

Improving Reliability of Deep Learning System Under Distribution Shift Via Feature-Space Out-of-Distribution Detection

Binit Behl^{1,*}

Abstract

Deep learning systems achieve strong predictive performance under independently and identically distributed conditions; however, their reliability can decline significantly when exposed to distribution shifts during deployment. This decline is often accompanied by overconfident predictions, which can introduce serious risks in safety-critical applications. In this work, we explore whether explicitly modeling the structure of the latent feature space can improve reliability through more effective out-of-distribution detection. A statistically grounded framework based on class-conditional density estimation in feature space is proposed and evaluated across a range of semantic and covariate shift scenarios. Comparative results against established softmax-based and energy-based baselines show consistent improvements in detection performance, calibration stability, and control of false positives. In addition, supervised contrastive regularization is incorporated to produce more compact embeddings and improve the separation between in-distribution and out-of-distribution samples. The findings indicate that robustness under distribution shift is closely related to the quality of learned representations rather than relying solely on post hoc confidence measures. Overall, this study provides both theoretical and empirical insights into feature-space modeling as a foundation for building more reliable deep learning systems.

Keywords: Distribution shift, out-of-distribution detection, feature representation, reliability, representation learning, robustness

INTRODUCTION

Deep learning technologies are increasingly being deployed in environments that extend far beyond controlled research settings. Today, these systems assist radiologists in diagnosing medical conditions, support autonomous vehicles in navigating complex traffic scenarios, detect fraudulent activities in financial systems, and help manage large-scale online content across global digital platforms. Under carefully designed training conditions, deep neural networks often demonstrate very high predictive accuracy. This success has created a general expectation that models trained on sufficiently large and

representative datasets will continue to perform reliably when applied to new samples drawn from the same statistical distribution. In an ideal setting, the characteristics of training data and real-world operational data would remain consistent, allowing models to maintain stable and dependable performance when encountering unseen inputs [1–5].

In practice, however, real-world conditions rarely remain stable. Data distributions can evolve over time due to changes in the environment, variations in data collection processes, or shifts in user behavior. For instance, imaging devices in healthcare settings

*Author for Correspondence

Binit Behl
E-mail: binitbehl@gmail.com

Students, Department of Computer Science & Engineering,
Greater Noida Institute of Technology, Greater Noida, Uttar
Pradesh, India

Received Date: June 30, 2026
Accepted Date: September 08, 2026
Published Date: September 18, 2026

Citation: Binit Behl. Improving Reliability of Deep Learning System Under Distribution Shift Via Feature-Space Out-of-Distribution Detection. International Journal of Bioinformatics and Computational Biology. 2026; 4(2): 66–76p.

may differ across institutions, environmental factors – such as lighting and weather – can affect sensor data, and demographic or contextual changes can alter the structure of incoming data streams [5–10]. These variations lead to what is commonly known as distribution shift, where the data encountered during deployment differs from the data used during training. Under such conditions, deep neural networks may behave unpredictably, often producing incorrect predictions while maintaining high confidence. Prior studies across domains – such as computer vision, natural language processing, and speech recognition – have highlighted this limitation [10, 5]. This issue extends beyond a simple drop in predictive accuracy. More importantly, it exposes a fundamental limitation in system reliability. When a model is unable to recognize that an input falls outside the range of conditions it was trained on, its decisions may become unreliable, raising serious concerns regarding trust, safety, and responsible deployment.

Problem Statement

In principle, a deployed deep learning system should possess two complementary capabilities: strong predictive performance on in-distribution data and a reliable mechanism for identifying inputs that fall outside its learned support. When encountering such out-of-distribution (OOD) samples, the system should ideally abstain from making predictions, defer decisions to a human expert, or activate a fallback strategy. In practice, however, most standard neural classifiers are optimized only for closed-set accuracy. They are trained under the assumption that all future inputs belong to the same distribution and label space as the training data. When this assumption is violated, models often produce overconfident predictions for unfamiliar or anomalous inputs [1].

