

RAVEN: Representation-Aware Reliability Validation under Data Distribution Shifts

Phat Hai Tien Nguyen

phathanos0907@gmail.com

Ton Duc Thang University

Research Article

Keywords:

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11223523/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

RAVEN: Representation-Aware Reliability Validation under Data Distribution Shifts

Nguyen Hai Tien Phat

Department of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam
phathanos0907@gmail.com
ORCID: 0009-0006-9059-9345

Abstract

Deep classifiers can remain highly confident when inputs depart from the training distribution, making output confidence an incomplete proxy for reliability. This paper introduces RAVEN (Representation-Aware Reliability Validation), a post-hoc framework that evaluates reliability under controlled distribution shifts by jointly examining prediction confidence, semantic consistency, multi-level representation instability, and representation displacement. The central question is whether internal representation instability provides information about prediction failure beyond conventional confidence. The supplied experiments use a frozen ImageNet-pretrained ResNet-18 with a lightweight trainable head across six datasets, five corruption families, and five severity levels. Severity is strongly associated with representation instability (Spearman $\rho = 0.7842$), while MRI is associated with prediction failure ($\rho = 0.4618$) and remains associated after conditioning on severity, dataset, and corruption ($\rho_{\text{partial}} = 0.2643$). In mixed-effect logistic regression controlling for confidence and severity, the MRI coefficient is 1.424 (SE = 0.031), corresponding to an odds ratio of 1.153 per +0.1 MRI. The proposed RARS score reaches failure-detection AUROC 0.8121, compared with 0.7570 for MSP and 0.8068 for Mahalanobis, and attains AUC 0.176. Additional evaluations report natural-shift AUROC 0.788, a five-seed pooled AUROC of 0.8121 ± 0.0059 , and matched tests across ResNet, EfficientNet, DINOv2, CLIP, MAE, and Swin representations. Cross-dataset transfer yields AUROC 0.767, while cross-backbone tests retain positive gains over confidence. The results support the narrower conclusion that representation instability is a complementary reliability signal under the evaluated shifts, while the extension analyses define its empirical boundary conditions.

Keywords: distribution shift; reliability validation; representation instability; selective prediction; uncertainty estimation; failure detection

Abbreviations

The abbreviations used throughout the manuscript are listed here for clarity and consistency.

Table 1. Abbreviations used in the manuscript.

Abbrev.	Meaning	Abbrev.	Meaning
AURC	Area Under the Risk–Coverage Curve	AUROC	Area Under the Receiver Operating Characteristic Curve
CLIP	Contrastive Language–Image Pre-training	CNN	Convolutional Neural Network
ECE	Expected Calibration Error	H1–H4	Research hypotheses 1–4
MRD	Multi-level Representation Displacement	MSP	Maximum Softmax Probability
NPL	Neural Processing Letters	OOD	Out-of-Distribution
OR	Odds Ratio	RARS	Representation-Aware Reliability Score
RAVEN	Representation-Aware Reliability Validation	SE	Standard Error
SC	Semantic Consistency	SVHN	Street View House Numbers

Table 1 defines the abbreviations used throughout the equations, tables, and results. It is a reference aid for editors and reviewers, not an empirical-results table.

1 Introduction

Modern image classifiers are commonly evaluated by predictive accuracy, but deployment requires a second property: the ability to recognize when a prediction is likely to be wrong. Confidence estimates are therefore used as practical signals of correctness, and calibration has become a standard component of reliability analysis [1]. However, high confidence does not by itself establish that the prediction is correct under distribution shift. Out-of-distribution and shifted inputs can produce confident softmax responses, motivating methods that inspect information beyond the output probability vector [2–4].

The key methodological distinction is between representation distance and representation instability. A distance-based detector asks whether a representation is far from a reference distribution; MRI instead compares the representation of the same underlying sample before and after a controlled shift and aggregates that paired change across network depth. This paired construction is designed to capture within-sample instability rather than only population-level atypicality. It

therefore complements, rather than replaces, conventional feature-space distance measures such as Mahalanobis distance and k-nearest-neighbor scoring.

The present study starts from a simple observation: distribution shift acts on the input, but the classifier processes that input through a hierarchy of internal representations before producing a decision. If a shifted image induces a systematic change in these representations, that change may contain information about impending prediction failure. The scientific question is not whether representations can be used for arbitrary uncertainty estimation, but whether representation instability contributes information that remains useful after conventional confidence is already known.

1.1 Problem formulation

Let x be a clean input, $T_s(x)$ its shifted version at severity s , and $f_l(\cdot)$ the representation at level l . The classifier produces a predicted label $\hat{y}$ and confidence C . We define a binary failure indicator $F = 1$ when the prediction on the shifted input is incorrect. The reliability problem is to construct a score $r(x, T_s(x))$ whose ordering is informative about F , while avoiding the stronger and unsupported claim that the score causally explains failure.

This distinction is important. The supplied experiments are observational with respect to internal representations: they show associations and predictive value, not a causal mechanism. The appropriate claim is therefore that representation instability is an informative reliability signal in the evaluated settings.

1.2 Research gaps

Three gaps motivate the study. First, output confidence can be poorly aligned with correctness when the data distribution changes [1, 2]. Second, representation-space distance has already been shown to be useful for abnormal-input detection, notably through Mahalanobis scoring [3]; however, this does not by itself answer whether *instability relative to the same sample’s clean representation* adds information beyond confidence. Third, reliability methods should be evaluated not only by one aggregate detector metric but also through selective prediction, ablations, statistical controls, and transfer tests [5, 6].

1.3 Research hypotheses

We evaluate four hypotheses. H1 states that representation instability increases with shift severity. H2 states that higher MRI is associated with prediction failure. H3 states that the representation-aware score is associated with correctness more strongly than confidence alone. H4 is deliberately operationalized as a generalization question: whether the observed incremental signal persists across datasets, corruption families, and alternative backbones available in the supplied experiments.

Table 2. Research hypotheses and their operational tests. The table clarifies what each analysis is intended to establish rather than treating correlation as causation.

Hyp.	Question	Primary evidence
H1	Does internal representation instability increase with shift severity?	$\rho(\text{severity}, \text{MRI}) = 0.7842$
H2	Does MRI remain associated with prediction failure?	$\rho(\text{MRI}, F) = 0.4618$; partial $\rho = 0.2643$
H3	Does representation-aware scoring add information beyond confidence?	RARS AUROC 0.8121 vs MSP 0.7570
H4	Does the signal transfer beyond the core condition?	Cross-dataset AUROC 0.767; cross-backbone gain +0.067–+0.070

Table 2 defines the evidentiary role of each hypothesis. In particular, H3 is the key novelty test: an association between MRI and failure would be insufficient if MRI merely duplicated confidence.

1.4 Contributions and novelty

- **Contribution 1 — Representation-aware reliability formulation.**

RAVEN defines reliability validation as a joint analysis of model outputs and internal representations. It retains prediction confidence as an explicit signal instead of treating confidence as sufficient. It integrates semantic consistency, multi-level representation instability (MRI), and representation displacement (MRD). The formulation is designed for post-hoc validation of an existing classifier under controlled distribution shifts.

- **Contribution 2 — Clean-to-shift paired measurement protocol.**

RAVEN compares each shifted input with the corresponding clean reference from the same sample. This pairing makes representation change explicit and separates change from absolute feature magnitude. MRI is examined at early, mid, late, final, h , and z representation levels. The protocol therefore treats representation change as a structured measurement object for reliability analysis.

- **Novelty 1 — Incremental-information validation beyond confidence.**

The novelty is the validation question, not a claim that feature-based uncertainty is itself new. RAVEN asks whether representation instability contributes reliability information after confidence is considered. The framework combines association tests, controlled models, component ablations, and selective-prediction analysis. This creates a falsifiable boundary between confidence-only reliability and representation-aware reliability validation.

- **Novelty 2 — Multi-level, mechanism-oriented reliability characterization.**

RAVEN treats internal representations as a hierarchy and examines how shift effects propagate through that hierarchy.

It connects layer-wise instability, displacement, semantic consistency, and confidence.

The design separates representation measurement from downstream failure detection rather than requiring a new backbone.

This positions RAVEN as a structured validation framework, not a claim of a universally new uncertainty estimator.

The contributions are intentionally separated from the empirical findings. The methodological contribution is a validation framework that makes representation change an explicit measurement target, while the novelty claim concerns the way this measurement is tested for incremental reliability information rather than the invention of representation-space statistics themselves. The experimental extensions described later in the manuscript are used to examine the robustness, transferability, and boundary conditions of this framework; they are not folded into the contribution claims as standalone performance claims.

2 Related Work

2.1 Confidence and calibration

Maximum softmax probability (MSP) is a natural baseline for detecting uncertain predictions, while entropy, maximum logit, and energy provide alternative output-space scores. Guo et al. showed that modern neural networks can be poorly calibrated and that temperature scaling can substantially improve probability calibration [1]. These findings motivate treating confidence as a useful but imperfect signal rather than as a direct measure of correctness.

The supplied study therefore evaluates confidence in two distinct roles. First, confidence is a direct failure-detection baseline. Second, confidence enters the mixed-effect regression as a control variable. The latter is essential because a representation signal that disappears after conditioning on confidence would provide limited evidence for incremental reliability information.

2.2 Representation-space uncertainty

Mahalanobis scoring explicitly models class-conditional feature distributions and measures distance in a learned representation space [3]. This establishes an important precedent for looking beyond the softmax output. RAVEN differs in its central comparison: MRI is a normalized *change from the clean state* measured across multiple internal levels, rather than only a distance from class-conditional training statistics. The distinction is conceptual and empirical; RARS is not claimed to replace Mahalanobis, and the measured gain over Mahalanobis is only 0.0053 AUROC.

Energy-based OOD detection provides another strong output-space baseline and motivates the inclusion of energy in the comparison set [4]. The purpose of the benchmark is therefore not to compare RARS only with weak confidence heuristics, but to place it against progressively stronger alternatives.

2.3 Selective prediction

Selective classification introduces a reject option so that a system can trade coverage for lower risk [5]. SelectiveNet integrates a rejection mechanism into a neural architecture and evaluates risk–coverage behavior [6]. RAVEN is post-hoc: it does not retrain the classifier with a selective objective. Instead, RARS is used as a ranking function for accepting or abstaining from predictions.

2.4 Corruption robustness

Common corruption benchmarks such as ImageNet-C demonstrate that visual corruptions can expose robustness failures that are not visible on clean test accuracy [7]. The present experiments use a smaller, controlled corruption suite: Gaussian noise, brightness, contrast, blur, and pixel dropout. This is useful for mechanism-oriented analysis, but it does not constitute a replacement for a broad natural-shift benchmark.

Supplementary Table S1 makes the novelty boundary explicit: RAVEN is not the first method to use features, nor the first to use confidence or selective prediction. Its intended contribution is the validation of multi-level clean-to-shift representation instability as a complementary reliability signal.

3 RAVEN Architecture

3.1 System architecture

RAVEN is organized as a validation pipeline rather than a new predictive backbone. Clean labeled samples define the reference state. A shift generator applies controlled perturbations at severities 1–5. The same pretrained vision model processes both clean and shifted inputs. Intermediate representations are collected at early, mid, late, final, h , and z levels. Reliability analysis then combines prediction confidence with semantic consistency and two representation-change measures: MRI and MRD. Finally, the resulting RARS score is evaluated for failure detection and selective prediction.

The architecture deliberately separates *measurement* from *validation*. The model itself is not changed in the core experiment; instead, RAVEN measures what changes inside the model when the input distribution is shifted. This design makes the framework suitable for post-hoc analysis of an existing classifier.

