



International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Confidence-Aware Machine Learning for Imbalanced and Noisy Datasets: A Joint Framework for Calibration, Minority-Class Recall, and Label-Noise Robustness

Neetu

Assistant Professor Department of Computer Science, Guru Nanak Khalsa College

Neetu.ciet@gmail.com

ABSTRACT :

Real-world classification tasks rarely provide the clean, balanced training data assumed by standard supervised learning pipelines; instead, they combine skewed class distributions with unreliable, noisily annotated labels. Existing research has largely treated class imbalance, confidence calibration, and label-noise robustness as separate problems, producing methods that improve one property while silently degrading another. A re-weighted loss that improves minority-class recall, for instance, can simultaneously amplify the influence of mislabeled examples and worsen calibration. This paper proposes Confidence-Aware Adaptive Learning (CAAL), a unified classification framework that jointly estimates prediction confidence, routes uncertain or suspect-noisy samples to a noise-adaptive loss re-weighting mechanism, and applies a dedicated minority-class recall booster within a single, iteratively updated training loop. The framework couples a shared feature encoder with a lightweight confidence-estimation head, enabling per-sample decisions about how much a given label should be trusted during optimization. A structured literature review spanning imbalance-handling, calibration, and noisy-label research identifies a concrete gap: no existing framework jointly evaluates confidence calibration, minority-class recall, and robustness to label noise under a shared experimental protocol. Building on this gap, the paper presents a proposed system architecture, a mathematical formulation of the training objective, and a detailed research methodology including controlled imbalance-noise experimental grids, component-wise ablation design, and calibration-aware evaluation metrics such as Expected Calibration Error and Area Under the ROC Curve intended to guide empirical validation of the framework on benchmark and real-world datasets.

Keywords: *confidence calibration; class imbalance; label noise; robust learning; uncertainty estimation; deep neural networks.*

1. INTRODUCTION AND OBJECTIVE

1.1 Background

Modern supervised machine learning systems are commonly trained under an implicit assumption that training data is approximately balanced across classes and that the associated labels are correct. In practice, this assumption rarely holds. Domains such as fraud detection, rare-disease diagnosis, industrial defect inspection, and crowd-sourced content moderation exhibit severe class imbalance, in which the class of genuine interest—fraud, disease, or defect—may represent well under one percent of all examples [1], [2]. At the same time, the labels attached to training examples are frequently noisy: they may originate from non-expert annotators, automated heuristics, or historical records containing errors, and empirical audits have shown that widely used benchmark datasets contain non-trivial fractions of mislabeled examples [9], [10].

These two conditions rarely occur in isolation, and their interaction is more damaging than either problem alone. A classifier trained to emphasize rare classes—for example through re-sampling [1] or re-weighting [3], [4]—implicitly increases the influence of every minority-class example on the learned decision boundary, including any that happen to be mislabeled. Conversely, many noise-robust training strategies identify likely-corrupted examples by treating atypical, hard-to-fit samples with suspicion [7], [9]; because minority-class examples are, by definition, atypical relative to the majority distribution, such strategies risk discarding or down-weighting exactly the rare, legitimate examples that imbalance-handling techniques are trying to protect.

1.2 Motivation

A further complication concerns predictive confidence. Deep neural networks are known to produce poorly calibrated probability estimates, typically over-confident relative to their true accuracy, even when trained on clean and balanced data [5]. Under class imbalance and label noise, this miscalibration is expected to worsen: majority-class predictions become artificially over-confident from sheer repetition, while minority-class predictions become unreliable because the model has too few clean examples from which to learn a stable decision boundary. A system that reports a confidence score without accounting for these two confounds risks misleading any downstream process—human reviewers, active-learning selectors, or automated

decision thresholds that depends on that score being trustworthy.

The motivation for this work is therefore not to introduce another isolated technique for imbalance or for noise robustness, but to recognize that calibration, minority-class performance, and noise robustness are three faces of the same underlying problem: knowing how much to trust a given prediction, and by extension, how much to trust the label used to produce it during training. A framework that reasons jointly about these three properties should, in principle, avoid the failure mode in which optimizing one metric degrades another.