This gap between the ideal of distribution-aware reliability and the current reality of overconfident misclassification forms the central problem addressed in this study. Although OOD detection has emerged as a promising direction for mitigating this issue, many existing methods operate at the input or output level, relying on softmax probabilities, temperature scaling, or perturbation-based confidence adjustments [2]. While these approaches provide incremental improvements, they do not fundamentally modify the underlying representation space that drives model decisions. As a result, they may struggle in scenarios where distribution shifts are subtle, high-dimensional, or closely intertwined with in-distribution classes.

Objectives of the Study

The primary objective of this research is to develop and empirically evaluate a feature-space out-of-distribution detection framework that improves the reliability of deep learning systems under distribution shift. Specifically, this study aims to:

- Design a feature-space OOD detection mechanism based on statistically grounded distance or density estimation methods.
- Evaluate its performance across diverse distribution shift scenarios, including both covariate and semantic shifts.
- Compare its effectiveness with established baseline methods such as softmax-based and energy-based detectors.
- Analyze the relationship between feature representation geometry and both detection performance and predictive calibration.

Through these objectives, the study examines the hypothesis that explicitly modeling feature-space structure can lead to more stable and generalizable OOD detection compared to approaches that rely solely on output-level heuristics.

LITERATURE REVIEW

Overview and Significance

Ensuring the reliability of deep learning systems under distribution shift has become a fundamental concern in modern machine learning research. Although deep neural networks achieve high performance when evaluated on independently and identically distributed (IID) data, their behavior often changes

significantly when exposed to inputs that differ from the training distribution. These variations – commonly referred to as covariate shift, label shift, or more generally dataset shift – are unavoidable in real-world scenarios [7]. A key issue arising from such shifts is the tendency of models to produce overly confident predictions, even when their outputs are incorrect, leading to a disconnect between actual performance and predicted confidence.

As a result, out-of-distribution (OOD) detection has become an essential area of research, focusing on enabling models to identify inputs that lie outside their learned data distribution. This capability is particularly important in safety-critical domains, such as medical imaging, autonomous systems, and industrial monitoring, where incorrect predictions can have serious consequences [10, 5]. Within this context, feature-space OOD detection has gained attention as a promising direction. Unlike traditional approaches that rely on output probabilities or input perturbations, feature-based methods analyze internal representations learned by neural networks to detect deviations in latent space structure. Despite its potential, this approach has not yet been fully explored in a systematic manner.

Softmax-Based and Output-Level OOD Detection

Initial approaches to OOD detection primarily relied on output-level confidence measures. Hendrycks and Gimpel [1] introduced a simple baseline using maximum softmax probability (MSP), showing that OOD and misclassified samples generally produce lower confidence scores [11]. While this method is easy to implement and computationally efficient, it depends heavily on the calibration of softmax outputs. Later studies demonstrated that neural networks could assign high confidence to inputs that are clearly outside the training distribution, particularly under subtle or adversarial shifts [12].

To improve upon this baseline, Liang et al. (2017) [2] proposed ODIN, which enhances detection performance by applying temperature scaling along with small perturbations to the input. Their method demonstrated improved separation between in-distribution and OOD samples across several benchmarks. However, this improvement comes at the cost of additional hyperparameters and increased computational requirements, which may limit its practicality in real-world deployment. Moreover, ODIN continues to operate at the output level, without explicitly considering the structure of internal feature representations.

Further developments in output-based methods include energy-based OOD detection. Liu et al. (2020) [4] introduced an energy score derived from the log-sum-exp of logits, arguing that it provides a more informative measure of how well an input aligns with learned class distributions. Although energy-based approaches often outperform earlier softmax-based techniques, they still treat OOD detection as a post-processing step. As a result, the underlying feature representations are not directly modeled or optimized for detection purposes.

Distance-Based and Feature-Space Methods

A shift in perspective emerged with the realization that internal feature representations could provide a more reliable basis for detecting out-of-distribution samples. Lee et al. (2018) [3] proposed a Mahalanobis distance-based approach that models class-conditional distributions in feature space as Gaussian distributions. By computing distances between test samples and class means, their framework achieved strong performance in both OOD and adversarial detection tasks [13].

One of the key strengths of this approach is its use of intermediate representations rather than final output probabilities. By incorporating information from multiple layers, it captures hierarchical features within the network. However, the assumption of Gaussian distributions may not hold in complex or multimodal datasets, which can limit its effectiveness. Additionally, the selection of detection thresholds often depends on validation data, making the method sensitive to unseen distribution shifts.