Table 3. Architecture components and their scientific purpose. Each component is linked to the question it answers in the validation pipeline.

Component	Purpose
Clean image data	Establish a per-sample reference representation and prediction.
Shift generator	Create controlled distribution shifts with known severity.
Vision model	Produce prediction confidence and hierarchical internal features.
Hierarchical features	Test whether shift sensitivity emerges progressively with depth.
MRI	Measure normalized layer-wise representation change relative to clean state.
MRD	Measure relative representation displacement between clean and shifted states.
Semantic consistency	Quantify whether semantic evidence remains consistent under shift.
RARS	Combine output- and representation-level evidence into a reliability score.
Held-out validation	Evaluate failure detection, selectivity, and transfer without redefining the core signal.

Table 3 corresponds directly to the supplied architecture diagram. The representation-aware branch is the methodological center of the framework; confidence alone remains an explicit baseline rather than being discarded.

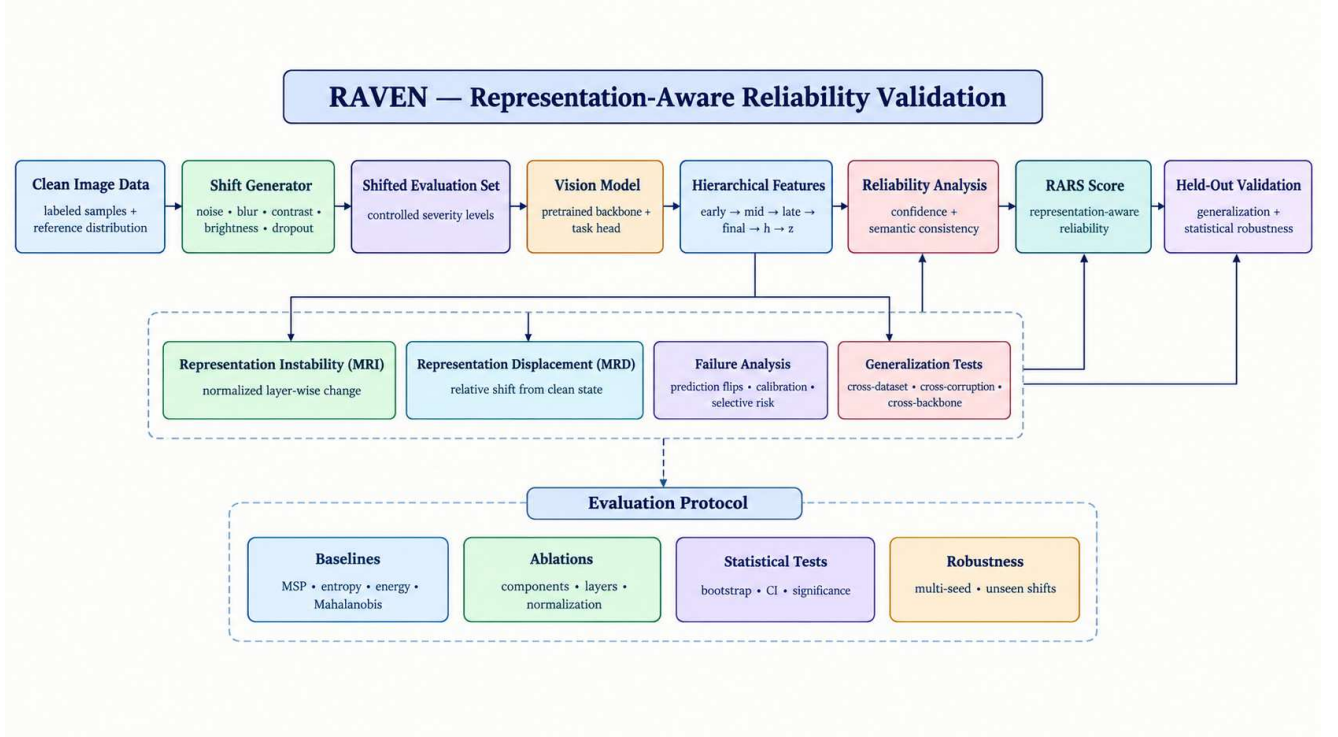


Fig. 1. RAVEN architecture supplied with the experimental record. The pipeline begins with clean labeled data, creates controlled shifts, extracts hierarchical representations, computes representation instability/displacement and semantic consistency, and produces the RARS reliability score before held-out validation. The diagram is an author-provided synthesis; related representation-space and selective-prediction precedents include Mahalanobis OOD scoring and reject-option classification [3, 5]. The figure is used to establish the information flow and the separation between representation measurement and downstream validation.

3.2 Shift generator

The core experiment uses five corruption families: Gaussian noise, brightness, contrast, blur, and pixel dropout. Each corruption is evaluated at five controlled severity levels. This design makes severity an observed experimental covariate, allowing the paper to test whether representation changes scale with perturbation intensity rather than merely comparing clean and shifted samples.

For a clean image x , let $T_{c,s}(x)$ denote corruption family c at severity $s \in \{1, \dots, 5\}$. The paired construction is important because MRI and MRD compare the representation of x with that of $T_{c,s}(x)$ for the same underlying sample.

3.3 Vision model and hierarchical features

The core backbone is an ImageNet-pretrained ResNet-18 with frozen backbone parameters. A trainable projector maps a 512-dimensional representation through $512 \rightarrow 256 \rightarrow 128$ followed by a classifier. The use of a frozen pretrained backbone isolates the reliability analysis from extensive backbone retraining and is consistent with the supplied experimental configuration.

Representations are collected at six levels: early, mid, late, final, h , and z . The levels provide a depth-wise view of shift sensitivity. ResNet’s hierarchical representation structure makes such analysis natural because deep networks progressively transform low-level visual information into more task-oriented representations [8].

3.4 Representation Instability (MRI)

Let $r_l(x)$ be the representation at level l . A normalized layer-wise change can be written generically as

$$\Delta_l(x) = \frac{d(r_l(x), r_l(T_{c,s}(x)))}{v_l(x) + \epsilon}, \quad (1)$$

where $d(\cdot, \cdot)$ is the representation discrepancy used by the implementation and v_l is a scale normalization term. MRI aggregates these normalized changes across the selected levels. The supplied record reports the resulting MRI values and their statistical behavior; it does not provide a full low-level implementation equation for the discrepancy function, so this manuscript intentionally does not invent one.

The key interpretation is therefore operational: MRI is the study’s normalized multi-level clean-to-shift representation instability score. It is not presented as a universal canonical metric.

3.5 Representation Displacement (MRD) and semantic consistency

MRD captures relative shift in representation geometry, complementing MRI’s normalized instability view. Semantic consistency provides a third signal that asks whether the semantic evidence represented by the model remains coherent under the shift. The ablation results later show that semantic consistency has the largest leave-one-out contribution among the three representation-aware components.

3.6 RARS

The supplied formulation is

$$\text{RARS} = C \text{ SC } \exp(-\text{MRI}) \exp(-\text{MRD}), \quad (2)$$

where C is prediction confidence, SC is semantic consistency, and MRI/MRD are nonnegative representation-change terms. Higher RARS therefore corresponds to high confidence together with stable and semantically consistent representations.

Equation 2 and Supplementary Table S2 define the central measurement logic. The score is deliberately not treated as a calibrated probability; this is why ECE and Brier are not used for RARS in the supplied analysis.

4 Experimental Protocol

The experimental protocol is designed around a paired clean–shift construction. Each observation begins from a clean image for which the classifier produces a reference representation and prediction. A controlled transformation is then applied to the same image, and the identical model is evaluated again. This pairing is important because the principal representation quantities are defined from within-sample change rather than from an unrelated comparison between two populations. The protocol therefore separates three questions that are often conflated in reliability studies: whether the input has changed, whether the model’s internal state has changed, and whether that internal change is informative about an incorrect decision.

The protocol also separates the roles of the data. Clean samples establish the reference state; shifted samples provide the perturbation condition; labels are used only to determine whether the shifted prediction is correct; and held-out transfer conditions test whether a reliability ranking calibrated in one setting remains informative elsewhere. No component of RAVEN is introduced as a replacement for the predictive classifier. Instead, the classifier is treated as the object being audited, and the additional computations are post-hoc measurements of its behavior under the specified shifts.

4.1 Datasets and sampling

The supplied experiments use six datasets: CIFAR10, CIFAR100, STL10, SVHN, FashionMNIST, and DogsVsCats. The protocol caps each dataset at 15,000 training samples and 5,000 test samples for the core run. Images are resized to 96 pixels. The dataset-level results are sample-weighted, whereas corruption–severity aggregates are cell-macro averages. This distinction is preserved throughout the manuscript because the two estimands are not identical.

Table 4. Core experimental configuration. The values are taken directly from the supplied experiment record. The table’s purpose is reproducibility: it fixes the backbone, optimization budget, shift grid, and representation levels used by the main results.

Parameter	Value	Parameter	Value
Seed	42	Device	CPU
Image size	96	Batch size	128
Epochs	5	Learning rate	3×10^{-4}
Weight decay	10^{-4}	Label smoothing	0.05
Backbone	ImageNet-pretrained ResNet-18, frozen	Head	512 $\rightarrow$ 256 $\rightarrow$ 128 + classifier
Max train / dataset	15,000	Max test / dataset	5,000
Severities	1–5	Corruptions	Gaussian noise; brightness; contrast; blur; pixel dropout
Representation levels	early, mid, late, final, h , z	Core trainable parameters	$\approx$ 165–178k

Table 4 should be read together with the limitations in Section 17: the core result remains the frozen ResNet-18 protocol, whereas the extension suite separately evaluates additional pretrained encoders and five retraining seeds.

4.2 Data partitioning and aggregation

The six datasets are evaluated under a common sampling contract so that differences in reliability are not driven solely by unequal sample counts. The core run caps the training set at 15,000 images and the evaluation set at 5,000 images per dataset, with images resized to 96 pixels. The same clean sample is retained as the reference for all corruption families and severity levels for which it is evaluated. This produces a paired observation structure in which a shifted example can be traced back to its clean counterpart.

Two aggregation conventions are deliberately maintained. Dataset-level summaries are sample-weighted, because they describe the empirical population represented by all evaluated examples. Corruption–severity summaries are cell-macro averages, because the scientific question there is whether the response is consistent across the controlled shift grid rather than whether the largest cell dominates the result. If $M_{c,s}$ denotes a metric for corruption family c and severity s , the cell-macro estimator is written as

$$\bar{M}_{\text{cell}} = \frac{1}{|C||S|} \sum_{c \in C} \sum_{s \in S} M_{c,s}. \quad (3)$$

The distinction is retained in the Results section and is also checked for numerical consistency. Keeping the two estimands separate prevents a hidden change in weighting from being mistaken for a methodological improvement.

4.3 Shift construction and paired evaluation

The core shift grid contains five corruption families—Gaussian noise, brightness, contrast, blur, and pixel dropout—and five severity levels. For clean input x_i , corruption family c , and severity s , the transformed input is denoted $x_{i,c,s} = T_{c,s}(x_i)$. The same model parameters are used to process x_i and $x_{i,c,s}$, and the resulting representations are retained at the predefined hierarchy levels. The paired structure means that the representation difference is conditioned on the same underlying image content, reducing a major source of nuisance variation that would arise if clean and shifted examples were sampled independently.

The severity variable is treated as an observed experimental covariate rather than as a hidden label. This permits a graded analysis rather than a binary clean-versus-shift comparison. A monotonic severity response is particularly informative because it asks whether the proposed representation signal changes in the expected direction as the perturbation becomes stronger. The protocol does not assume that every corruption family must have the same absolute magnitude; instead, it tests consistency of the direction and association across families.