1.3 Problem Statement

Let a labelled training set be denoted $D = \{(x_i, \hat{y}_i)\}$, $i = 1, \dots, N$, where $\hat{y}_i$ is an observed label that may differ from an unknown true label y_i with an unknown, possibly class-conditional noise rate, and where the empirical class prior $P(y)$ is highly skewed toward one or more majority classes. The problem addressed in this paper is to design a classification framework f_θ that simultaneously: (a) maximizes recall on minority classes without disproportionately fitting to mislabeled examples; (b) produces confidence scores $c(x)$ that are well calibrated, such that $P(\hat{y} = \arg\max f_\theta(x) \mid c(x) = p) \approx p$ for any confidence level p ; and (c) adaptively identifies and down-weights the influence of likely-noisy labels during training, while explicitly avoiding the systematic mis-classification of rare-but-genuine minority-class examples as noise.

1.4 Research Objectives

1. To review and consolidate existing literature on class-imbalance handling, confidence calibration, and learning with noisy labels, in order to characterize how each line of work treats the other two dimensions of the problem.
2. To identify, on the basis of this review, the specific gap left by treating imbalance, calibration, and noise robustness as independent problems.
3. To propose a unified system architecture Confidence-Aware Adaptive Learning (CAAL) that couple's confidence estimation with noise-adaptive loss re-weighting and a minority-class recall booster.
4. To formalize the training objective of the proposed framework mathematically, including the routing mechanism that separates trusted from suspect-noisy samples.
5. To design a rigorous research methodology, including a controlled imbalance–noise experimental grid, baseline comparisons, ablation design, and a joint evaluation-metric suite, suitable for guiding future empirical validation of the framework.

1.5 Organization of the Paper

The remainder of this paper is organized as follows. Chapter 2 reviews the literature on class-imbalance handling, confidence calibration, and noisy-label learning. Chapter 3 synthesizes this review into an explicit statement of the research gap. Chapter 4 presents the proposed Confidence-Aware Adaptive Learning framework, including its architecture and mathematical formulation. Chapter 5 details the proposed research methodology for validating the framework. Chapter 6 concludes the paper and outlines directions for future work, followed by the list of references.

2. LITERATURE REVIEW

This chapter reviews four bodies of work relevant to the proposed framework: methods for handling class imbalance, methods for confidence calibration, methods for learning with noisy labels, and the comparatively small body of work that begins to combine these dimensions.

2.1 Class Imbalance Handling

The earliest and most widely adopted approach to class imbalance is resampling. Chawla et al. [1] introduced the Synthetic Minority Over-sampling Technique (SMOTE), which generates synthetic minority-class examples by interpolating between a minority sample and its nearest minority-class neighbours in feature space, rather than simply duplicating existing points. He and Garcia [2] provide a comprehensive survey of the imbalanced-learning problem, organizing methods into sampling-based, cost-sensitive, and kernel/active-learning-based categories, and noting that simple predictive accuracy is an inappropriate metric once class distributions are skewed.

A second family of approaches operates at the loss-function level rather than the data level. Lin et al. [3] proposed focal loss, which reshapes the standard cross-entropy loss to down-weight the contribution of easily classified (typically majority-class) examples, concentrating gradient signal on harder examples. Cui et al. [4] later argued that simple inverse-frequency re-weighting over-corrects when classes contain highly redundant samples, and instead proposed weighting each class by its effective number of samples, a quantity that saturates as class size grows. Both approaches assume that every

training label is correct an assumption this paper argues is frequently violated precisely in the rare-class regime where labelling is hardest and most error-prone.

2.2 Confidence Calibration

Guo et al. [5] demonstrated that modern, high-capacity neural networks are systematically miscalibrated: their predicted confidence values do not match their empirical accuracy, and this miscalibration worsens with network depth and width. They proposed temperature scaling, a simple post-hoc method that rescales logits using a single learned scalar parameter fitted on a held-out validation set, and showed it to be surprisingly effective relative to more complex calibration methods. Kendall and Gal [6] took a complementary, training-time view, distinguishing aleatoric uncertainty (irreducible noise in the data) from epistemic uncertainty (uncertainty in the model itself due to limited data), and proposing a Bayesian deep-learning framework that models both jointly. Neither line of work, however, was evaluated under conditions of severe class imbalance or label noise, leaving open the question of how calibration behaves when the very labels used for calibration may themselves be unreliable.