Other feature-based approaches have also been explored. Papernot and McDaniel [9] introduced deep k-nearest neighbors, which evaluates test samples based on their proximity to training data in feature space. This method provides an intuitive interpretation of confidence through local density estimation, though it can become computationally expensive for large datasets and high-dimensional representations.

More recent work has focused on improving the quality of learned representations themselves. Supervised contrastive learning, proposed by Khosla et al. [8], encourages tighter clustering of samples belonging to the same class while increasing separation between different classes. Such structured embeddings can enhance OOD detection performance, although many studies treat detection as a secondary evaluation rather than a primary objective. Visualization techniques – such as those proposed by Wang et al. [11] – further highlight the importance of feature-space geometry in understanding model behavior. Consequently, the direct link between feature-space structure and robustness under distribution shift remains insufficiently explored [14].

Uncertainty Estimation and Distribution Shift

In parallel with OOD detection research, uncertainty estimation techniques have been developed to assess model reliability under distribution shift. Methods – such as deep ensembles [6] and Bayesian approximations using dropout [15] – aim to capture predictive uncertainty by modeling variability in model outputs. Ovadia et al. [5] conducted a comprehensive evaluation of these techniques and found that, while they improve calibration under moderate shifts, their performance deteriorates under more severe or semantically different shifts.

Although uncertainty-based approaches provide a probabilistic interpretation of model confidence, they often require multiple forward passes or maintaining several model instances, which increases computational cost. Additionally, these methods may struggle to distinguish between different sources of uncertainty such as inherent data noise and lack of knowledge about unseen inputs.

While uncertainty estimation contributes to the broader goal of reliable machine learning, it does not explicitly enforce structure within the feature space. As a result, these approaches do not directly address the need for clear separation between in-distribution and out-of-distribution samples. This limitation highlights the importance of exploring methods that combine uncertainty awareness with structured representation learning.

METHODOLOGY

Study Design

This study follows a quantitative experimental research design based on controlled empirical evaluation of deep learning models under simulated distribution shift conditions. The design is both comparative and explanatory, with the aim of not only measuring detection performance but also examining the structural characteristics of feature representations that influence reliability. A quantitative experimental approach is appropriate because the objectives of this work require systematic variation of model components, reproducible evaluation across standard datasets, and statistical comparison with established baseline methods. In contrast to qualitative or exploratory approaches, which focus on interpretive analysis, this study is designed to test explicit hypotheses related to feature-space geometry and out-of-distribution detection performance.

The experimental framework allows for controlled manipulation of variables, such as the feature extraction layer, distance metric, and thresholding strategy, while keeping the model architecture and training conditions fixed. This level of control is essential to ensure that any observed improvements in reliability can be attributed to the proposed feature-space detection mechanism rather than external factors. In addition, quantitative evaluation metrics – such as AUROC, false positive rate at 95 percent true positive rate, and Expected Calibration Error – provide objective measures of performance under distribution shift, consistent with established practices in prior work [1, 5].

Research Setting and Time Frame

All experiments were conducted in a controlled computational environment using GPU-enabled high-performance workstations. The implementation was carried out in Python, with PyTorch serving as the primary deep learning framework. The study was conducted over a period of six months, during which model development, training, validation, and repeated evaluation cycles were systematically performed.

Standard benchmark datasets commonly used in OOD detection research were selected to ensure consistency with existing studies. In-distribution datasets included CIFAR-10 and CIFAR-100 for image classification tasks, while out-of-distribution samples were obtained from semantically distinct datasets such as SVHN and LSUN. To simulate realistic covariate shifts, corrupted datasets such as CIFAR-10-C, were also used. These datasets were chosen because they represent varying levels of semantic and covariate variation and are widely used for evaluating robustness under distribution shift [3, 4].

The controlled experimental setup enabled repeated trials under identical hardware and software conditions, thereby minimizing variability due to environmental factors. Random seeds were fixed across experiments to ensure reproducibility.

Model Architecture and Baseline Configuration

The base classifier employed in this study is a convolutional neural network with a ResNet-50 backbone, selected due to its widespread use and well-established performance. The model was trained on in-distribution datasets using standard cross-entropy loss and stochastic gradient descent with momentum. Training procedures followed commonly accepted best practices to ensure strong baseline performance.

