https://doi.org/10.64142/jeai.1.2.22

Journal of Engineering and Artificial Intelligence

Research Article

JEAI-25-22

Systematic Evaluation of Label Noise Effects on Accuracy and Calibration in Deep Neural Networks

Christopher Boseak*

Independent Research, USA

***Corresponding author:** Christopher Boseak, Independent Research, USA, E-mail: cboseak@gmail.com

Received date: July 24, 2025; **Accepted date:** July 29, 2025; **Published date:** August 24, 2025

Citation: Boseak C (2025) Systematic Evaluation of Label Noise Effects on Accuracy and Calibration in Deep Neural Networks. J Eng Artif Intell Vol.1 No.2: 22.

Abstract

Label noise is a pervasive issue in real-world datasets that can de- grade both the accuracy and calibration of deep neural networks. In this study, we systematically examine how symmetric (random) and asymmetric (class-dependent) label noise influence model accuracy and confidence calibration in image classification using the CIFAR-10 dataset and a ResNet-18 architecture. We apply five levels of label noise (0%, 10%, 20%, 40%, 60%) and evaluate their effects using metrics such as test accuracy, Expected Calibration Error (ECE) and predictive entropy. Our findings show that increasing noise levels significantly degrade classification accuracy and impair model calibration. In particular, asymmetric noise at a 60% corruption level causes test accuracy to drop to approximately 38.7% while ECE surges above 35%, indicating extreme overconfidence in incorrect predictions. By contrast, symmetric noise at the same noise level yields higher predictive entropy (uncertainty) and a comparatively modest miscalibration (ECE ~9%). These results highlight the importance of distinguishing noise types when assessing model robustness and reliability. All experiments are reproducible, with code and data publicly available to facilitate further investigation.

Keywords: Deep learning; Machine learning; Decentralisation; Artificial Intelligence (AI)

Introduction

Deep Neural Networks (DNNs) have achieved remarkable success in image classification tasks thanks to the availability of large-scale labeled datasets such as CIFAR-10 and ImageNet. However, real-world datasets are rarely free of label noise incorrect annotations arising from human error, ambiguous cases or automated labeling. Learning from datasets contaminated with noisy labels is especially concerning in safety-critical domains (*e.g.*, medical diagnosis or autonomous driving) because it can compromise model accuracy, generalization and calibration.

Previous studies have demonstrated that modern over-parameterized neural networks possess sufficient capacity to memorize even completely random labels [1]. Such memorization undermines the model's ability to generalize and leads to a deterioration in classification accuracy on unseen, correctly labeled data. Moreover, label noise can severely degrade model calibration, resulting in predictions that are made with unjustifiably high confidence despite being incorrect. This issue is particularly dangerous in safety-critical applications where trustworthy probability estimates are essential.

Model calibration, often quantified by metrics such as Expected Calibration Error (ECE), measures how closely a model's predicted probabilities align with the true likelihood of correctness [2]. While considerable research has explored methods to enhance calibration on clean (noise-free) datasets (*e.g.*, temperature scaling, Bayesian inference and ensemble techniques), there has been relatively limited systematic investigation into how different types and magnitudes of label noise affect calibration.

This study aims to bridge that gap by conducting a systematic analysis of the effects of symmetric (random) and asymmetric (structured, class-dependent) label noise on both the accuracy and confidence calibration of DNN classifiers. Using the well-established CIFAR-10 dataset and a standard ResNet-18 model under controlled conditions, we train

https://doi.org/10.64142/jeai.1.2.22

Journal of Engineering and Artificial Intelligence

Research Article

models with varying noise levels (0%, 10%, 20%, 40%, 60%) and evaluate performance in terms of test accuracy, ECE and predictive entropy.