2.3 Learning with Noisy Labels

A substantial body of research addresses training with unreliable labels directly. Han et al. [7] proposed co-teaching, in which two networks are trained simultaneously and each selects a subset of likely clean, small-loss examples to teach its peer network, exploiting the empirical observation that deep networks tend to memorize clean labels before noisy ones. Zhang and Sabuncu [8] proposed a generalized cross-entropy (GCE) loss that interpolates between the noise-robust mean absolute error and the noise-sensitive but well-behaved categorical cross-entropy, offering a tunable trade-off between robustness and convergence speed. Northcutt et al. [9] introduced confident learning, a framework that directly estimates the joint distribution between noisy observed labels and unknown true labels using per-class prediction confidence, allowing likely label errors to be identified and pruned prior to training. Song et al. [10] provide a comprehensive survey of over sixty robust-training methods, categorizing them by mechanism (robust loss, sample selection, meta-learning, and combined approaches) and explicitly flagging the interaction between label noise and class imbalance as a comparatively under-explored research direction.

2.4 Toward Joint Approaches

Across the reviewed literature, a consistent pattern emerges: imbalance-handling methods [1]–[4] are validated on datasets with clean labels; calibration methods [5], [6] are validated on balanced, clean benchmarks; and noisy-label methods [7]–[9] are validated on artificially corrupted but otherwise class-balanced datasets. The survey by Song et al. [10] is, to the knowledge of this review, the clearest acknowledgement in the literature that these three dimensions interact and are rarely studied together. Table 1 summarizes the reviewed methods, their primary focus, and their principal limitation with respect to the joint problem addressed in this paper.

Table 1. Summary of reviewed literature and limitations with respect to the joint imbalance–noise–calibration problem.

Ref.	Focus Area	Key Technique	Primary Limitation w.r.t. Joint Problem
[1]	Class imbalance	SMOTE: k-NN-based synthetic minority over-sampling	Synthetic points can be generated near or inside noisy/mislabeled minority samples, amplifying noise.
[2]	Class imbalance (survey)	Taxonomy of resampling and cost-sensitive methods	Does not consider label noise or model confidence calibration.
[3]	Class imbalance	Focal loss: down-weights well-classified (easy) examples	Assumes observed labels are correct; hard examples and noisy examples are conflated.
[4]	Class imbalance	Class-balanced loss via effective number of samples	Static, count-based re-weighting; no per-sample confidence or noise signal.
[5]	Confidence calibration	Temperature scaling (post-hoc calibration)	Calibrated on a clean held-out set; behaviour under imbalance/noise is untested.
[6]	Uncertainty estimation	Joint aleatoric + epistemic uncertainty modelling	Computationally expensive (Bayesian/MC-based); not evaluated under label noise.
[7]	Noisy labels	Co-teaching: two networks cross-	Requires training two networks; interaction with

Ref.	Focus Area	Key Technique	Primary Limitation w.r.t. Joint Problem
		select likely-clean samples	class imbalance unexplored.
[8]	Noisy labels	Generalized cross-entropy (GCE) robust loss	Fixed robustness level (single hyper-parameter); no calibration guarantee.
[9]	Noisy labels	Confident learning: identifies label errors via the confident joint	Assumes class-conditional noise; rare classes can be mistaken for noisy ones.
[10]	Noisy labels (survey)	Categorizes 62 robust-training methods	Explicitly notes the imbalance–noise interaction as a comparatively under-studied open problem.

3. RESEARCH GAP

The literature reviewed in Chapter 2 reveals four specific and interrelated gaps that motivate the framework proposed in this paper.

Gap 1 Imbalance-handling methods assume clean labels.

Resampling [1] and re-weighting [3], [4] techniques increase the training influence of minority-class examples but contain no mechanism for distinguishing a genuine rare example from a mislabeled one; every minority-class sample, correct or not, is amplified equally.

Gap 2 Noise-robust methods can mistake rare classes for noise.