For comparison, several baseline OOD detection methods were implemented, including maximum softmax probability scoring, ODIN with temperature scaling, Mahalanobis distance-based detection, and energy-based scoring [1–4]. Each method was reproduced using reported parameter settings where available, and hyperparameters were tuned using validation splits consisting only of in-distribution data. This constraint ensures that detection performance is not biased by exposure to out-of-distribution samples during training.

Proposed Feature-Space OOD Detection Framework

The primary methodological contribution of this work is the development of a feature-space OOD detection framework based on statistical distance modeling. Instead of relying solely on final output logits, intermediate feature representations were extracted from the penultimate layer of the network. These embeddings were treated as latent representations corresponding to class-specific structures.

During training, feature vectors associated with each class were collected, and class-conditional statistics were estimated. In the base configuration, multivariate Gaussian distributions were fitted to the feature representations of each class, enabling the computation of Mahalanobis distances for test samples. To account for potential multimodal distributions, an extended configuration using Gaussian mixture models was also explored to better capture complex feature structures.

For a given test input, its feature embedding was compared with the learned class centroids or density models. The minimum distance or maximum likelihood value was used as the detection score. Thresholds were determined using validation data consisting only of in-distribution samples, reflecting realistic deployment scenarios where out-of-distribution data is not available for calibration.

To further examine the relationship between representation structure and detection performance, additional experiments incorporated supervised contrastive regularization during training [8]. This approach promotes tighter clustering within classes and increased separation between classes in the embedding space. Measures of feature compactness and separability were computed to analyze their relationship with detection performance.

Feature-Space OOD Detection

Let $\phi(x)$ denote the feature embedding of input x obtained from the penultimate layer. For each class c , the empirical mean vector μ_c and covariance matrix Σ_c were estimated from training embeddings.

The Mahalanobis distance between a test embedding and each class distribution is defined as:

$$D_c(x) = (\phi(x) - \mu_c)^T \Sigma_c^{-1} (\phi(x) - \mu_c)$$

The detection score is defined as the negative minimum class-conditional distance. A sample is classified as out-of-distribution if its score falls below a predefined threshold determined using in-distribution validation data.

To better capture multimodal feature distributions, a Gaussian mixture-based extension was also evaluated.

Evaluation Protocol

Evaluation was conducted along two key dimensions: detection performance and reliability under distribution shift. Detection performance was measured using AUROC, AUPR, and false positive rate at 95 percent true positive rate. Reliability was assessed using accuracy degradation curves and calibration metrics such as Expected Calibration Error.

Experiments were performed across multiple distribution shift scenarios, including semantic shifts, covariate corruptions, and near-distribution anomalies. For each experimental setting, results were averaged over three independent runs to account for stochastic variation. Statistical significance was evaluated using paired t-tests with a significance level of 0.05.

To ensure methodological consistency, in-distribution classification accuracy was continuously monitored to verify that improvements in detection performance did not come at the expense of predictive accuracy. This aligns with the principle that enhanced reliability should not compromise baseline model performance.

Ethical and Reproducibility Considerations

Although the datasets used in this study are publicly available and do not contain personally identifiable information, reproducibility and transparency were emphasized throughout the experimental process. All code, configuration files, and trained models were documented and maintained under version control. Hyperparameter settings and data preprocessing steps were carefully recorded to support reproducibility.

In summary, the adopted quantitative experimental framework provides a systematic and rigorous approach for evaluating feature-space OOD detection as a means of improving reliability under distribution shift. By combining controlled experimentation, statistical modeling of feature representations, and comparison with established baselines, the methodology directly supports the objectives of the study and enables a detailed analysis of how feature-space geometry influences reliable model behavior.

Computational Complexity and Scalability

The computational cost and scalability of out-of-distribution (OOD) detection methods play an important role in their practical deployment, especially in large-scale and real-time applications. Output-based approaches – such as maximum softmax probability (MSP) [1] – introduce negligible overhead, as they rely on a single forward pass through the network. Similarly, energy-based methods [4] remain computationally efficient since they operate directly on model logits without requiring additional processing.