Our contributions include:

- A thorough empirical investigation of how symmetric *vs.* asymmetric label noise affects both classification accuracy and confidence calibration in an image classification setting.
- Demonstration of the differential impacts on model confidence: Symmetric noise tends to make the model more cautious (higher uncertainty, reduced tendency toward overconfidence) whereas asymmetric noise can lead to the model being confidently wrong (severe miscalibration), especially at high noise levels.
- Public release of comprehensive code and datasets to facilitate reproducibility and further research on noisy-label training.

Through these contributions, we aim to clarify the specific challenges posed by noisy labels and underscore the importance of addressing label noise in both research and practical applications.

Related work

Learning with noisy labels has been extensively studied in the machine learning literature. Zhang et al., famously showed that deep networks can memorize random labels, underscoring the need for strategies to mitigate the impact of mislabeled data. Numerous approaches have since been proposed to make training more robust to label noise, including loss correction techniques and semi-supervised or multi-network methods such as co-teaching [3,4]. These methods primarily aim to improve model accuracy under noisy labels (for example, by estimating the noise distribution or selecting potentially clean samples during training) and generally pay less attention to the confidence calibration of the model's predictions.

Separately, the problem of probabilistic calibration of DNNs has gained attention in recent years. Guo et al., demonstrated that modern neural networks are often poorly calibrated and introduced temperature scaling as an effective post-hoc calibration method for models trained on clean data [2]. Other techniques, including Bayesian neural networks and deep ensembles, have also been explored to produce better-calibrated uncertainty estimates.

To date, however, the interplay between label noise and model calibration remains underexplored. Our work is situated at this intersection: whereas prior studies have improved robustness to noisy labels or improved calibration in isolation, we provide a systematic examination of how label noise itself affects both accuracy and the fidelity of predicted probabilities.

Methodology

Dataset

We use the CIFAR-10 dataset, a widely used benchmark for image classification that contains 60,000 color images of size 32 × 32 pixels, evenly divided into 10 classes (airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck). The dataset is split into 50,000 training images and 10,000 test images. We leave the test set unmodified and always use it as a clean evaluation benchmark. CIFAR- 10's manageable size and standardized format make it well-suited for controlled experiments involving synthetic label corruption.

Model

All experiments employ a standard ResNet-18 convolutional neural network implemented in PyTorch [5]. ResNet-18 consists of 18 layers with residual skip connections that facilitate stable training of deeper networks. We train the model from scratch (random initialization, no pre-training) under identical hyperparameter settings across all noise types and levels to ensure fair comparisons.

Label Noise Types

We simulate two common types of label noise:

- **Symmetric noise:** A randomly selected fraction p of the training labels is replaced with a uniformly random label from the other 9 classes. This models annotation errors that are independent of the input features (*e.g.*, an annotator randomly clicking the wrong class).
- **Asymmetric noise:** A fraction p of the training labels is replaced ac- cording to a fixed class-to-class mapping designed to mimic more realistic label confusion. Specifically, we use the following mapping (adapted from prior work): truck → automobile; bird → airplane; deer → horse; cat ↔ dog (cats and dogs are relabeled as each other). This structured form of noise reflects common mislabeling patterns where visually similar classes are confused.

Label noise is introduced only in the training set. The corruption is applied once before training and remains fixed throughout training (*i.e.,* the noisy labels are not corrected or changed during the training process). For each experiment, we set $p \in \{0\%, 10\%, 20\%, 40\%, 60\%\}$, yielding a 2 × 5 grid of experiments (10 total runs covering each combination of noise type and noise level).

https://doi.org/10.64142/jeai.1.2.22
Journal of Engineering and Artificial Intelligence
Research Article

Training procedure

All models are trained for 50 epochs using Stochastic Gradient Descent (SGD) with momentum 0.9 and weight decay 5×10^{-4}. The initial learning rate is 0.1 and it is reduced by a factor of 10 after epoch 40. We use a batch size of 128. Standard data augmentation is applied during training (random horizontal flips and random cropping with 4-pixel padding) to improve generalization.

To ensure reproducibility, all experiments are run with a fixed random seed controlling weight initialization, data shuffling and noise injection.