Small-loss and confidence-based noise filters [7], [9] rely on the observation that networks fit clean examples faster and with lower loss than noisy ones. Genuinely rare, atypical minority-class examples can exhibit the same slow-fitting, high-loss signature as mislabeled examples during early training, so noise-filtering criteria calibrated on balanced data risk systematically discarding exactly the examples that matter most for minority-class recall.

Gap 3 Calibration methods are validated outside the imbalance–noise regime.

Post-hoc calibration [5] and uncertainty-modelling approaches [6] are evaluated on clean, balanced benchmarks. It remains unclear from the existing literature whether a single global temperature parameter, or a class-agnostic uncertainty model, remains valid when minority classes are both under-represented and disproportionately affected by label noise, or whether calibration must instead be modelled per class or per confidence region.

Gap 4 No shared evaluation protocol exists for the joint problem.

Because the three lines of work in Chapter 2 are evaluated with different metrics – accuracy and F1 for imbalance, Expected Calibration Error (ECE) for calibration, and noise-detection precision/recall for noisy-label methods – no existing study reports all three properties for the same model under the same controlled experimental conditions. Song et al. [10] identify this absence explicitly as an open direction.

These four gaps jointly motivate the central contribution of this paper: a framework that treats confidence estimation as the shared mechanism linking minority-class recall and noise robustness, together with a research methodology that evaluates all three properties under one controlled experimental protocol, described in Chapters 4 and 5 respectively.

4. PROPOSED FRAMEWORK

4.1 Design Principles

The proposed framework, Confidence-Aware Adaptive Learning (CAAL), is guided by three design principles drawn directly from the gaps identified in Chapter 3. First, confidence must be estimated per sample and used as the shared signal that governs both noise-adaptive loss re-weighting and minority-class emphasis, rather than treating these as independent modules. Second, the routing mechanism that separates trusted from suspect-noisy samples must be aware of class frequency, so that rare-but-genuine examples are not systematically treated as noise. Third, the training objective must include an explicit calibration term, so that the resulting confidence scores remain interpretable rather than being a by-product of optimizing accuracy alone.

4.2 System Architecture

Figure 1 illustrates the proposed architecture. Raw input data first passes through a pre-processing stage that applies mild, class-aware resampling

and augmentation. A shared feature encoder then produces a latent representation that feeds two components: a standard classification head, which produces class probabilities, and a confidence-estimation module, which produces a scalar confidence score for the current prediction. This confidence score drives a routing decision: samples with high confidence are treated as trusted and contribute to a standard, class-re-balanced cross-entropy term, while samples with low confidence are passed to a noise-adaptive loss re-weighting mechanism that reduces their influence in proportion to estimated noise likelihood. In parallel, a minority-class recall booster applies an additional class-frequency-aware adjustment so that low-confidence minority-class samples are not discarded outright, but are down-weighted more conservatively than low-confidence majority-class samples. The outputs of both branches are combined into a single parameter-update step, and the cycle repeats iteratively as training progresses, allowing the confidence estimates themselves to improve as the encoder learns a more reliable representation.