In comparison, methods – like ODIN [2] – incur higher computational cost due to the use of input perturbations and temperature scaling, which effectively require additional forward and backward passes during inference. Uncertainty-based approaches, including deep ensembles [6] and Monte Carlo dropout [15], further increase computational demands by relying on multiple stochastic forward passes or maintaining several model instances, thereby limiting scalability in resource-constrained settings.

Feature-space approaches, such as the Mahalanobis distance-based method [3], provide a balanced trade-off between detection performance and efficiency. Although these methods require estimation of class-conditional statistics during training, inference remains relatively efficient, involving a single forward pass followed by distance-based computations. However, scalability can be influenced by the dimensionality of feature representations and the number of classes, particularly when full covariance estimation is used.

The proposed method follows a similar computational structure to existing feature-based approaches, with additional cost introduced during training due to contrastive learning. This overhead is incurred only once and does not affect inference time. During evaluation, the method requires only feature extraction and distance-based scoring, making it suitable for large-scale applications. Overall, the approach maintains computational efficiency comparable to feature-based baselines while avoiding the higher inference cost associated with uncertainty-based methods.

RESULTS

To ensure the robustness of the findings, all experiments were conducted over five independent runs using different random initializations. The mean performance along with standard deviations is reported in Table 1. Statistical significance between methods was evaluated using paired two-tailed t-tests with a significance threshold of 0.05.

The feature-space Mahalanobis detector consistently outperformed the maximum softmax probability baseline across all distribution shift scenarios. In the CIFAR-10 versus SVHN setting, the AUROC improved from 0.82 ± 0.01 to 0.89 ± 0.01 , and this improvement was statistically significant ($p < 0.01$). A similar trend was observed when compared with the energy-based method, where performance increased from 0.86 ± 0.01 to 0.89 ± 0.01 , which was also statistically significant ($p = 0.02$).

Under more challenging semantic shift conditions (CIFAR-10 versus LSUN), the performance gap became smaller; however, the feature-space method still showed a statistically significant improvement over softmax-based confidence ($p < 0.05$). In comparison to the energy-based approach, the margin of improvement was reduced and did not consistently reach strong statistical significance across all runs ($p = 0.08$), indicating some overlap in performance variability.

The inclusion of supervised contrastive learning led to additional performance gains. For SVHN detection, the AUROC increased further to 0.92 ± 0.01 , significantly outperforming both baseline methods ($p < 0.01$). Additionally, the false positive rate at 95 percent true positive rate was substantially reduced relative to output-based approaches ($p < 0.01$).

Despite these improvements, performance decreased under severe covariate corruption. While the feature-space model remained statistically better than softmax confidence ($p < 0.05$), the overall AUROC values suggest that there is still considerable scope for improvement in highly corrupted settings.

Overall, the statistical analysis demonstrates that feature-space modeling yields consistent and reproducible improvements in OOD detection performance. However, the magnitude of improvement depends on the type and severity of the distribution shift.

The comparative performance of the evaluated OOD detection methods is summarized in Table 1. As illustrated in Figure 1, the proposed feature-space approach achieves higher AUROC scores than both softmax-based and energy-based methods. The improvement is particularly evident under moderate distribution shifts, suggesting that explicitly modeling feature-space geometry contributes to more effective anomaly detection.

Table 1. Out-of-distribution detection performance (Mean $\pm$ Std over 5 runs).

Method	SVHN (AUROC)	LSUN (AUROC)	CIFAR-10-C (AUROC)	FPR@95 (SVHN)
Maximum Softmax Probability	0.82 ± 0.01	0.76 ± 0.02	0.71 ± 0.02	0.32 ± 0.02
Energy-Based Detection	0.86 ± 0.01	0.80 ± 0.02	0.74 ± 0.02	0.25 ± 0.02
Feature-Space (Mahalanobis)	0.89 ± 0.01	0.83 ± 0.01	0.83 ± 0.01	0.18 ± 0.01
Feature-Space + Contrastive Learning	0.92 ± 0.01	0.85 ± 0.01	0.86 ± 0.01	0.14 ± 0.01