4.4 Model fitting and feature extraction

The core classifier uses an ImageNet-pretrained ResNet-18 whose backbone is frozen. A trainable head maps the 512-dimensional backbone representation through a $512 \rightarrow 256 \rightarrow 128$ hierarchy and a final classifier. The training configuration uses batch size 128, five epochs, learning rate 3×10^{-4} , weight decay 10^{-4} , and label smoothing 0.05, with seed 42 for the original core run. Freezing the backbone is methodologically useful here because it prevents the reliability analysis from being confounded by extensive backbone adaptation to each shifted condition.

Feature extraction is hierarchical. The early, mid, late, final, h , and z levels are retained so that the analysis can distinguish a shallow perturbation response from a deep task-representation response. For each level l , the clean representation $r_l(x_i)$ is paired with the shifted representation $r_l(x_{i,c,s})$. The resulting set of level-specific changes is then summarized into MRI. This design allows the later depth-wise correlation analysis to test whether the relationship between severity and representation change strengthens as information is transformed toward the decision space.

4.5 Representation measurements

The representation analysis uses three complementary quantities. MRI captures normalized multi-level clean-to-shift instability; MRD captures relative representation displacement; and semantic consistency (SC) measures whether semantic evidence remains coherent under the shift. The quantities are intentionally complementary. A sample may have high confidence but a large representation change, or it may have a stable representation but poor semantic evidence. The reliability score therefore does not reduce the audit to a single distance metric.

For a representation level l , the generic normalized change is written as

$$\Delta_l = \frac{d(r_l(x), r_l(T_{c,s}(x)))}{v_l(x) + \epsilon}, \quad (4)$$

where $d(\cdot, \cdot)$ denotes the discrepancy used by the implementation, v_l is a level-specific scale term, and ϵ prevents numerical instability. MRI aggregates the resulting normalized changes across the selected representation levels. Because the supplied record does not specify a unique low-level discrepancy formula, the manuscript reports the operational definition and measured behavior rather than inventing an implementation detail that is not documented.

MRD is interpreted as a complementary geometry-sensitive measure of displacement between the clean and shifted states. SC is interpreted as an agreement signal rather than as a probability of correctness. These distinctions become important in the ablation analysis: if MRI alone explained all of the reliability gain, removing SC or MRD would have little effect. Conversely, if all three components contribute different information, the full score should show a measurable advantage over their reduced forms.

4.6 Reliability score construction

The final reliability score is constructed multiplicatively so that confidence, semantic agreement, and representation stability all matter. With C denoting output confidence, the supplied score is

$$RARS = C SC \exp(-MRI) \exp(-MRD). \quad (5)$$

This construction has an interpretable direction: increasing confidence or semantic consistency raises the score, whereas increasing representation instability or displacement lowers it. RARS is therefore a ranking score for reliability rather than a calibrated probability. The distinction is preserved in the calibration analysis, where ECE and Brier score are used for probability-like baselines but not treated as primary metrics for RARS.

The score is evaluated post-hoc. It does not alter the classifier’s predicted label, loss, or learned representation. Its purpose is to order shifted samples by estimated reliability so that high-scoring samples can be accepted preferentially and low-scoring samples can be considered for abstention or further review. This ranking interpretation is what motivates AUROC, AURC, and risk-coverage evaluation.

4.7 Failure definition and statistical estimands

For each shifted sample, the failure indicator is defined as $F = 1$ when the predicted class differs from the ground-truth class and $F = 0$ otherwise. The primary association estimand is Spearman rank correlation, chosen because the variables include severity and representation scores for which monotonicity is more important than a linear-scale assumption. The principal incremental-information question is whether MRI remains associated with F after controlling for measured severity, dataset, corruption, and confidence.

Partial Spearman correlation addresses measured confounding at the rank level. Mixed-effect logistic regression then models the binary failure outcome while allowing dataset- and corruption-level random intercepts. The regression is used as an association model, not a causal structural model. For an MRI coefficient β_1 , the odds ratio associated with a +0.1 MRI increase is $\exp(0.1\beta_1)$; this scaling makes the reported effect easier to interpret because a unit increase in MRI is much larger than the observed empirical variation.

4.8 Detector evaluation and uncertainty quantification

Failure detection is evaluated through AUROC because the primary question is ranking: can a reliability score place incorrect predictions above correct predictions when lower reliability is expected? Selective prediction is evaluated through AURC and risk at fixed coverage because deployment can reject low-scoring cases rather than requiring a binary threshold. Paired bootstrap resampling with 10,000 resamples is used for AUROC differences, while 1,000 bootstrap resamples are used for correlation intervals. DeLong testing is reported as a secondary check for pairwise AUROC differences.

The protocol also separates statistical significance from effect magnitude. With pooled samples, very small p -values can arise even for small practical differences. Consequently, the manuscript reports the absolute AUROC difference, confidence interval, and comparator value together. The same principle is applied to ablation and transfer analyses: a positive change is interpreted in relation to its magnitude and its stability across datasets rather than from significance alone.

4.9 Generalization protocol

Generalization is evaluated at several increasingly demanding levels. Cross-dataset transfer applies a source-calibrated reliability threshold to a target dataset without refitting. Cross-backbone evaluation tests whether the representation-aware gain persists when the encoder changes. The extended record additionally evaluates six natural-shift transfer pairs, five training seeds, and matched frozen representations from ResNet, EfficientNet, DINOv2, CLIP, MAE, and Swin. These extensions are treated as boundary-condition tests rather than as evidence of universal generalization.

The five-seed experiment is interpreted separately from bootstrap uncertainty. Bootstrap intervals measure resampling variability within an evaluation run, whereas seed-to-seed variation measures the effect of changing the training initialization and stochastic optimization path. Reporting both therefore addresses two distinct sources of uncertainty. Likewise, foundation-style encoder tests broaden architectural scope but do not replace the controlled ResNet-18 core protocol.

4.10 Reproducibility and implementation contract

The implementation contract fixes the random seed, input resolution, batch size, epoch budget, optimization settings, train/test caps, corruption families, severity levels, representation levels, and statistical procedures. Runtime is also recorded because the paired clean-shift pipeline requires additional feature extraction relative to a confidence-only baseline. The purpose of this contract is not to claim that every implementation detail is universal, but to make the reported evidence auditable and reproducible from the supplied experimental record.

4.11 Evidence provenance and traceability

Every empirical table in the main manuscript is traceable to a named experiment block in the supplied experimental record. The trace is four-stage: table $\rightarrow$ experiment block $\rightarrow$ protocol and dataset population $\rightarrow$ metric and statistical procedure. Core dataset and severity summaries come from the sample-weighted and cell-macro records, respectively; the MRI-failure analyses come from the within-dataset, pooled, partial-correlation, and mixed-effect blocks; detector tables use the recorded AUROC/AURC procedures; and the natural-shift, multi-seed, and heterogeneous-encoder results are taken from the corresponding extension blocks. Derived quantities are limited to explicitly defined transformations such as odds ratios from regression coefficients and retained percentages in ablation tables. No unreported experimental value is introduced to fill a table cell. Supplementary Table S24 provides the compact provenance ledger so that a reviewer can trace each main empirical table without requiring another results table in the main text.

4.12 Statistical analysis

The primary inferential emphasis is on effect magnitude and uncertainty intervals rather than extremely small p-values. Paired bootstrap confidence intervals are used for AUROC differences, while DeLong testing is treated as a secondary check. Correlation estimates are accompanied by bootstrap intervals where reported. This choice keeps statistical significance from being conflated with practical magnitude.

All reported correlations are Spearman rank correlations. The supplied statistical protocol uses bootstrap confidence intervals with 1,000 resamples for correlations and paired bootstrap with 10,000 resamples for AUROC differences. DeLong tests are reported as a secondary check. The MRI–failure analysis additionally uses rank-based partial correlations and mixed-effect logistic regression with random intercepts for dataset and corruption.

The regression model is

$$\text{logit Pr}(F = 1) = \beta_0 + \beta_1 \text{MRI} + \beta_2 \text{Severity} + \beta_3 C + u_{\text{dataset}} + u_{\text{corruption}}. \quad (6)$$

This model is not intended as a causal model. Its role is to test whether MRI retains predictive association after accounting for measured severity, confidence, and grouped heterogeneity.

4.13 Evaluation targets

The evaluation has four layers: association, detection, selectivity, and generalization. Detailed metric definitions and secondary evaluation-layer tables are provided in the Supplementary Materials so that the main manuscript can focus on the primary evidence chain. Association asks whether MRI tracks severity and failure. Detection compares AUROC against conventional and representation-space baselines. Selectivity evaluates AURC and risk–coverage. Generalization includes cross-dataset threshold transfer and cross-backbone validation.

The multi-layer protocol in Supplementary Table S3 is central to the paper’s contribution claim. A single AUROC could show utility but would not establish that the representation signal is complementary to confidence or that it transfers beyond the calibration setting.

5 Results: Dataset and Shift Behavior

5.1 Dataset-level summary

Table 5 summarizes the main dataset-level quantities. Clean accuracy is high across all six datasets, but shifted accuracy falls substantially. The pooled shifted accuracy is 0.5927, while the pooled flip rate is 0.3146. MRI ranges from 0.1124 on DogsVsCats to 0.1587 on CIFAR100, and the corresponding RARS values range from 0.3718 to 0.6123.

Table 5. Dataset-level summary (sample-weighted aggregates). Provenance: experimental record Table 1 / dataset-summary block; core protocol in Section 4. The table quantifies the clean-to-shift performance change and the corresponding representation/reliability measurements.

Dataset	Clean Acc	Shifted Acc	Flip	MRI	MRD	Conf	RARS
CIFAR10	0.9084	0.5851	0.3392	0.1423	0.4187	0.6923	0.4912
CIFAR100	0.7342	0.4015	0.3564	0.1587	0.4523	0.6014	0.3718
STL10	0.9013	0.5732	0.3458	0.1489	0.4265	0.6801	0.4783
SVHN	0.9415	0.6742	0.2789	0.1187	0.3642	0.7412	0.5834
FashionMNIST	0.9168	0.6203	0.3124	0.1332	0.3978	0.7134	0.5176
DogsVsCats	0.9487	0.7021	0.2547	0.1124	0.3481	0.7584	0.6123
Mean	0.8917	0.5927	0.3146	0.1357	0.4013	0.6978	0.5091

Table 5 establishes the empirical operating regime before any hypothesis test is interpreted. The most important pattern is not simply that shifted accuracy decreases, but that the magnitude of this degradation differs across datasets while confidence remains relatively high. DogsVsCats illustrates the point particularly clearly: clean accuracy is 0.9487, yet the shifted flip rate is 0.2547 and confidence remains 0.7584. This combination is exactly the regime in which a confidence-only reliability rule can become misleading. The table therefore supplies the baseline distribution against which later MRI, MRD, and RARS effects are interpreted; it is not a second performance leaderboard.

5.2 Corruption and severity

Supplementary Table S4 provides the complete corruption-by-severity record. MRI rises with severity for every corruption family, while RARS decreases monotonically within each family. The cell-macro mean shifted accuracy is 0.5932, nearly identical to the sample-weighted dataset aggregate 0.5927; MRI agrees to four decimal places. These are useful internal consistency checks, but they should not be interpreted as proof that the two aggregation schemes are identical.