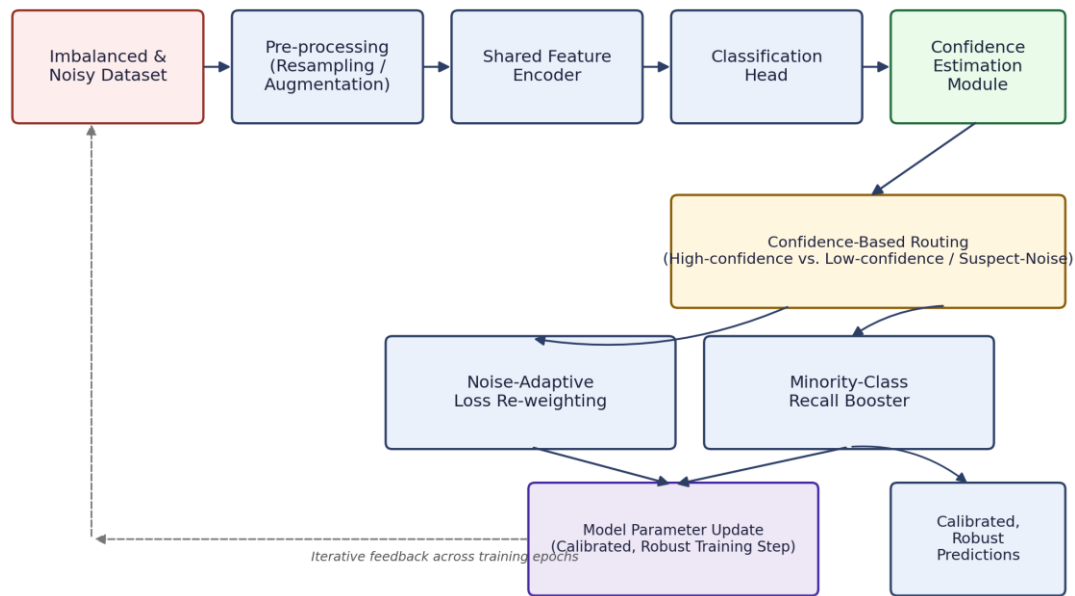


Figure 1. Proposed Confidence-Aware Adaptive Learning (CAAL) architecture.

4.3 Mathematical Formulation

The shared encoder maps each input to a latent representation:

$$z_i = f_{enc}(x_i; \theta_e)$$

The classification head produces a probability vector $p_i = \text{softmax}(f_{head}(z_i; \theta_h))$, and the confidence module produces a scalar $c(x_i) = f_{conf}(z_i; \theta_c) \in [0, 1]$, trained jointly with the classifier. Each sample is routed according to two thresholds $\tau_h > \tau_l$:

- If $c(x_i) \geq \tau_h$, the sample is treated as trusted and contributes to a class-re-balanced cross-entropy term, using the effective-number-of-samples weighting of Cui et al. [4].
- If $c(x_i) < \tau_l$, the sample is treated as suspect-noisy and contributes through a generalized cross-entropy term [8], whose robustness parameter q down-weights the influence of examples with high loss and low confidence.
- If $\tau_l \leq c(x_i) < \tau_h$, the sample lies in an intermediate band and receives a linearly interpolated weight between the two regimes, preventing a hard discontinuity at the routing boundary.

The minority-class recall booster modifies the effective per-sample weight w_i by a class-frequency factor so that the down-weighting applied to low-confidence samples is attenuated for rare classes:

$$w_i = \left[\sigma(c(x_i); \tau_h, \tau_l) \right] \cdot \left[(1 - \beta) / (1 - \beta^{(n_{yi})}) \right]^{(1 - c(x_i))}$$

where n_{yi} is the number of training examples in the class of sample i , $\beta \in [0, 1]$ is the effective-number hyper-parameter, and $\sigma(\cdot)$ is the routing function that maps confidence to a base weight in $[0, 1]$. The exponent $(1 - c(x_i))$ ensures that the class-rarity correction has the greatest effect precisely when confidence is low. The overall training objective combines a re-weighted classification loss with an explicit calibration regularizer and a confidence-

penalty term that discourages the confidence module from collapsing to a trivial constant output:

$$L_{total} = \lambda_1 \cdot \sum_i w_i \cdot L_{CE}(p_i, \hat{y}_i) + \lambda_2 \cdot L_{calib}(p, \hat{y}; T) + \lambda_3 \cdot L_{conf}(c, correctness)$$

Here L_{calib} is a temperature-scaling-style regularizer following Guo et al. [5], fitted on a small clean held-out subset, and L_{conf} is a confidence-penalty term in the spirit of Kendall and Gal's loss-attenuation formulation [6], which rewards the confidence module for assigning high confidence to correctly classified samples and low confidence to misclassified ones. The coefficients $\lambda_1, \lambda_2, \lambda_3$ balance the three objectives and are treated as tunable hyper-parameters in the methodology of Chapter 5. Table 2 summarizes the notation used throughout this section.

Table 2. Notation used in the proposed framework.