We do not employ any special label-noise mitigation techniques (such as label smoothing, loss correction or noise-robust training) so as to isolate the effects of the noise itself on model training dynamics.

Evaluation metrics

After training each model, we evaluate performance on the clean (uncorrupted) CIFAR-10 test set. The following metrics are computed:

- **Test accuracy:** The fraction of correctly classified samples in the test set, which measures standard generalization performance.
- **Expected Calibration Error (ECE):** A scalar measure of calibration, defined as the weighted average gap between predicted confidence and empirical accuracy across confidence bins [2]. We compute ECE using 15 equal-width confidence bins. A lower ECE indicates better calibration (*i.e.,* the model's predicted probabilities are more in line with the true frequencies of correctness).
- **Predictive entropy:** The average Shannon entropy of the model's predicted class probability distribution over all test samples. Higher entropy corresponds to greater predictive uncertainty.
- **Training performance:** We track training loss and accuracy over epochs to observe learning dynamics, including signs of overfitting or memorization of noisy labels.

All metrics are computed using standard implementations (*e.g.,* from the TorchMetrics library or custom scripts). Complete training logs and metric outputs are saved for each experiment to enable post-hoc analysis and verification.

Experiments

To systematically study the impact of label noise on model accuracy and confidence calibration, we conducted a series of controlled experiments. Each experiment is defined by a pair of:

- Noise type: either symmetric or asymmetric, as described above;
- Noise level: one of {0%, 10%, 20%, 40%, 60%} label corruption.

This yields a total of 10 distinct training runs (5 noise levels × 2 noise structures). Each model is trained and evaluated independently using the common methodology described in the previous section. In all cases, label noise is injected into the training set according to the specified noise type and level and remains fixed during training.

We implemented all experiments in PyTorch and executed training on a single GPU. After training each model for 50 epochs, we evaluate it on the standard CIFAR-10 test set (which has no label noise) and record the test accuracy, ECE and entropy.

We also log training metrics (loss and accuracy) to help diagnose the learning behavior under noise. All stochastic aspects (dataset shuffling, initialization, noise selection) are controlled by fixed seeds to ensure that the results are repeatable.

Results

We now present the results of our experiments, highlighting how different types and levels of label noise affect model accuracy, calibration and uncertainty.

The key evaluation metrics are test accuracy, Expected Calibration Error (ECE) and predictive entropy, all measured on the clean CIFAR-10 test set after training.

Accuracy declines with increased noise

Figure 1 illustrates the test accuracy as a function of noise level for both symmetric and asymmetric noise. As expected, test accuracy decreases monotonically as the proportion of noisy labels increases. This trend holds under both noise types, but it is more pronounced with asymmetric noise. Without label noise (0% noise), the model achieves about 92–93% test accuracy.

At the highest noise level of 60%, the model trained with symmetric noise still correctly classifies just over half of the test samples (53.9% accuracy), suggesting it can extract some meaningful signal from the remaining clean portion of the data. In contrast, 60% asymmetric noise causes accuracy to plummet to only 38.7%, meaning the model misclassifies the majority of test samples.

https://doi.org/10.64142/jeai.1.2.22
Journal of Engineering and Artificial Intelligence
Research Article

The stark difference at high noise levels is likely due to the model learning persistent but misleading patterns from systematically corrupted labels (for example, confusing cats with dogs or trucks with automobiles due to the noise mapping).

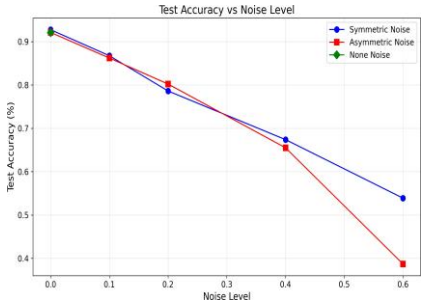

Figure 1: Test accuracy on CIFAR-10 under varying levels of symmetric vs. asymmetric label noise. Accuracy drops as noise level increases, with asymmetric noise leading to more severe degradation than symmetric noise at equivalent noise rates.