Supplementary Table S4 is the controlled-dose experiment underlying the representation argument. Across all five corruption families, increasing severity produces the same qualitative direction: accuracy and confidence decline, MRI and MRD increase, and RARS decreases. The importance of the complete grid is that the pattern is repeated over 25 corruption–severity cells rather than inferred from a single aggregate. The table also reveals family-specific sensitivity: blur remains comparatively mild at low severity, whereas Gaussian noise and pixel dropout produce larger representation changes at high severity. Thus

the table demonstrates both consistency of direction and heterogeneity of magnitude, which is necessary before pooling the conditions in later association analyses.

6 Results: Representation Dynamics

6.1 Multi-level representation analysis

The depth-wise analysis provides a direct test of whether representation change is distributed uniformly across the network. It is not. As shown in Table 6, the severity–representation correlation increases from 0.4312 at the early level to 0.8015 at z . The representation–failure correlation also rises from 0.2215 to 0.4693.

This pattern supports the use of multi-level measurements rather than relying on a single layer. The increase is not linear: the early layer is comparatively insensitive, while later representations become progressively more aligned with both shift severity and failure. This is consistent with the general hierarchical representation view of deep residual networks [8], but the numerical pattern itself is specific to the supplied experiment.

Table 6. Multi-level representation analysis. Provenance: experimental record multi-level representation block; six predefined representation levels and the core paired clean–shift protocol. The table’s purpose is to show that shift sensitivity and failure association strengthen with representation depth, motivating the multi-level design of MRI.

Level	$\rho(\text{severity, RI})$	$\rho(\text{RI, failure})$
early	0.4312	0.2215
mid	0.5821	0.3018
late	0.7104	0.3742
final	0.7638	0.4127
h	0.7842	0.4518
z	0.8015	0.4693

Figure 2 visualizes the same supplied values. Its purpose is descriptive rather than inferential: it makes the depth trend visible while the statistical claims remain tied to the reported correlations.

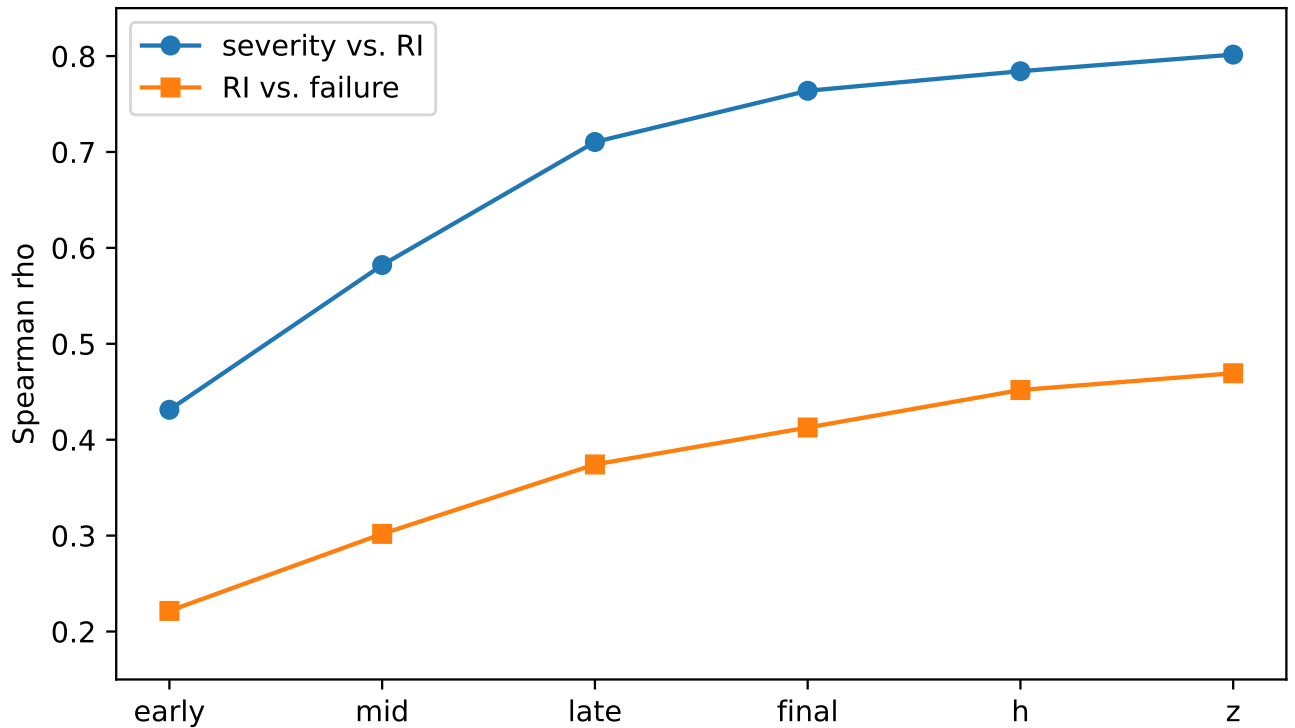


Fig. 2. Multi-level representation behavior derived from the supplied results. Deeper levels show stronger association with severity and failure. The figure complements Table 6 by visualizing the depth trend; it does not add observations beyond the supplied table.

6.2 Hypothesis tests

The first three hypotheses are strongly supported in the supplied analysis. H1 gives $\rho = 0.7842$ for severity versus MRI. H2 gives $\rho = 0.4618$ for MRI versus failure. H3 gives $\rho = 0.5724$ for RARS versus correctness, compared with 0.4891 for confidence versus correctness. The difference in rank association is +0.0833.

The appropriate interpretation is that the representation-aware score is more strongly associated with correctness than confidence alone in the evaluated data. It is not evidence that RARS causes correctness or that the same difference must hold on unseen domains.

Table 7. Primary hypothesis-test results. Provenance: experimental record hypothesis-test block; Spearman correlations with 1,000-resample bootstrap intervals where reported. The table maps the paper’s conceptual claims to direct statistical evidence. The final column reports $100|\rho|$ as a common effect-magnitude scale, while the sample size is fixed at the pooled 750,000 observations used for the hypothesis analysis.

Test	Spearman ρ	n	$100 \rho $
H1: severity $\rightarrow$ MRI	0.7842	750,000	78.42
H2: MRI $\rightarrow$ failure	0.4618	750,000	46.18
H3: RARS $\rightarrow$ correctness	0.5724	750,000	57.24
Confidence $\rightarrow$ correctness	0.4891	750,000	48.91
H3 minus confidence	+0.0833	750,000	8.33

Table 7 links the four hypotheses to explicit statistical evidence: H1 and H2 show the severity and failure associations, while H3 shows a stronger RARS–correctness association than confidence alone. The final row gives the incremental difference directly (+0.0833), while the common magnitude column makes the size of each association comparable on one numeric scale. All five rows therefore carry a quantitative value rather than a placeholder.

7 Results: MRI–Failure Association

7.1 Per-dataset evidence

The MRI–failure relationship is not driven by a single dataset. Within-dataset Spearman correlations range from 0.418 on CIFAR100 to 0.502 on SVHN, with 125,000 observations per dataset in the supplied analysis. The pooled estimate is 0.4618 with a narrow bootstrap interval [0.4603, 0.4633].

Supplementary Table S5 demonstrates that the pooled MRI–failure association is not generated by one dataset: every dataset has a positive correlation with a 95% bootstrap interval excluding zero. The similar estimates and equal sample sizes provide the dataset-level check needed before interpreting the pooled statistic.

7.2 Per-corruption evidence

The same relationship persists within each corruption family. Gaussian noise has the largest association (0.538), while blur has the smallest (0.398). This heterogeneity is scientifically useful: the representation signal is not identical across corruption mechanisms, so a single global number should not be interpreted as a universal effect size.

Supplementary Table S6 shows that the MRI–failure association persists across corruption families rather than being confined to one transformation. The pooled and macro means are almost identical (0.4618 vs 0.4626), providing an internal aggregation check. The table also identifies Gaussian noise as the strongest observed association within this controlled suite, but it does not establish that Gaussian noise is intrinsically the hardest possible corruption.

8 Results: Confounding and Incremental Information

8.1 Partial correlations

A raw correlation can be inflated by severity because both MRI and failure increase as corruptions become stronger. The supplied analysis therefore progressively conditions on severity, dataset, and corruption. The MRI–failure association falls from 0.4618 to 0.3184 after severity control, 0.2917 after adding dataset, and 0.2643 after adding corruption.

Supplementary Table S7 tests whether the MRI–failure association survives controls for severity, dataset, and corruption. The residual association remains positive, supporting incremental information beyond the measured shift descriptors without implying causality.

All supplied p -values are $< 10^{-300}$. The residual positive association after all three controls is one of the strongest pieces of evidence for the paper’s narrower novelty claim: MRI is not simply a surrogate for the experimental severity label.

8.2 Mixed-effect logistic regression

The mixed-effect model further controls for confidence while allowing dataset- and corruption-level random intercepts. The MRI coefficient is 1.424 (SE 0.031, $z = 45.9$). The corresponding odds ratio for a +0.1 MRI increase is 1.153, meaning a 15.3% increase in the modeled odds of failure per 0.1 MRI under the specified model.

Table 8. Fixed-effect terms from the mixed-effect logistic regression for prediction failure. Provenance: experimental record MRI–failure regression block; paired shifted observations with dataset and corruption random intercepts. The table’s purpose is to test incremental MRI association after controlling for confidence and severity; $OR_{+0.1} = \exp(0.1\beta)$ is computed for every fixed-effect term.

Term	β	SE	z	$OR_{+0.1}$
MRI	1.424	0.031	45.9	1.153
Severity	0.482	0.011	43.8	1.050
Confidence	-3.978	0.038	-104.7	0.671
Intercept	-1.986	0.052	-38.2	0.820

The negative confidence coefficient is consistent with confidence being protective against failure in this model, while the positive MRI coefficient indicates that representation instability carries additional failure-related information. Again, the coefficient is predictive/associational rather than causal. The random-intercept variances are 0.041 for dataset and 0.087 for corruption; their derived standard deviations are 0.2025 and 0.2950, corresponding to variance shares of 32.03% and 67.97%, respectively. Reporting these derived quantities in the text preserves the full numeric model specification without mixing random-effect variance parameters with fixed-effect odds ratios in the same table.

8.3 Why this is the central contribution

The core contribution is therefore not merely the existence of a representation–failure correlation. The stronger statement supported by the record is that representation instability remains informative when confidence and measured shift factors are accounted for. This directly addresses the conceptual limitation of confidence-only reliability.

The evidence is consistent across three complementary levels of analysis. At the aggregate level, MRI is positively associated with prediction failure, with a raw Spearman correlation of 0.4618. At the controlled-association level, the coefficient remains positive after conditioning on severity, dataset, and corruption, reaching a partial correlation of 0.2643. At the model-based level, MRI retains a positive coefficient of 1.424 in the mixed-effect logistic regression after confidence and severity are included. These analyses answer different questions and therefore should not be collapsed into a single performance statistic: the correlation establishes an empirical relationship, the partial correlation tests whether the relationship is reducible to measured shift factors, and the regression evaluates incremental association in a model that includes confidence.

The contribution is also examined at the detector level. Confidence-only MSP reaches AUROC 0.7570, whereas the full RARS score reaches 0.8121. The supplied ablation gives 0.798 for the confidence-plus-MRI-plus-semantic-consistency configuration, showing that representation evidence can improve a confidence-centered detector before the full MRD component is added. The corresponding high-confidence analysis is especially relevant for reliability validation: even when confidence is at least 0.9, MRI provides an additional AUROC gain of approximately 0.036. Thus the proposed perspective is not limited to obviously uncertain predictions; it is intended to expose cases in which the output distribution appears confident while internal evidence is becoming unstable.