Symbol	Description
$x_i, \hat{y}_i$	Input feature vector and observed (possibly noisy) label for training example i
y_i	Unknown true label for example i
$f_{enc}(\cdot; \theta_e)$	Shared feature encoder mapping x_i to a latent representation z_i
p_i	Predicted class-probability vector, $p_i = \text{softmax}(f_{head}(z_i))$
$c(x_i)$	Predicted confidence score from the confidence-estimation module, $c(x_i) \in [0, 1]$
τ_h, τ_l	High- and low-confidence routing thresholds
w_i	Effective per-sample training weight after routing and class re-balancing
β	Effective-number-of-samples hyper-parameter (class re-balancing)
q	GCE robustness hyper-parameter, $q \in (0, 1]$
T	Temperature-scaling parameter for calibration
$\lambda_1, \lambda_2, \lambda_3$	Loss-mixing coefficients for classification, calibration, and confidence-penalty terms

4.4 Component Description

(a) Shared Feature Encoder.

Any standard backbone (e.g., a convolutional network for images or a transformer encoder for text/tabular embeddings) may serve as f_{enc} ; the framework is architecture-agnostic and is designed as a training-time wrapper rather than a bespoke network topology.

(b) Confidence-Estimation Module.

A lightweight auxiliary head attached to the shared representation, trained jointly with the classifier via L_{conf} , so that confidence is learned rather than derived post hoc solely from the softmax maximum, which is known to be an unreliable confidence proxy under miscalibration [5].

(c) Confidence-Based Routing.

A two-threshold routing function that separates trusted, suspect-noisy, and intermediate samples, avoiding the sharp trusted/untrusted split used in prior sample-selection methods [7], [9].

(d) Noise-Adaptive Loss Re-weighting.

A generalized cross-entropy component [8] applied specifically to the suspect-noisy branch, so that robustness is concentrated where it is needed rather than applied uniformly to the entire dataset.

(e) Minority-Class Recall Booster.

A class-frequency correction, informed by the effective-number-of-samples principle [4], that is deliberately coupled to the confidence signal so that rare classes are protected specifically in the low-confidence regime where they are most vulnerable to being discarded as noise.

4.5 Novelty Statement

The novelty of the proposed framework lies not in any single component—confidence estimation, class-balanced weighting, and robust losses each have precedent in isolation [4], [5], [8]—but in the explicit coupling of all three under one routing mechanism and one training objective, together with a research methodology (Chapter 5) that evaluates calibration, minority-class recall, and noise robustness jointly rather than in isolation. As summarized in Table 1, no reviewed method reports all three properties under a shared experimental protocol; this triangulated evaluation is the specific contribution this paper puts forward for empirical validation.

5. RESEARCH METHODOLOGY

5.1 Research Design

This study adopts a quantitative, controlled experimental design intended to isolate the effect of imbalance, label noise, and their interaction on classification performance, calibration, and noise robustness. Rather than relying on a single benchmark, the methodology combines synthetic imbalance–noise injection, which allows precise control over the two experimental factors, with real-world datasets that exhibit naturally occurring imbalance and label uncertainty, in order to assess external validity.

5.2 Dataset Selection

Three categories of datasets are proposed for evaluation. First, a long-tailed image classification benchmark (e.g., a class-imbalanced variant of CIFAR-10/CIFAR-100) provides a controlled setting in which imbalance ratio and injected label noise can both be varied systematically. Second, a naturally imbalanced tabular dataset drawn from a high-stakes application domain, such as credit-card fraud detection or a clinical diagnosis dataset, is proposed to evaluate the framework under real-world, uncontrolled noise. Third, a text-classification dataset with an imbalanced label distribution (e.g., an imbalanced sentiment or topic-classification corpus) is proposed to assess whether the architecture-agnostic design of the framework generalizes beyond vision tasks.

5.3 Controlled Imbalance–Noise Injection Protocol

For the synthetic benchmark, imbalance is induced by sub-sampling minority classes to fixed majority-to-minority ratios, and label noise is injected using both symmetric noise (labels flipped uniformly at random among incorrect classes) and asymmetric, class-conditional noise (labels flipped preferentially toward visually or semantically similar classes), consistent with the noise models used in prior noisy-label research [7]–[9]. Table 3 defines the proposed experimental grid, crossing four imbalance ratios with four noise levels, yielding sixteen controlled conditions per dataset, each evaluated with the proposed framework and every baseline listed in Section 5.4.

Table 3. Proposed experimental design grid crossing imbalance ratio and label-noise rate.