Calibration error rises under noise (especially asymmetric noise)

Label noise also has a strong impact on model calibration. **Figure 2** shows the Expected Calibration Error (ECE) across noise levels for each noise type. Under symmetric noise, ECE increases only modestly as noise grows: from around 3.7% with no noise to about 9.0% at 60% noise. Interestingly, at low symmetric noise levels (10-20%), we observed ECE values slightly lower than the clean baseline. For example, at 10% symmetric noise the ECE dropped to roughly 2.0%, better (lower) than the 3.7% ECE with 0% noise. This suggests that a small amount of random label noise can make the model more cautious, yielding slight improvements in calibration. However, at higher symmetric noise fractions, ECE worsens again, indicating that excessive noise eventually harms calibration.

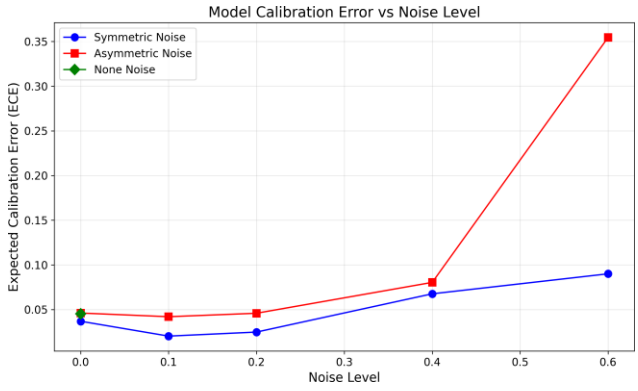

Figure 2: Expected Calibration Error (ECE) on the test set *vs.* noise level. Symmetric noise (random label flips) causes only a slight increase in ECE at the highest noise levels, whereas asymmetric noise (structured flips) leads to a dramatic spike in ECE, indicating severe miscalibration under heavy noise.

In contrast, asymmetric noise produces a sharp breakdown in calibration at high noise rates. ECE stays around 4-5% for mild asymmetric noise (0-20%), then jumps to roughly 8% at 40% noise and surges to 35.4% at 60% noise. In heavy structured-noise conditions, the model becomes grossly miscalibrated. In other words, it is often confidently wrong on the test set: The network makes many high-confidence predictions that turn out to be incorrect, which inflates the calibration error dramatically.

Predictive entropy reveals confidence patterns

To further understand the model's confidence under noisy training conditions, we examine the average predictive entropy of the model's outputs on test data (**Figure 3**). Entropy provides insight into how uncertain the model is in its predictions. We find that symmetric noise leads to a large increase in entropy as the noise level grows. For example, the average prediction entropy rises from about 0.10 nats at 0% noise to 1.64 nats at 60% symmetric noise. This trend indicates that when many training labels are randomly corrupted, the model becomes much less confident (more uncertain) in its predictions on clean test data.

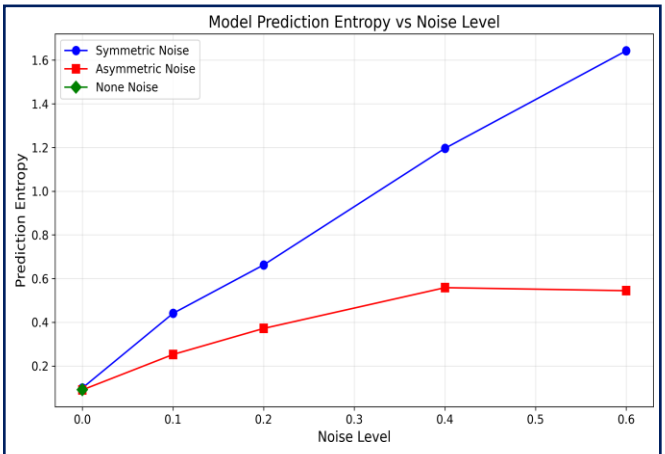

Figure 3: Average predictive entropy of model outputs vs. noise level. Higher entropy indicates greater uncertainty. Under symmetric noise, the model's predictive entropy increases substantially with noise level, reflecting growing uncertainty. Under asymmetric noise, entropy increases only slightly, indicating the model remains relatively confident (low uncertainty) despite being trained on corrupted labels.