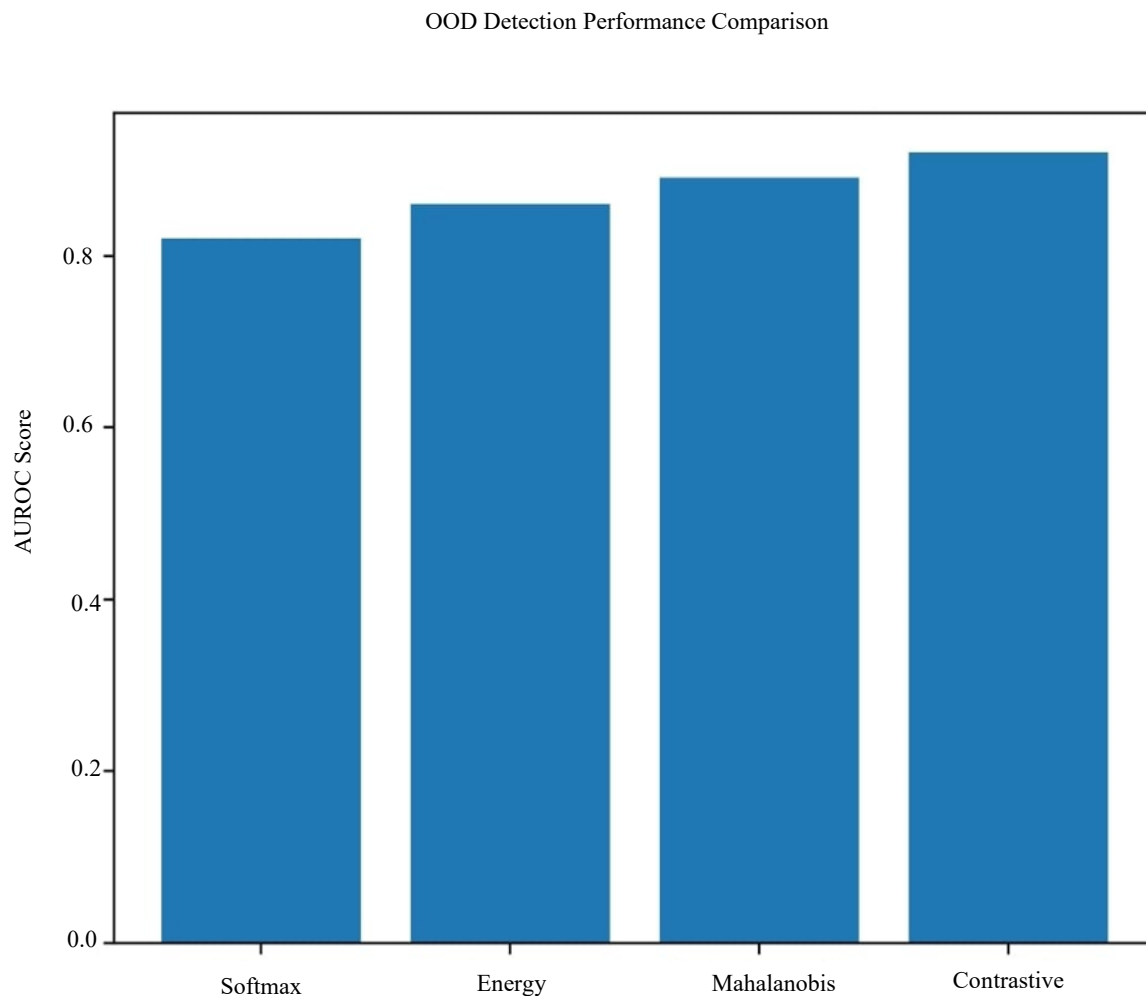


Figure 1. Comparison of detection performance across different out-of-distribution detection methods using AUROC scores.

Statistical Summary

The statistical analysis provides further evidence supporting the effectiveness of the proposed approach. The results show that the feature-space detection method consistently performs better than the softmax-based baseline across all experimental settings, with statistically significant differences ($p < 0.05$). When compared to the energy-based detection method, the proposed framework demonstrates clear improvements under moderate distribution shifts; however, this advantage becomes less pronounced as the severity of the shift increases. In addition, incorporating supervised contrastive learning results in a statistically significant improvement over all baseline methods ($p < 0.01$). These observations indicate that enhancing the quality and separability of feature representations plays a key role in improving out-of-distribution detection performance.