Imbalance Ratio (majority : minority)	0% Noise	10% Noise	20% Noise	40% Noise
1 : 1 (balanced)	full training run + evaluation suite	full training run + evaluation suite	full training run + evaluation suite	full training run + evaluation suite
10 : 1	full training run + evaluation suite	full training run + evaluation suite	full training run + evaluation suite	full training run + evaluation suite
50 : 1	full training run + evaluation suite	full training run + evaluation suite	full training run + evaluation suite	full training run + evaluation suite
100 : 1	full training run + evaluation suite	full training run + evaluation suite	full training run + evaluation suite	full training run + evaluation suite

5.4 Baseline Methods

The proposed framework is to be compared against representative methods from each reviewed category: plain cross-entropy training (naïve baseline); SMOTE-based resampling [1] combined with cross-entropy; focal loss [3]; class-balanced loss [4]; temperature scaling applied post hoc to the naïve baseline [5]; co-teaching [7]; and generalized cross-entropy loss alone [8]. This baseline set spans imbalance-only, calibration-only, and noise-only

methods, enabling the joint evaluation described in Section 5.5 to quantify what is gained specifically from combining these mechanisms.

5.5 Evaluation Metrics

Consistent with the joint nature of the research problem, no single metric is treated as primary. Table 4 lists the proposed evaluation-metric suite, grouped by the dimension of the problem each metric is intended to capture.

5.6 Ablation Study Design

To isolate the contribution of each architectural component, four ablated variants of the framework are proposed: (i) the confidence-estimation module replaced with the raw softmax maximum as a naïve confidence proxy; (ii) the noise-adaptive re-weighting branch removed, routing all samples through standard cross-entropy; (iii) the minority-class recall booster removed, leaving confidence-based routing unadjusted for class frequency; and (iv) the calibration regularization term (λ_2) set to zero. Each ablated variant is to be evaluated across the same experimental grid described in Table 3, allowing the marginal contribution of each component to be quantified against the full framework.

Table 4. Proposed evaluation metrics, grouped by problem dimension.

Metric	What It Captures	Relevant Dimension
Overall Accuracy	Proportion of correctly classified test samples	Baseline reference only
Macro-averaged F1	Unweighted average of per-class F1 scores	Minority-class performance
Minority-Class Recall	True positive rate restricted to minority class(es)	Minority-class performance
Expected Calibration Error (ECE)	Weighted gap between confidence and empirical accuracy across bins	Confidence calibration
Brier Score	Mean squared error between predicted probability and outcome	Confidence calibration
AUROC	Separability of correct vs. incorrect predictions by confidence score	Confidence calibration
Noise-Detection Precision/Recall	Agreement between routed "suspect-noisy" samples and known injected noise	Noise robustness
Minority-Recall Retention under Noise	Minority recall at increasing injected noise rates, relative to 0% noise	Joint imbalance–noise interaction

5.7 Ablation Study Design

To isolate the contribution of each architectural component, four ablated variants of the framework are proposed: (i) the confidence-estimation module replaced with the raw softmax maximum as a naïve confidence proxy; (ii) the noise-adaptive re-weighting branch removed, routing all samples through standard cross-entropy; (iii) the minority-class recall booster removed, leaving confidence-based routing unadjusted for class frequency; and (iv) the calibration regularization term (λ_2) set to zero. Each ablated variant is to be evaluated across the same experimental grid described in Table 3, allowing the marginal contribution of each component to be quantified against the full framework.

5.8 Statistical Validation

Each configuration in the experimental grid is to be run with multiple random seeds (a minimum of five is recommended) to characterize variance due to initialization and mini-batch ordering. Differences between the proposed framework and each baseline are to be assessed using a paired statistical test appropriate to non-normal metric distributions, such as the Wilcoxon signed-rank test, with results reported alongside effect sizes rather than p-values alone, to avoid overstating practical significance from statistically significant but small differences.

5.9 Implementation Details

The framework is proposed to be implemented using a standard deep-learning library (e.g., PyTorch), with hyper-parameters λ_1 , λ_2 , λ_3 , the routing

thresholds τ_h and τ_l , the effective-number parameter β , and the GCE robustness parameter q selected via a held-out validation split using a bounded random or Bayesian hyper-parameter search rather than exhaustive grid search, given the size of the experimental grid in Table 3. Figure 2 summarizes the end-to-end methodology workflow, from dataset selection through final reporting.