By contrast, asymmetric noise yields only a slight increase in entropy with noise level. The entropy goes from about 0.09 nats at 0% to 0.54 nats at 60% asymmetric noise a relatively small change compared to the symmetric case. Thus, under structured noise, the model maintains a much higher confidence (lower uncertainty) even as it is trained on systematically wrong labels.

https://doi.org/10.64142/jeai.1.2.22
Journal of Engineering and Artificial Intelligence
Research Article

This divergence is particularly important: symmetric noise appears to induce honest uncertainty (the model recognizes its uncertainty under heavy random noise), whereas asymmetric noise results in overconfident misclassification (the model remains confident even when many of its predictions are wrong).

Overfitting and memorization of noisy labels

The training dynamics reveal that models tend to overfit more severely under asymmetric noise than under symmetric noise. In the experiment with 60% symmetric noise, the final training accuracy reached only about 47.2%, reflecting the model's difficulty in memorizing the purely random incorrect labels. Notably in this case the test accuracy (53.9%) was actually higher than the training accuracy, suggesting that the model effectively "gave up" on many of the noisy training examples and thus generalized better on the clean test set than it fit the noisy training set. In contrast, with 60% asymmetric noise, the model achieved a high training accuracy of approximately 77.7%, indicating that it managed to learn (*i.e.*, memorize) a large portion of the incorrect labels. This memorization coincided with very poor generalization: the test accuracy was only 38.7% and ECE was extremely high. These observations align with the memorization hypothesis in deep learning, which posits that DNNs fit the clean patterns in the data first and only later absorb the noise if training continues long enough. Here we see that without any noise-specific regularization or early stopping, a sufficiently flexible model will eventually overfit structured noise, reinforcing wrong patterns in the training data and dramatically harming test performance and calibration.

Class-specific performance under asymmetric noise

Asymmetric noise introduces systematic biases that unevenly affect different classes. We observed that certain classes involved in the directed label flips suffered particularly large drops in test accuracy under asymmetric noise. For example, in the 60% asymmetric noise setting, test accuracy on the truck and cat classes was extremely low compared to other classes, since many training examples of "truck" had been mislabeled as "automobile" and cats were frequently mislabeled as dogs in the training set. Meanwhile, classes that were not part of any noise mapping (such as frog) retained relatively high accuracy even at the highest noise level. In contrast, symmetric noise (being random) tended to impact all classes more uniformly (roughly in proportion to their frequency in the training data). These class-specific effects highlight that structured label noise can create critical failure modes concentrated in certain categories, which is an important consideration for real-world applications where some classes may be far more prone to systematic labeling errors than others.

Discussion

Our experimental results reveal a nuanced relationship between label noise, classification accuracy and confidence calibration in deep neural networks. The effects differ substantially depending on the nature of the noise whether it is symmetric (randomized) or asymmetric (structured) and on the proportion of noisy labels. Below, we elaborate on several key observations and their broader implications.

Symmetric *vs.* asymmetric noise

While both noise types degrade model accuracy, their impacts on model calibration diverge significantly. Under symmetric noise, accuracy drops steadily as noise increases and the model's predictive uncertainty grows (as reflected by higher entropy). The model becomes less confident as the labels become less trustworthy, which paradoxically helps preserve calibration (ECE remains relatively low even at high noise levels). By contrast, asymmetric noise causes the model to learn and trust incorrect label patterns, leading to a model that is confidently wrong. Calibration suffers much more in this scenario (ECE values reach above 35% at high noise). These patterns underscore the importance of characterizing the structure of label noise: in real-world datasets where mislabeling often follows a pattern (*e.g.*, consistent annotation errors for specific classes), asymmetric noise is likely both a more realistic and a more dangerous scenario for model reliability.