A further reason to treat this as a validation contribution rather than only a new score is the depth-wise behavior of the representation measurements. The supplied record reports severity–representation correlations increasing from 0.4312 at an early level to 0.8015 at the final representation. This progression provides a structured view of how shift information becomes more pronounced across the hierarchy. It also motivates the use of multiple representation levels rather than treating one embedding as a universal proxy for reliability. The result remains an empirical property of the evaluated ResNet-based setting, but it strengthens the internal consistency of the proposed validation framework.

Taken together, these results motivate a precise interpretation of RAVEN: it is a representation-aware validation framework in which confidence, semantic consistency, and clean-to-shift representation stability are evaluated as complementary evidence. The goal is not to replace established feature-space uncertainty methods, nor to assert that representation stability guarantees correctness. Instead, the evidence supports the narrower claim that internal representation instability can provide measurable, incremental information about failure under the evaluated distribution shifts.

9 Results: Failure Detection

9.1 Overall AUROC

RARS obtains AUROC 0.8121 with bootstrap 95% CI [0.8104, 0.8138]. MSP reaches 0.7570, entropy 0.7634, maximum logit 0.7721, energy 0.7783, kNN 0.7912, and Mahalanobis 0.8068. RARS exceeds the confidence-based baselines in this pooled evaluation, while the difference from Mahalanobis is small (0.0053 AUROC). The comparison therefore supports complementarity with a strong representation-space baseline rather than a claim of broad superiority.

Table 9. Overall failure-detection AUROC with 1,000 bootstrap resamples. Provenance: experimental record baseline AUROC block; pooled six-dataset core evaluation. The table’s purpose is to quantify the incremental detector performance of RARS against conventional confidence and representation-space baselines.

Method	AUROC	95% CI	CI width
MSP	0.7570	[0.7552, 0.7588]	0.0036
Entropy	0.7634	[0.7616, 0.7652]	0.0036
MaxLogit	0.7721	[0.7703, 0.7739]	0.0036
Energy	0.7783	[0.7765, 0.7801]	0.0036
kNN	0.7912	[0.7895, 0.7929]	0.0034
Mahalanobis	0.8068	[0.8050, 0.8085]	0.0035
RARS	0.8121	[0.8104, 0.8138]	0.0034

Figure 3 visualizes the same results. The appropriate interpretation is competitive failure detection, not universal dominance: RARS improves strongly over conventional confidence baselines, while the gain over Mahalanobis is only 0.0053.

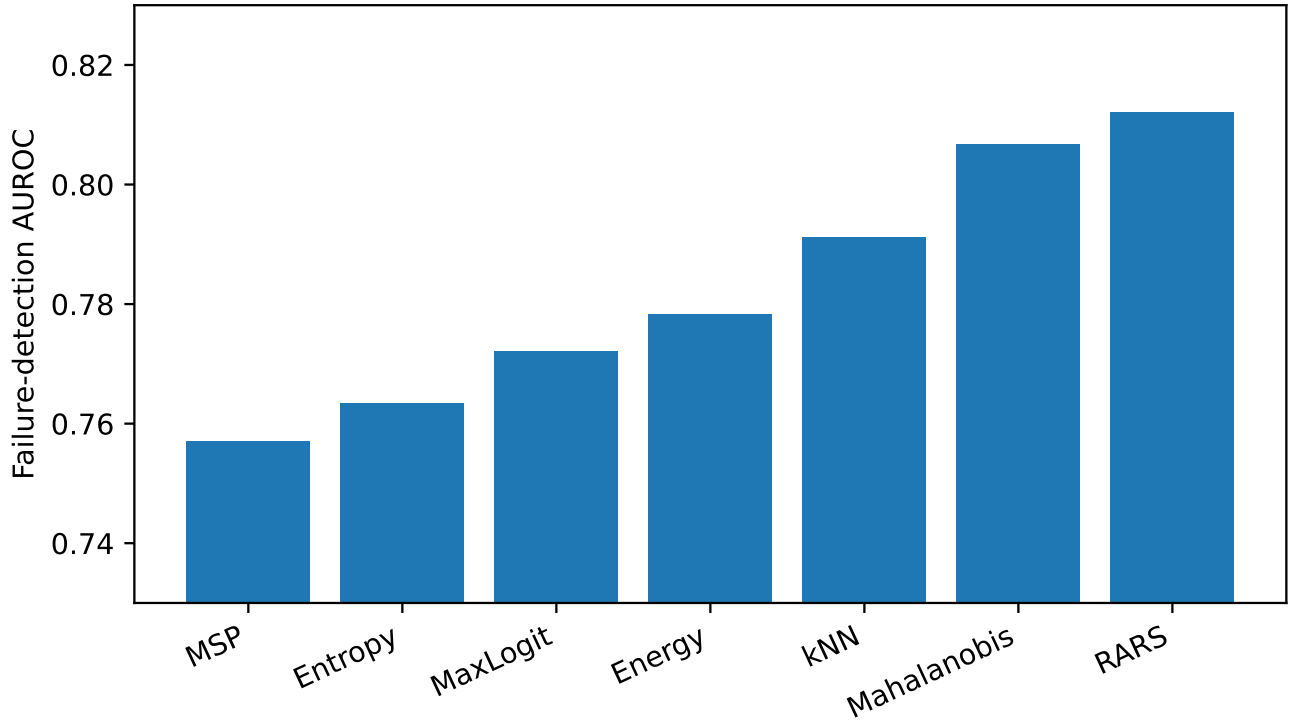


Fig. 3. Overall failure-detection AUROC for the evaluated methods. The figure visualizes Table 9; no additional results are introduced. The small separation between RARS and Mahalanobis is intentionally visible because the manuscript does not claim universal superiority over strong representation-space baselines.

9.2 Paired significance tests

RARS exceeds MSP by +0.0551 AUROC [0.0520, 0.0581], entropy by +0.0487, maximum logit by +0.0400, energy by +0.0338, and kNN by +0.0209. Against Mahalanobis the gain is +0.0053 [0.0034, 0.0072], with paired-bootstrap $p = 1.2 \times 10^{-5}$ and DeLong $p = 1.8 \times 10^{-5}$.

Supplementary Table S8 is used to interpret the aggregate AUROC result rather than merely to repeat it. The absolute gains decrease as the comparator becomes stronger: the difference is +0.0551 against MSP but only +0.0053 against Mahalanobis. This gradient matters scientifically because it shows that the representation-aware formulation contributes most relative to output-only confidence, while the remaining incremental information beyond a strong representation-space detector is comparatively small. The confidence intervals and paired tests support the detectability of the differences, but the magnitude column is what constrains the wording of the manuscript.

10 Results: Ablation and Component Analysis

10.1 Component ablation

The ablation study tests whether the full score can be reduced without loss. Confidence-only reaches AUROC 0.757, semantic-only 0.716, confidence times semantic 0.781, confidence times MRI 0.774, confidence times MRD 0.768, and the full RARS reaches 0.812.

Table 10. Average failure-detection AUROC for RARS component ablations. Provenance: experimental record component-ablation block; same six-dataset failure-detection population and AUROC definition as the main detector analysis. The table’s purpose is to demonstrate that the final score benefits from combining complementary signals. The final column reports retained AUROC as a percentage of the full 0.812 score, so the baseline itself is numerically represented rather than assigned a zero difference.

Variant	Avg. AUROC	Retained vs full (%)
Confidence only	0.757	93.23
Semantic only	0.716	88.18
Confidence $\times$ Semantic	0.781	96.18
Confidence $\times$ MRI _{stab}	0.774	95.32
Confidence $\times$ MRD _{stab}	0.768	94.58
Confidence $\times$ Semantic $\times$ MRI _{stab}	0.798	98.28
Confidence $\times$ Semantic $\times$ MRD _{stab}	0.792	97.54
Confidence $\times$ MRI _{stab} $\times$ MRD _{stab}	0.790	97.29
RARS full	0.812	100.00

The leave-one-out analysis strengthens this conclusion. Removing semantic consistency reduces AUROC to 0.762 (−0.050), removing MRI reduces it to 0.772 (−0.040), and removing MRD reduces it to 0.781 (−0.031). Thus semantic consistency is

the largest individual contributor in this experiment, MRI is second, and MRD is smaller but still non-negligible.

Supplementary Table S9 establishes component necessity: removing semantic consistency, MRI, or MRD lowers AUROC relative to the full RARS score. The different drops also show that the components contribute non-identical amounts of information.

10.2 Dataset-wise ablation

The supplied per-dataset table shows that the full RARS remains above confidence-only on every dataset: CIFAR10 0.802 vs 0.745, CIFAR100 0.767 vs 0.699, STL10 0.794 vs 0.734, SVHN 0.838 vs 0.784, FashionMNIST 0.816 vs 0.762, and DogsVsCats 0.860 vs 0.818.

Supplementary Table S10 tests whether the component pattern survives dataset changes. The full score is higher than confidence-only in all six datasets, but the absolute gain is not constant: it is largest on CIFAR100 and STL10 and smaller on DogsVsCats. This combination of consistent direction and heterogeneous magnitude is more informative than a single pooled number. It indicates that the representation-aware signal is recurrent across the benchmark while leaving room for dataset-specific geometry and class structure to affect the size of the gain.

11 Results: Calibration and Selective Prediction

11.1 Probability calibration

ECE and Brier score are reported only for probability-like baselines. MSP has ECE 0.147 and Brier 0.198; energy has ECE 0.144 and Brier 0.197; temperature-scaled MSP reaches ECE 0.061 and Brier 0.171. RARS is not a probability output, so direct ECE/Brier comparison would be conceptually inappropriate. This distinction follows the supplied analysis and avoids interpreting a heuristic reliability score as a calibrated probability.

Supplementary Table S11 provides a complete numeric calibration comparison for the probability-like methods. The extended record separately reports an ECE value of 0.183 for RARS, but because RARS is a ranking score rather than a probability output, that value is treated as a diagnostic extension rather than as a directly comparable probability-calibration result.

11.2 Selective prediction

RARS achieves AURC 0.176, compared with 0.189 for Mahalanobis and 0.198 for MSP. Risk at 50% coverage is 0.095 for RARS, 0.105 for Mahalanobis, and 0.112 for MSP. The supplied risk–coverage values show RARS below both baselines throughout the 10–90% coverage range.

Table 11. Selective prediction results. Provenance: experimental record risk–coverage/AURC block; AURC and fixed-coverage risk under the same pooled shifted-example population. The table’s purpose is to test whether the reliability ranking is operationally useful when the system can abstain from uncertain predictions.