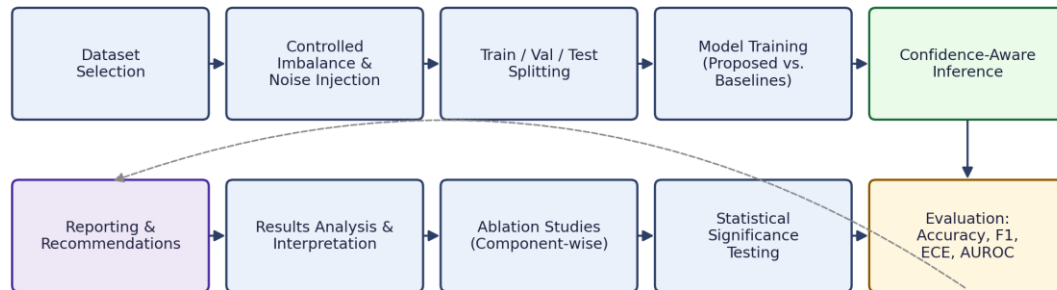


Figure Flow: Stage 1 (top row) precedes Stage 2 (bottom row, right-to-left)

Figure 2. Proposed end-to-end research methodology workflow.

6. CONCLUSION

This paper has argued that class imbalance, confidence calibration, and label-noise robustness are not independent problems but interacting facets of a single underlying question: how much should a learning system trust a given training label, and how should that trust be reflected in the confidence it reports at inference time. Through a structured review of the literature on resampling and re-weighting methods [1]–[4], calibration methods [5], [6], and noisy-label learning methods [7]–[10], this paper identified four specific gaps left by treating these problems in isolation, most notably the risk that noise-robust methods systematically mistake rare, genuine minority-class examples for mislabeled ones.

In response, this paper proposed Confidence-Aware Adaptive Learning (CAAL), a unified framework that couples a learned, per-sample confidence signal with class-frequency-aware routing between a class-re-balanced loss and a noise-robust loss, together with an explicit calibration regularizer embedded directly in the training objective. component-wise ablations, and a joint evaluation-metric suite spanning accuracy, minority-class recall, calibration error, and noise-detection quality intended to guide rigorous empirical validation of the proposed approach.

The framework, as presented, carries acknowledged limitations that should guide its evaluation. The confidence-estimation module introduces additional parameters and training overhead relative to single-branch baselines, and effective calibration regularization is expected to depend on access to a small, reliably clean held-out subset, an assumption that may not hold in every deployment setting. Future work should extend the proposed methodology in three directions. First, empirical validation across the full experimental grid proposed in Chapter 5 is required to establish whether the joint framework outperforms single-axis baselines specifically in the regime where imbalance and noise co-occur, rather than in either condition alone. Second, the framework could be extended to multi-label and streaming or online-learning settings, where label noise and class distribution may both shift over time. Third, integrating the confidence-routing mechanism with active-learning or human-in-the-loop annotation review could allow the low-confidence, suspect-noisy sample stream identified by the framework to be used constructively, directing limited human review effort toward the training examples most likely to be both informative and mislabeled.

REFERENCES

- [1] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
- [2] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.
- [3] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 2980–2988.
- [4] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 9268–9277.
- [5] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Machine Learning (ICML)*, Sydney, Australia, 2017, pp. 1321–1330.
- [6] 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)*, vol. 30, Long Beach, CA, USA, 2017.

-
- [7] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, Montréal, Canada, 2018, pp. 8527–8537.
- [8] Z. Zhang and M. R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, Montréal, Canada, 2018, pp. 8778–8788.
- [9] C. G. Northcutt, L. Jiang, and I. L. Chuang, "Confident learning: Estimating uncertainty in dataset labels," *Journal of Artificial Intelligence Research*, vol. 70, pp. 1373–1411, 2021.
- [10] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning from noisy labels with deep neural networks: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 11, pp. 8135–8153, Nov. 2023.