Confidence degradation *vs.* overconfidence

One notable contrast between the noise types is that symmetric noise tends to make the model cautious, whereas asymmetric noise leaves it overconfident. In our experiments, symmetric noise greatly increased the model's output entropy, indicating that the network was often unsure about its predictions when trained on a substantial fraction of random labels. This reduction in confidence can actually be beneficial from a calibration standpoint, as the model is less likely to be confidently wrong. In practical terms, a model trained with a great deal of symmetric noise at least signals its uncertainty (*e.g.*, by outputting a more diffuse probability distribution). On the other hand, a model trained on heavy asymmetric noise maintained low entropy (high confidence) even as accuracy plummeted, implying severe overconfidence. In safety-critical domains such as healthcare or autonomous driving, this is a serious concern: A wrong prediction is problematic, but a wrong prediction made with high certainty is potentially

https://doi.org/10.64142/jeai.1.2.22
Journal of Engineering and Artificial Intelligence
Research Article

disastrous. Our results highlight that label noise can either degrade overall confidence or perversely inflate unwarranted confidence, depending on the noise structure.

Overfitting to noisy labels

The tendency of deep networks to eventually fit or memorize noisy data was clearly observed in our study, particularly under asymmetric noise. At 60% asymmetric noise, the model attained over 77% training accuracy but only ~39% test accuracy, indicating that it had memorized a significant fraction of the wrong labels. In contrast, with 60% symmetric noise, the training accuracy remained much lower (around 47%), showing that the model struggled to fit purely random noise and effectively "gave up" on many corrupted examples. Interestingly, this led to better generalization in that condition (the test accuracy exceeded the training accuracy). These outcomes are consistent with the memorization dynamics described by Zhang et al.: DNNs fit the clean patterns in the data first and only later (if at all) do they begin to absorb the noise. Our results reinforce that without interventions such as early stopping or strong regularization, prolonged training on highly noisy data (especially structured noise) will drive the model to overfit the noise, severely impairing test performance and calibration.

Calibration as a diagnostic tool

Our findings also show that calibration metrics like ECE can serve as valuable diagnostics for training issues. In conditions with low or moderate noise, we observed only modest ECE (a few percent), indicating the model's confidence was reasonably aligned with its accuracy. However, at high asymmetric noise levels, ECE spiked to over 35%, flagging a serious misalignment between confidence and correctness. This occurred even as the training accuracy remained relatively high, which might otherwise mislead one to believe the model was performing adequately. Thus, monitoring calibration errors can provide an early warning sign of a model that is "succeeding" on the training data by memorizing noise, yet failing to be trustworthy on true clean data. In general, we advocate that practitioners evaluate model calibration (not just accuracy) when working with potentially noisy datasets, as it can reveal problems that accuracy alone would not show.

Practical implications

In practical settings where some amount of label noise is inevitable, our results suggest several actionable recommendations:

- **Characterize your label noise:** Attempt to determine whether label errors are mostly random or follow systematic patterns. This knowledge can inform mitigation strategies (for example, targeted relabeling efforts for certain classes versus more global noise-robust training techniques).
- **Always monitor calibration:** When training on potentially noisy data, track not only accuracy but also calibration metrics (*e.g.,* ECE) or inspect reliability diagrams. A model that is wrong and confident is much more problematic than one that is wrong but aware of its uncertainty.
- **Use regularization or early stopping:** In high-noise regimes, techniques such as early stopping, stronger regularization or noise-robust loss functions can help prevent the model from completely memorizing noisy labels and becoming overconfident. Our results showed that if unchecked, a network will eventually fit structured noise; interventions are necessary to preserve generalization and reliable confidence estimates.