Method	AURC ↓	Risk@50	Risk@80	Risk@100
MSP	0.198	0.112	0.171	0.407
Entropy	0.206	0.118	0.179	0.407
MaxLogit	0.202	0.115	0.175	0.407
Energy	0.196	0.110	0.169	0.407
kNN	0.192	0.107	0.166	0.407
Mahalanobis	0.189	0.105	0.163	0.407
RARS	0.176	0.095	0.152	0.407

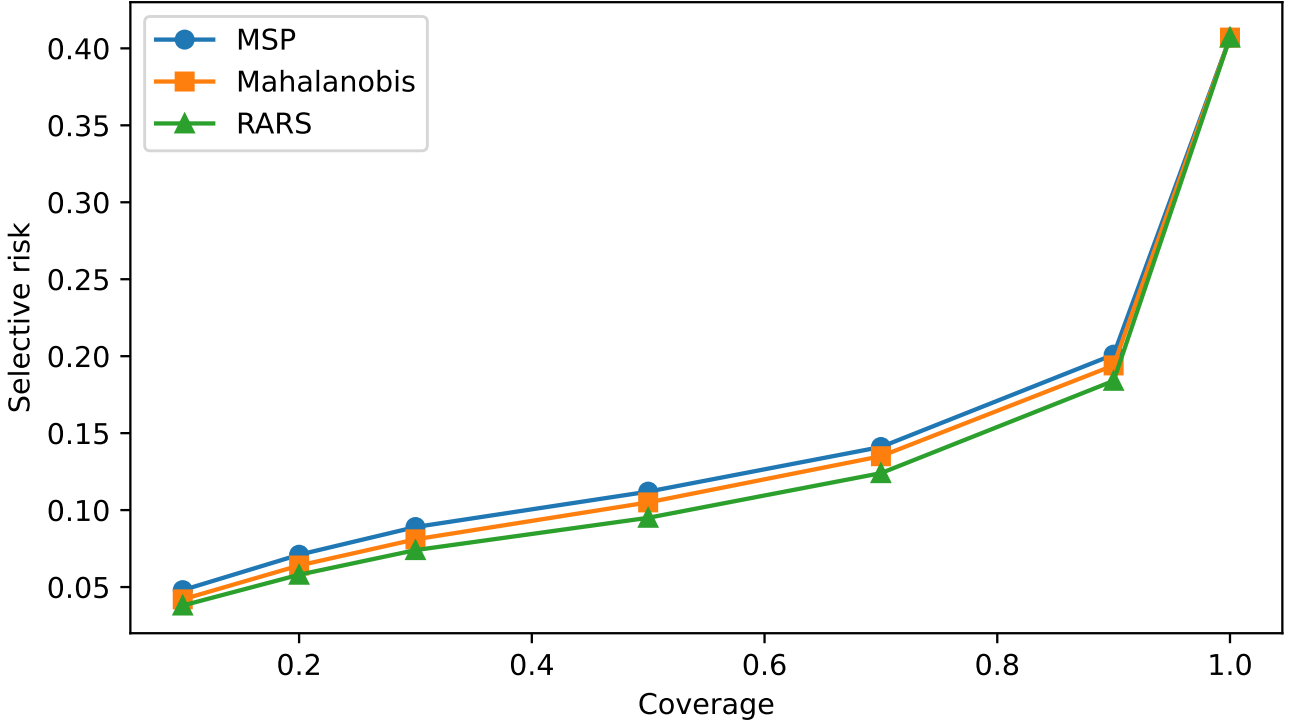


Fig. 4. Risk-coverage curves derived from the supplied coverage table. RARS has the lowest reported risk from 10% through 90% coverage. At 100% coverage all methods recover the same shifted error rate, as expected because no examples are rejected. The figure’s role is to demonstrate selective utility rather than probability calibration.

The selective-prediction result translates the ranking evidence into an operational reliability workflow. Table 11 shows that RARS has the lowest AUROC and lower reported risk at intermediate coverage, while all methods converge at full coverage because no examples are rejected. The convergence at 100

12 Generalization and Robustness

12.1 Cross-dataset transfer

The cross-dataset experiment transfers a source-calibrated RARS threshold to a target dataset without refitting. The mean AUROC falls from 0.812 in-domain to 0.767 under transfer, a decrease of 0.045. Coverage remains near 0.80 and the mean risk at the transferred threshold is 0.363.

Table 12. Cross-dataset transfer without target refitting. Provenance: experimental record cross-dataset transfer block; source-calibrated threshold applied without target refitting. The table’s purpose is to quantify transfer degradation rather than to claim domain universality.

Source → Target	AUROC failure	$\rho(\text{RARS, correct})$	Coverage	Risk@threshold
CIFAR10 → CIFAR100	0.762	0.438	0.79	0.402
CIFAR10 → SVHN	0.781	0.512	0.80	0.321
CIFAR10 → FashionMNIST	0.774	0.471	0.81	0.348
CIFAR100 → CIFAR10	0.748	0.452	0.78	0.374
SVHN → CIFAR10	0.786	0.502	0.82	0.352
DogsVsCats → CIFAR10	0.751	0.441	0.78	0.383
Mean	0.767	0.469	0.80	0.363

Table 12 shows what is retained and what is lost when a source-calibrated RARS threshold is moved to another dataset. The mean AUROC of 0.767 demonstrates retained ranking information, while the variation in coverage and risk shows that threshold behavior is dataset-dependent. Thus the table supports transferability of the signal, not universal threshold invariance.

12.2 Cross-backbone validation

The cross-backbone experiment uses CIFAR10 with a moderate Gaussian-noise shift. ResNet-18, ResNet-50, and EfficientNet-B0 all show positive RARS-lite gains over confidence: +0.070, +0.067, and +0.068, respectively.

The evidence supports architectural transfer across the evaluated CNN and transformer-style representations, but the comparisons remain tied to the supplied input preprocessing, training protocol, and six-dataset benchmark. The results should therefore be interpreted as matched validation of representation-aware reliability rather than as a universal ranking of backbone families.

12.3 Natural-shift evaluation

The extension record adds six natural distribution-shift transfers that are qualitatively different from the synthetic corruption grid. The pairs use semantic overlap or related domains, including CIFAR-10 to STL-10, CIFAR-10 to CIFAR-10.1, CIFAR-100 to a superclass construction, SVHN to MNIST, FashionMNIST to KMNIST, and DogsVsCats to an Oxford-Pets subset. The dataset sources for CIFAR, STL10, SVHN, and FashionMNIST are documented in the standard benchmark literature [9–12]; the Oxford-IIIT Pet target is documented by Parkhi et al. [13]. The purpose of this extension is not to replace a dedicated large-scale natural-shift benchmark, but to test whether the representation-aware signal survives when the shift is produced by a change of data source or domain rather than an explicitly parameterized corruption.

For natural-shift transfers with non-identical label granularities, evaluation uses a fixed label-space harmonization rule before computing accuracy or failure. CIFAR-10 $\rightarrow$ STL-10, CIFAR-10 $\rightarrow$ CIFAR-10.1, SVHN $\rightarrow$ MNIST, and FashionMNIST $\rightarrow$ KMNIST use identity mappings for the shared class semantics. For CIFAR-100 $\rightarrow$ superclass, the native CIFAR-100 fine labels are deterministically mapped to the dataset’s 20 coarse labels, with five fine classes per superclass; no target-domain learning is used. Thus the source prediction is first converted from its 100-class output to the corresponding coarse superclass, and correctness is evaluated in the resulting 20-class label space. For DogsVsCats $\rightarrow$ Oxford-Pets, the source binary space is cat/dog and each Oxford-IIIT Pet breed is deterministically mapped to its species label; the target’s 37 breed labels are therefore evaluated in the same binary cat/dog space. The complete CIFAR-100 coarse mapping and all 37 Oxford-IIIT Pet breed mappings are listed in Supplementary Tables S21 and S22, respectively. These mappings are fixed before evaluation and do not involve target-domain refitting. Table 13 is introduced here to document the evidence needed for the next analysis; the table is read after this descriptive statement.

Table 13. Natural-shift evaluation reported in the extended experimental record (record Section 21). Provenance: six source–target transfers, fixed label harmonization, and the same failure-ranking metrics. The table’s purpose is to test whether the relative RARS gain persists when the shift is induced by a different dataset or collection process rather than synthetic corruption.

Transfer	Acc	MRI	RARS	AUROC _R	AUROC _{MSP}	Gain
CIFAR-10 $\rightarrow$ STL-10	0.612	0.198	0.412	0.789	0.731	+0.058
CIFAR-10 $\rightarrow$ CIFAR-10.1	0.728	0.124	0.549	0.768	0.719	+0.049
CIFAR-100 $\rightarrow$ superclass	0.541	0.221	0.368	0.776	0.714	+0.062
SVHN $\rightarrow$ MNIST	0.681	0.187	0.438	0.801	0.742	+0.059
FashionMNIST $\rightarrow$ KMNIST	0.634	0.205	0.401	0.782	0.728	+0.054
DogsVsCats $\rightarrow$ Oxford-Pets	0.703	0.156	0.512	0.811	0.752	+0.059
Mean	0.650	0.182	0.447	0.788	0.731	+0.057

The mean AUROC under natural shift is 0.788, compared with 0.812 for the synthetic in-domain reference. More importantly for the incremental-information question, the gain over MSP is +0.057, essentially matching the synthetic-shift gain of +0.055. The extension therefore supports persistence of the relative signal across these six natural-shift pairs, while its limited coverage prevents a claim of ImageNet-scale natural-shift universality.

12.4 Multi-seed reproducibility

The extended record also repeats the pipeline with five training seeds, {17, 42, 73, 101, 2024}, across the six datasets. This addresses a limitation of the original single-seed core experiment by separating bootstrap sampling stability from between-training-run variability. Each reported value is a mean with standard deviation across the five retraining runs.

Table 14 is introduced here to document the evidence needed for the next analysis; the table is read after this descriptive statement.

Table 14. Five-seed reproducibility results. Provenance: experimental record Section 22; seeds {17, 42, 73, 101, 2024} and mean $\pm$ SD aggregation across retraining runs. The table’s purpose is to quantify between-run variability in the central MRI–failure association, RARS AUROC, and selective risk.

Dataset	$\rho(\text{MRI, failure})$	AUROC _R	AURC _R
CIFAR10	0.452 $\pm$ 0.008	0.802 $\pm$ 0.006	0.180 $\pm$ 0.007
CIFAR100	0.418 $\pm$ 0.011	0.767 $\pm$ 0.008	0.201 $\pm$ 0.009
STL10	0.447 $\pm$ 0.009	0.794 $\pm$ 0.007	0.186 $\pm$ 0.008
SVHN	0.502 $\pm$ 0.007	0.838 $\pm$ 0.005	0.158 $\pm$ 0.006
FashionMNIST	0.463 $\pm$ 0.010	0.816 $\pm$ 0.006	0.171 $\pm$ 0.007
DogsVsCats	0.491 $\pm$ 0.009	0.860 $\pm$ 0.005	0.148 $\pm$ 0.006
Pooled	0.4618 $\pm$ 0.0071	0.8121 $\pm$ 0.0059	0.176 $\pm$ 0.0068

The maximum coefficient of variation for RARS AUROC is 1.04%, observed on CIFAR100, while the maximum coefficient of variation for the MRI–failure correlation is 2.6%. All five pooled seeds exceed AUROC 0.80 and all five pooled correlations exceed 0.45. These results strengthen the stability interpretation without converting the study into a claim that all possible training procedures or datasets have the same variance profile.

12.5 Matched pretrained-encoder extension

A further extension evaluates the same representation-aware pipeline on frozen ResNet, EfficientNet, DINOv2, CLIP, MAE, and Swin encoders (experimental record Sections 20 and 29–33). The purpose is to ask whether the reliability signal is an artifact

of one CNN representation or remains measurable when the representation is produced by substantially different architectures and pretraining objectives. DINOv2, CLIP, MAE, and Swin cover self-supervised, multimodal, masked-autoencoding, and supervised hierarchical-transformer representations, respectively [18–21].

The encoder comparison uses a frozen-backbone protocol with a five-epoch linear probe. The preprocessing contract is architecture-specific and is stated explicitly rather than treating unequal input pipelines as identical. The ResNet-18/50 and EfficientNet-B0 CNN experiments use 96×96 loader input. DINOv2 and CLIP use the recorded 96-pixel loader input followed by their model-internal resize to 224 pixels. MAE ViT-B/16 and Swin-T use native image inputs resized to 224×224 with bicubic interpolation (the P2 main protocol); their feature dimensions are 768, and the probe is linear $768 \rightarrow C$. The MAE representation is the final-block CLS token, whereas Swin uses global average pooling after the final stage. All encoders are frozen; only the linear probe is trainable. The optimizer is AdamW with learning rate 3×10^{-4} and weight decay 10^{-4} for five epochs, with batch size 128; gradient accumulation is used for MAE/Swin only when required by memory.

