.