These considerations apply not only in academic experiments but also in real world deployments across computer vision, natural language processing, healthcare AI and other domains where data quality cannot be guaranteed. Ensuring that models remain both accurate and well-calibrated in the face of noisy data is crucial for building reliable machine learning systems.

Limitations and future work

While this study provides detailed insights into the effects of label noise on model accuracy and calibration, it has several limitations that suggest directions for future work. First, due to computational constraints, we ran only a single training trial for each noise configuration; repeating experiments multiple times would allow us to assess the variability of outcomes and attach statistical significance to the conclusions. Second, our experiments were limited to one dataset (CIFAR-10) and one model architecture (ResNet-18). Although this setup sufficed to illustrate clear trends, the quantitative impact of noise may differ with more complex datasets (*e.g.,* CIFAR-100 or ImageNet) or with different architectures (such as Vision transformers or larger ResNets). Third, we intentionally did not employ any noise-handling techniques (like label smoothing, sample re-weighting or semi-supervised label cleaning) in order to isolate the natural effect of noise; incorporating such methods could potentially improve both accuracy and calibration and analyzing their benefits would be a valuable extension. Fourth, our noise injection was synthetic and static: Real-world label noise might evolve over time or affect classes unevenly and models might need strategies to handle time-varying or non-uniform noise. Finally, our analysis mostly focused on the end of training; examining how calibration and accuracy evolve during training (especially early on, as the model fits clean *vs.* noisy labels) could offer deeper

understanding and inform interventions like adaptive early stopping.

Conclusion

In this work, we conducted a systematic investigation into how label noise affects both the accuracy and confidence calibration of deep image classifiers. Using the CIFAR-10 dataset and a ResNet-18 model, we evaluated the impact of symmetric and asymmetric label noise across five noise levels ranging from 0% to 60%. Our analysis considered key metrics including test accuracy, Expected Calibration Error (ECE) and predictive entropy. The results demonstrate that label noise not only degrades classification accuracy but also compromises the reliability of the model's predicted confidences. Symmetric (random) noise leads to substantially higher prediction uncertainty (entropy) but maintains relatively moderate calibration error. In contrast, asymmetric (structured) noise causes the model to become overconfident in its incorrect predictions, resulting in severe miscalibration particularly at high noise levels where ECE exceeded 35%. These findings reinforce the need to distinguish between different types of label noise when designing training protocols or evaluating model trustworthiness. Moreover, our work highlights the value of calibration metrics such as ECE and entropy as diagnostic tools in noisy-label scenarios. We have openly released all code, data and evaluation scripts used in this study to facilitate replication and extension of our results. In the future, it would be worthwhile to explore calibration-preserving training methods under noisy labels, such as noise-aware loss functions, confidence-based example filtering or adaptive early stopping strategies. Investigating larger-scale datasets and modern architectures can also help assess the generality of the trends observed here. Ultimately, we conclude that label noise affects not only what a model learns to predict, but also how confident it is in those predictions and under- standing this dual effect is critical for developing robust, trustworthy machine learning systems.

Reproducibility

To facilitate transparency and reproducibility, all code, data generation scripts and evaluation metrics used in this study are publicly available at:

https://github.com/cboseak/noisy-label-calibration

This repository includes instructions for reproducing the experiments, training models with varying label noise levels and generating the figures and calibration metrics reported in this paper.

References

1. Chiyuan Z, Samy B, Moritz H, Benjamin R oriol V (2017) Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*. [Crossref], [Google Scholar]
2. Chuan G, Geoff P, Yu S, Kilian Q (2017) On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)* 1321-1330. [Crossref], [Google Scholar]
3. Giorgio P, Alessandro R, Aditya KM, Richard N, Lizhen Q (2017) Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 1944-1952. [Google Scholar]
4. Bo H, Quanming Y, Xingrui Y, Gang N, Mingyuan X, et al. (2018) Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems (NeurIPS)* 8527-8537. [Crossref], [Google Scholar]
5. Kaiming H, Xiangyu Z, Shaoqing R, Jian S (2016) Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770-778. [Google Scholar]