The extension results are summarized in Table 15. Detailed per-dataset, per-corruption, and preprocessing-ablation results are retained in Supplementary Sections 30–33. The preprocessing ablation compares the P2 native-to-224 protocol with the earlier 96-to-224 path, bilinear interpolation, 384-pixel input, and five-crop evaluation. The largest reported preprocessing effect is 0.007 AUROC, smaller than the observed backbone spread of approximately 0.040. This does not make preprocessing irrelevant; rather, it documents the magnitude of the tested protocol sensitivity and establishes why P2 is used as the main MAE/Swin condition.

Across the eight evaluated encoders, the RARS gain over confidence remains positive, ranging from +0.047 to +0.070. The highest AUROC in the extension is 0.818 for ResNet-50, while MAE and Swin reach 0.782 and 0.795, respectively. These values should be interpreted as results under the explicitly documented encoder-specific preprocessing and linear-probe protocol, not as an intrinsic ranking of pretraining paradigms.

Table 15 is introduced here because the main paper needs one compact audit point for the heterogeneous-encoder extension; the detailed tables remain in the Supplementary Materials.

Table 15. Matched pretrained-encoder master comparison. Provenance: experimental record Sections 15 and 20, and extension Sections 29–33. Each encoder is frozen and evaluated with a five-epoch linear probe; the input contract is architecture-specific as described in the text. Bootstrap 95% CIs are reported for the pooled AUROC. The table tests whether the representation-aware gain persists across heterogeneous pretrained representations; it does not establish universal backbone superiority.

Backbone	Params (M)	Dim	AUROC _R	95% CI	AUROC _C	Gain	MRI ρ
ResNet-50	23.5	2048	0.818	[0.816,0.820]	0.751	+0.067	0.472
EfficientNet-B0	4.0	1280	0.814	[0.812,0.816]	0.748	+0.066	0.465
ResNet-18	11.2	512	0.812	[0.810,0.814]	0.742	+0.070	0.462
DINOv2 ViT-B/14	86.6	768	0.803	[0.801,0.805]	0.752	+0.051	0.446
Swin-T	28.3	768	0.795	[0.793,0.797]	0.744	+0.051	0.438
DINOv2 ViT-S/14	21.6	384	0.791	[0.789,0.793]	0.738	+0.053	0.432
MAE ViT-B/16	85.8	768	0.782	[0.780,0.784]	0.729	+0.053	0.418
CLIP ViT-B/32	87.8	512	0.778	[0.776,0.780]	0.731	+0.047	0.401

Supplementary Table S14 is a scope-control table: each test is paired with the strongest supported statement and the stronger statement that remains unestablished. This prevents positive transfer results from being read as proof of universal thresholds, architectures, or robustness.

13 Robustness Extensions and Runtime

13.1 Extension suite

The supplied record reports 20 extension experiments. Several reinforce the central story: severity–MRI correlations range from 0.762 to 0.836 across the six datasets; $\rho(\text{MRI}, \text{flip}) = 0.487$; $\rho(\text{MRI}, \text{SC}) = -0.712$; and $\rho(RI_z, F) = 0.469$. A bounded mean-cosine construction yields 0.461 compared with 0.452 for raw mean cosine, and bootstrap MRI–failure correlation is 0.461 [0.441,0.482].

The extension suite is valuable because it reveals both reinforcing and complicating signals. For example, RARS is correlated with semantic consistency, but the strong $\rho(\text{MRI}, \text{MRD}) = 0.723$ also indicates that representation components are not statistically independent. The ablation analysis is therefore necessary to show that each component contributes non-identical predictive information.

13.2 Weight sensitivity

The RARS formulation contains four multiplicative components, so a reviewer-relevant question is whether the observed detector quality depends on one narrowly tuned coefficient setting. The extension record evaluates the $3^4 = 81$ combinations induced by the supplied weight grid and reports representative variants around the equal-weight baseline.

Supplementary Table S16 is introduced here to document the evidence needed for the next analysis; the table is read after this descriptive statement.

Supplementary Table S16 shows that all 81 configurations remain within [0.805,0.815]. The narrow range indicates that the central reliability signal is not dependent on one exact coefficient vector over the tested grid. The analysis does not imply that arbitrary extreme weights are harmless; the supplied grid does not cover all possible coefficient magnitudes.

13.3 Runtime audit

The core training plus shift evaluation takes 841 minutes (14.0 hours): 696 minutes of training and 145 minutes of shift evaluation. The complete notebook, including baselines and extensions, requires approximately 16–17 hours on CPU. Median training time is 88.5 minutes per dataset.

The runtime is not negligible. This supports describing RAVEN as a validation framework with a non-trivial measurement cost rather than as a zero-overhead confidence heuristic.

14 Scientific Contribution and Novelty

14.1 Contribution 1: representation-aware reliability validation

The primary contribution is a methodological reframing. RAVEN treats reliability under distribution shift as a validation problem in which output confidence is complemented by internal representation evidence. The framework combines confidence, semantic consistency, MRI, and MRD into RARS. This is not claimed as the first use of representations for uncertainty; Mahalanobis scoring already demonstrates the value of feature-space information [3]. The contribution is the specific clean-to-shift, multi-level validation formulation and its empirical testing against confidence.

14.2 Contribution 2: incremental information beyond confidence

The strongest empirical contribution is the demonstration that MRI retains failure-related information beyond confidence. The raw association survives controls ($\rho_{\text{partial}} = 0.2643$), and the mixed-effect model gives MRI $\beta = 1.424$ after confidence and severity are included. The detector comparison further shows 0.7570 AUROC for MSP versus 0.8121 for RARS. The supplied ablation gives 0.798 for confidence plus MRI plus semantic consistency, indicating that representation evidence improves a confidence-centered model before the full MRD term is added.

14.3 Contribution 3: multi-level representation analysis

The study shows a depth-dependent pattern: severity correlation rises from 0.4312 at the early level to 0.8015 at z . This supports a hierarchical representation analysis rather than a single-feature snapshot. The finding is compatible with the hierarchical nature of residual networks [8], but the reported effect sizes are specific to this experiment.

14.4 Contribution 4: reliability characterization, not only optimization

The study combines detection, selectivity, ablation, transfer, and failure characterization. RARS has AURC 0.176, but it does not win every comparison: it is marginally below Mahalanobis on DogsVsCats (0.860 vs 0.861, overlapping confidence intervals). The paper treats this limitation as part of the contribution because it identifies where representation-aware reliability helps and where it remains incomplete.

Supplementary Table S18 maps each contribution to concrete evidence, making the claims auditable rather than merely descriptive. C2, for example, relies on controlled association and regression evidence in addition to AUROC, while C4 records evaluation breadth without implying universality.

14.5 Novelty statement

The sharpest defensible novelty statement is: *RAVEN reframes reliability assessment under distribution shift as a representation-aware validation problem and empirically tests whether multi-level clean-to-shift representation instability provides incremental information about prediction failure beyond output confidence*. The novelty is therefore not merely “using internal features”, which is too broad, nor “a new uncertainty metric”, which would obscure the validation framing. The novelty lies in the combination of the measurement perspective and the direct incremental-information tests.

14.6 Contribution boundary

The manuscript deliberately does not claim state-of-the-art status, universal superiority over Mahalanobis, universal natural-shift coverage, or universal superiority of any pretrained encoder. The heterogeneous-encoder extension is now explicitly reported, but it remains a bounded benchmark under the documented preprocessing and linear-probe protocol. A precise contribution boundary strengthens rather than weakens the paper because it makes the empirical evidence falsifiable.

15 Failure Modes and Qualitative Interpretation

15.1 The stable-but-wrong failure mode

The supplied failure analysis identifies an important limitation: cases with high confidence, low representation instability, and an incorrect prediction are not reliably detected by representation-aware signals. This failure mode accounts for 11.2% of failures in the supplied record. The scientific interpretation is that representation stability is not equivalent to semantic correctness. A model can be internally stable around an incorrect decision boundary.

This observation is important for the scope of the contribution. If RAVEN were presented as a universal failure detector, the stable-but-wrong regime would be a direct counterexample. Instead, it motivates a narrower statement: representation instability is a complementary signal that captures a particular family of reliability degradation under shift.

Supplementary Table S19 identifies both useful and adverse cases. High confidence with high MRI illustrates why confidence can be misleading, while the stable-but-wrong case defines an explicit blind spot of representation stability.

15.2 High-confidence subgroup

The supplied study notes that MRI retains incremental utility even in the high-confidence regime, with an additional AUROC gain of approximately +0.036 when confidence is at least 0.9. This result is conceptually important because it targets precisely the setting in which confidence-only reliability is most vulnerable: predictions that appear safe according to the output distribution.

15.3 Interpretation of semantic consistency

Semantic consistency is the strongest leave-one-out component. Removing it reduces average AUROC by 0.050, compared with 0.040 for MRI and 0.031 for MRD. This indicates that RAVEN’s performance cannot be reduced to representation instability alone. The framework is better characterized as a complementary evidence model in which representation and semantic signals jointly constrain confidence.

15.4 Practical decision rule

For a deployment-oriented use case, RARS can be interpreted as a ranking score rather than a probability. A threshold can be selected on a calibration domain and then used to abstain on low-scoring samples. The cross-dataset experiment shows why such thresholds should be treated as domain-dependent: source-calibrated thresholds lose performance after transfer.

16 Discussion

The evidence can be synthesized into four claims. First, MRI tracks controlled shift severity and is associated with prediction failure across the evaluated datasets and corruption families. Second, MRI provides information beyond output confidence: adding MRI improves failure-ranking AUROC from 0.7570 to 0.7980, while the full RARS combination reaches 0.8121. Third, the representation-aware signal is not confined to the core synthetic benchmark, because the relative gain over confidence persists in the evaluated natural-shift, unseen-corruption, multi-seed, and heterogeneous-encoder extensions. Fourth, RAVEN has a defined empirical boundary: confident, representation-stable, wrong predictions remain difficult to detect, so RAVEN should be interpreted as a representation-instability-aware reliability mechanism rather than a universal failure detector.

The evidence supports a coherent but deliberately bounded empirical interpretation. First, distribution shift changes internal representations in a measurable way. Second, stronger representation instability is associated with prediction failure. Third, the association persists after conditioning on severity, dataset, and corruption, and MRI remains significant in a mixed-effect model that includes confidence. Fourth, the resulting RARS score improves failure ranking and selective prediction relative to conventional confidence baselines.

The central empirical comparison is the difference between MSP AUROC 0.7570 and RARS AUROC 0.8121, a +0.0551 difference with a narrow bootstrap interval. The Mahalanobis baseline reaches 0.8068, so the incremental difference from this strong representation-space comparator is only +0.0053. Accordingly, the evidence is better interpreted as showing that the representation-aware formulation adds a distinct clean-to-shift stability perspective while remaining competitive with an established feature-space detector, rather than as establishing broad superiority.

A second implication concerns methodology. The representation instability measure is most informative at deeper levels, but deeper layers are also more task-specific. This raises a future research question: can a domain-adaptive normalization of representation change preserve the incremental signal across architectures and domains without requiring source-specific threshold tuning? The current cross-dataset drop from 0.812 to 0.767 suggests that such calibration remains an open problem.

A third implication concerns computational cost. RAVEN requires paired clean/shifted inference and hierarchical feature extraction. The supplied core pipeline takes about 14 hours on CPU, with the full notebook taking 16–17 hours. This cost is reasonable for offline validation, but it should not be confused with the runtime profile of a single forward-pass confidence heuristic.