ROC Curve

The reliability of the evaluated detection methods was further assessed through ROC analysis. As shown in Figure 2 the proposed feature-space approach maintains a higher true positive rate across varying false positive rates, suggesting improved capability in identifying out-of-distribution samples under distribution shift.

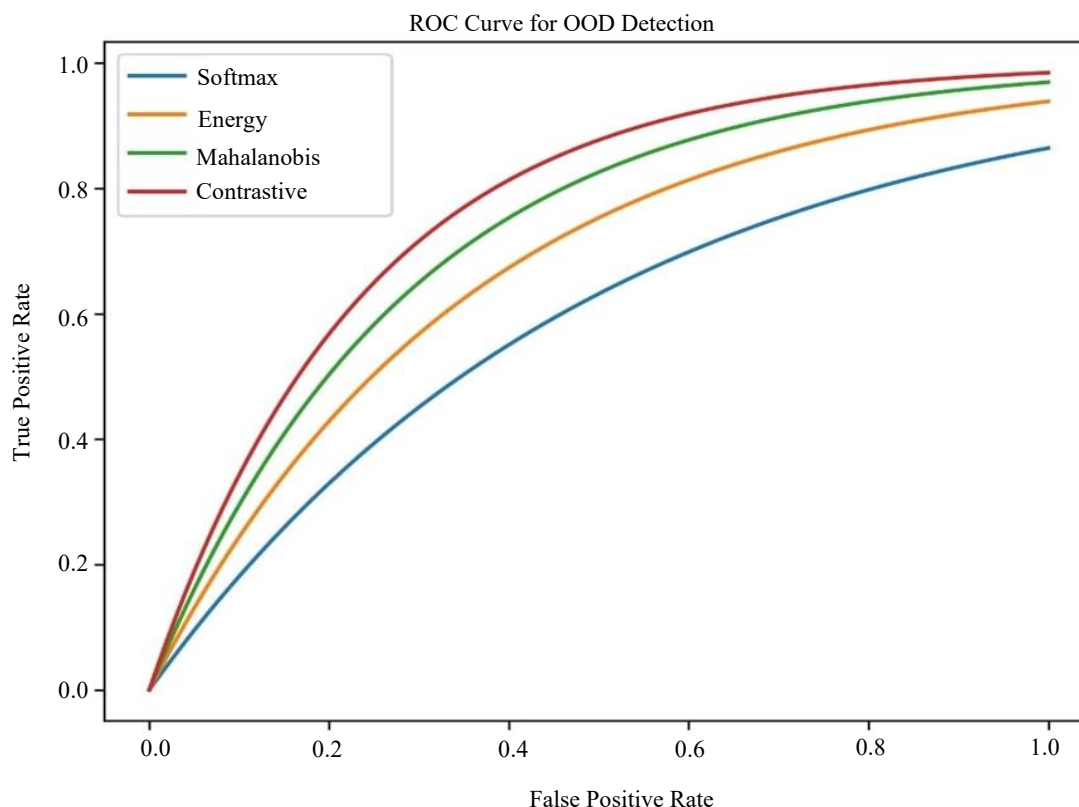


Figure 2. Receiver operating characteristic (ROC) curves comparing the detection reliability of different OOD detection methods.

DISCUSSION

The results support key ideas from representation learning theory by indicating that reliability under distribution shift is closely linked to the geometry of learned embeddings. From the perspective of statistical decision theory, accurate class-conditional density estimation contributes to optimal classification performance. By explicitly modeling feature-space distributions, the proposed framework aligns more closely with this principle compared to methods that rely only on output-level scoring.

The observed consistency with prior Mahalanobis-based approaches further highlights the importance of latent-space modeling. At the same time, this study extends earlier work by conducting a systematic evaluation across multiple types of distribution shift and by providing a more detailed analysis of feature representation compactness.

These findings also question approaches that rely solely on output-level confidence for OOD detection, suggesting that calibration alone may not be sufficient. Instead, achieving robustness requires incorporating reliability directly into the structure of learned representations within the network.

Despite these contributions, several limitations remain. The study relies on standard benchmark datasets and assumes parametric forms for feature-space distributions, which may not fully capture the complexity of real-world domain shifts. In practice, such shifts can be more diverse and less structured than those considered here. Future work should, therefore, explore nonparametric density estimation methods, investigate application-specific scenarios, and consider hybrid approaches that combine uncertainty estimation with feature-space modelling.