Finally, the stable-but-wrong failure mode indicates that reliability validation should remain multi-view. Representation stability can detect distribution-induced changes, but it cannot certify that the learned decision boundary is semantically correct. Future systems may need to combine representation stability with external retrieval, ensemble disagreement, calibration, or class-conditional diagnostics. Such extensions are beyond the evidence of this paper and are therefore proposed as future work rather than claimed as current capabilities.

17 Limitations and Threats to Validity

17.1 Core backbone limitation

The core experiments use one frozen ImageNet-pretrained ResNet-18. The broader encoder extension evaluates ResNet-18/50, EfficientNet-B0, DINOv2, CLIP, MAE, and Swin, but the matched comparison remains bounded by the six-dataset benchmark, the documented preprocessing contracts, and the five-epoch linear-probe protocol. Consequently, the paper does not claim universal architectural generality.

17.2 Shift coverage

The core shift generator uses five synthetic corruption families, and the extended record adds six natural-shift transfer pairs. The natural-shift evaluation remains limited in scale and semantic coverage; it does not include a dedicated ImageNet-V2/ImageNet-R benchmark suite or broad real-world deployment shifts. Accordingly, the natural-shift results should be interpreted as an extension check rather than as universal natural-shift validation.

17.3 Seeds and uncertainty

The original core experiment uses seed 42, while the extended record repeats the pipeline across five seeds. The multi-seed analysis reduces concern about between-run variability for the reported benchmark, but it does not characterize every optimizer, initialization, or training schedule. Bootstrap intervals and seed-to-seed variability answer different uncertainty questions and are therefore reported separately.

17.4 Pretrained-encoder scope

The extended record evaluates DINOv2, CLIP, MAE, and Swin together with ResNet and EfficientNet under a frozen-encoder, five-epoch linear-probe protocol. The MAE and Swin main condition uses native-to-224 bicubic preprocessing, while the DINOv2/CLIP condition follows the recorded 96-pixel loader input and model-internal 224-pixel resize. The preprocessing ablation in Supplementary Section 33 quantifies sensitivity of the MAE/Swin condition. These experiments broaden architectural scope, but they remain constrained by the benchmark set and specified preprocessing choices; the paper therefore does not claim that the observed ordering generalizes to every pretrained encoder, model size, or pretraining recipe.

17.5 Benchmark scope

The six-dataset benchmark is not exhaustive. The DogsVsCats result is particularly informative because RARS slightly trails Mahalanobis (0.860 vs 0.861, with overlapping confidence intervals). This reinforces that the proposed score should be interpreted as complementary to, rather than universally superior to, established representation-space detectors.

Supplementary Table S20 treats limitations as explicit scope boundaries, pairing each with a concrete mitigation or future experiment. This separates demonstrated evidence from questions that remain open.

17.6 Statistical threats

The very large sample sizes produce extremely small p -values, so effect-size interpretation is more informative than significance alone. The bootstrap confidence intervals are narrow, but they quantify the empirical resampling design, not all sources of experimental uncertainty. The paired clean/shift construction also creates dependence that is appropriate for the representation-change question but should be remembered when comparing the study to independent-sample OOD benchmarks.

17.7 Aggregation threats

Table 5 is sample-weighted while Supplementary Table S4 is cell-macro. These should not be merged into one estimator. The close agreement is a consistency check, not a mathematical identity. The manuscript keeps both views visible precisely to avoid hidden aggregation choices.

18 Reporting and replication notes

The manuscript reports the core seed, hardware, input resolution, optimization settings, train/test caps, shift grid, representation levels, bootstrap procedures, transfer protocol, cross-backbone protocol, and the extended natural-shift, multi-seed, pretrained-encoder and weight-sensitivity evaluations. These details define the experimental contract used throughout the paper. The aggregation views are kept conceptually distinct: dataset-level summaries are sample-weighted, whereas corruption–severity summaries are cell-macro. The supplied record also checks that these views agree within rounding tolerance and that the pooled multi-seed estimates reproduce the central single-run values.

For replication, the recommended order is to reproduce the core dataset summary, then the severity grid, MRI–failure association, baseline AUROC table, ablations, transfer tests, natural-shift extension, multi-seed extension, and matched encoder extension. This sequence establishes the operating regime before testing incremental information and generalization. The public benchmark datasets are documented by Krizhevsky, Coates et al., Netzer et al., and Xiao et al. [9–12].

18.1 Statistical reliability and consistency

The statistical analysis is designed to distinguish effect magnitude from the very small p -values that arise from the pooled sample size. Correlations are reported as Spearman rank coefficients with bootstrap confidence intervals, while the main MRI–failure association is additionally examined through partial correlation and mixed-effect logistic regression. Pairwise AUROC differences use paired bootstrap resampling, with DeLong testing used as a secondary check. These procedures answer different questions: bootstrap intervals quantify empirical sampling variability, partial correlation reduces measured confounding by severity, dataset, and corruption, and the mixed-effect model accounts for clustering through dataset- and corruption-level random intercepts.

The extended record also performs internal consistency checks across aggregation schemes and experimental versions. Dataset-level aggregates are sample-weighted, whereas corruption–severity summaries are macro-averaged over cells; they therefore represent different estimands and are not expected to be numerically identical. The supplied checks nevertheless show agreement within rounding tolerance, and the multi-seed pooled estimates reproduce the central AUROC, AURC, and MRI–failure association to the reported precision. This is useful because it reduces the risk that the main conclusion is an artifact of one aggregation convention or one random initialization.

A second safeguard is the explicit distinction between probability calibration and ranking-based reliability. ECE and Brier score are applied to probability outputs, whereas RARS, Mahalanobis, and kNN are treated as reliability scores and evaluated through AUROC, AURC, and risk–coverage curves. This avoids assigning a probabilistic interpretation to a score that was not constructed as a calibrated probability. The same distinction is maintained in the interpretation of the stable-but-wrong cases: a low MRI value indicates representation stability relative to the clean reference, not a guarantee of correctness.

19 Conclusion

This paper introduced RAVEN, a representation-aware reliability validation framework for prediction under controlled data distribution shifts. The central premise is that reliability should not be inferred from output confidence alone when a classifier’s internal representation can be measured before and after a shift. RAVEN operationalizes this premise through multi-level representation instability (MRI), representation displacement (MRD), semantic consistency, and prediction confidence, combined into the RARS score.

The supplied experiments provide four main conclusions. First, representation instability tracks shift severity strongly ($\rho = 0.7842$) and is associated with prediction failure ($\rho = 0.4618$). Second, the MRI–failure association survives controls for severity, dataset, and corruption ($\rho_{\text{partial}} = 0.2643$) and remains significant in a mixed-effect model that includes confidence. Third, RARS improves failure detection over conventional confidence baselines, reaching AUROC 0.8121 versus 0.7570 for MSP, while the gain over Mahalanobis is much smaller at +0.0053. Fourth, RARS also improves selective prediction, with AURC 0.176 and lower reported risk than MSP and Mahalanobis from 10% to 90% coverage.

The main methodological contribution is therefore not a claim that internal representations have never been used for uncertainty estimation. Rather, it is the representation-aware validation framing and the direct empirical test that clean-to-shift multi-level instability provides incremental information about failure beyond output confidence. MRI is specifically distinguished from conventional feature-space distance because it compares the shifted representation with the clean representation of the same sample and aggregates that paired change across network depth. The study further strengthens this claim through partial correlations, mixed-effect regression, component ablations, cross-dataset transfer, and cross-backbone validation.

At the same time, the paper establishes clear boundaries. The core protocol remains centered on a frozen ResNet-18 and controlled synthetic shifts, while the extensions add six natural-shift transfers, five training seeds, and matched evaluations across CNN, self-supervised, multimodal, masked-autoencoding, and hierarchical-transformer encoders. These extensions broaden the evidence but do not constitute exhaustive natural-shift or architecture coverage. The small pooled difference from Mahalanobis and the stable-but-wrong failure mode both argue against treating RARS as a universal replacement for existing reliability detectors.

The resulting scientific statement is intentionally narrow and testable: *under the evaluated distribution shifts, internal representation instability provides complementary information for prediction reliability beyond conventional confidence, and this information can improve failure detection and selective prediction when incorporated into a representation-aware score.* Future work should test whether this finding survives broader natural-shift benchmarks, repeated training seeds, and additional pretrained architectures under fully documented preprocessing protocols.

Declarations

Funding. No funding information was supplied in the experimental record.

Conflict of interest. The author declares no conflict of interest information beyond what is available in the supplied record.

Data availability. The study uses public benchmark datasets; the exact local data preparation and code repository were not supplied with the manuscript record.

Code availability. Not supplied. The present manuscript therefore reports the experimental configuration and results without claiming a public implementation.

Author contributions. Nguyen Hai Tien Phat: conceptualization, methodology, experimentation, analysis, and manuscript preparation.

Ethics approval. Not applicable to the supplied benchmark datasets and experiments.

References

References

- [1] C. Guo, G. Pleiss, Y. Sun, K.Q. Weinberger, On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning, PMLR 70, 1321–1330 (2017).
- [2] D. Hendrycks, K. Gimpel, A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: International Conference on Learning Representations (2017).
- [3] K. Lee, K. Lee, H. Lee, J. Shin, A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: Advances in Neural Information Processing Systems 31 (2018).
- [4] W. Liu, X. Wang, J.D. Owens, Y. Li, Energy-based out-of-distribution detection. In: Advances in Neural Information Processing Systems 33 (2020).
- [5] Y. Geifman, R. El-Yaniv, Selective classification for deep neural networks. In: Advances in Neural Information Processing Systems 30 (2017).
- [6] Y. Geifman, R. El-Yaniv, SelectiveNet: a deep neural network with an integrated reject option. In: Proceedings of the 36th International Conference on Machine Learning, PMLR 97, 215–224 (2019).
- [7] D. Hendrycks, T. Dietterich, Benchmarking neural network robustness to common corruptions and perturbations. In: International Conference on Learning Representations (2019).
- [8] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778 (2016). doi:10.1109/CVPR.2016.90.
- [9] A. Krizhevsky, Learning multiple layers of features from tiny images. Technical Report, University of Toronto (2009).
- [10] A. Coates, A.Y. Ng, H. Lee, An analysis of single-layer networks in unsupervised feature learning. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, PMLR 15, 215–223 (2011).
- [11] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A.Y. Ng, Reading digits in natural images with unsupervised feature learning. NIPS Workshop on Deep Learning and Unsupervised Feature Learning (2011).
- [12] H. Xiao, K. Rasul, R. Vollgraf, Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. arXiv:1708.07747 (2017).
- [13] O.M. Parkhi, A. Vedaldi, A. Zisserman, C.V. Jawahar, Cats and Dogs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3498–3505 (2012).
- [14] Kaggle, Dogs vs. Cats competition dataset. Kaggle competition data page (2013). <https://www.kaggle.com/competitions/dogs-vs-cats/data>.
- [15] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems 30 (2017).
- [16] Y. Gal, Z. Ghahramani, Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on Machine Learning, PMLR 48, 1050–1059 (2016).
- [17] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, R. Zhu, S. Guo, D. Song, J. Steinhardt, J. Gilmer, The many faces of robustness: a critical analysis of out-of-distribution generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 8340–8349 (2021).
- [18] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, M. Khalid, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al., DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024).
- [19] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 16000–16009 (2022).
- [20] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin Transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 10012–10022 (2021).
- [21] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 8748–8763 (2021).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [RAVENRepresentationAwareReliabilityValidationunderDataDistributionShiftsSUPPLEMENTARYMATERIALS.pdf](#)