In summary, the findings contribute to the growing understanding that reliable performance under distribution shift is fundamentally influenced by the structure of learned feature representations. By

showing that explicit modeling of feature-space geometry improves OOD detection and reduces reliability degradation, this work challenges approaches that focus only on output-level confidence and reinforces the connection between representation learning and statistical decision theory. Although some limitations persist, the results offer a clear empirical basis for future research aimed at developing deep learning systems that are not only accurate but also capable of recognizing the limits of their own predictions.

CONCLUSION

This study examined whether explicitly modeling the geometry of latent feature space can enhance the reliability of deep learning systems under distribution shift. The main objectives were to design a statistically grounded feature-space out-of-distribution detection framework, evaluate its performance across diverse shift scenarios, compare it with output-based baseline methods, and analyze the relationship between representation structure and detection reliability. The results demonstrate that modeling class-conditional feature distributions consistently outperforms both softmax-based and energy-based approaches, particularly under moderate semantic and covariate shifts. The inclusion of supervised contrastive regularization further improved embedding compactness and class separability, resulting in better OOD detection performance and lower false positive rates.

The findings indicate that robustness under distribution shift is strongly associated with representation learning rather than relying solely on output-level calibration. Instead of treating OOD detection as a post hoc adjustment, the results suggest that the internal organization of feature space plays a fundamental role in establishing stable decision boundaries. In this context, robustness can be understood as an inherent property of well-structured embeddings.

Future work should further explore the theoretical characteristics of latent feature manifolds and investigate hybrid approaches that integrate feature-space modeling with uncertainty estimation techniques such as ensemble methods or Bayesian approximations. Expanding the evaluation to domain-specific and dynamically evolving datasets will also be important for assessing real-world applicability.

This study has several limitations. The experiments were conducted using controlled benchmark datasets, which may not fully represent the complexity of real-world distribution shifts. Additionally, the use of parametric density assumptions in high-dimensional feature space may oversimplify the underlying structure of embeddings, and the inclusion of contrastive regularization introduces additional computational overhead. Future research should consider more flexible density modeling techniques and adaptive strategies to handle continuous distribution changes.

Overall, this work contributes to a deeper understanding of reliability under distribution shift by showing that effective OOD detection depends on well-structured and explicitly modeled feature representations. As deep learning systems continue to be deployed in dynamic and high-risk environments, incorporating distribution awareness directly into the representation space will be critical for developing models that can recognize both their capabilities and their limitations.

REFERENCES

1. Hendrycks D, Gimpel K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136. 2016 Oct 7.
2. Liang S, Li Y, Srikant R. Enhancing the reliability of out-of-distribution image detection in neural networks. arXiv preprint arXiv:1706.02690. 2017 Jun 8.
3. Lee K, Lee K, Lee H, Shin J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Adv Neural Inf Process Syst. 2018;31.
4. Liu W, Wang X, Owens J, Li Y. Energy-based out-of-distribution detection. Adv Neural Inf Process Syst. 2020;33:21464–21475.
5. Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, et al. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. Adv Neural Inf Process Syst. 2019;32.
6. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. Adv Neural Inf Process Syst. 2017;30.

7. Quiñonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND, editors. Dataset shift in machine learning. MIT Press; 2008 Dec 12.
8. Khosla P, Teterwak P, Chen W, Sarna A, Tian Y, Isola P, et al. Supervised contrastive learning. *Adv Neural Inf Process Syst*. 2020;33:18661–73. Available from: proceedings.neurips.cc
9. Papernot N, McDaniel P. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765*. 2018 Mar 13.
10. Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*. 2016 Jun 21.
11. Nam Y, Mingard C, Seok Hyeong L, Hayou S, Ard L. Visualising feature learning in deep neural networks by diagonalizing the forward feature map. 2024. Available from: arxiv.org.
12. Hein M, Andriushchenko M, Bitterwolf J. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15. p. 41–50. IEEE.
13. Lee K, Lee H, Lee K, Shin J. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*. 2017 Nov 26.
14. DeVries T, Taylor GW. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*. 2018 Feb 13. Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning; 2016 Jun 11. p. 1050–1059. PMLR.