

A Comprehensive Review On Self Supervised Failure Prediction For AI Models Under Unseen Distribution Shifts

Shivika Singh*

Computer Science Engineering, VIT Vellore, Tamil Nadu, India

ABSTRACT

Artificial intelligence systems are increasingly deployed in environments where the data encountered after deployment may differ substantially from the data used during model development. A central reliability question is therefore whether a model can identify signals that indicate an increased probability of its own future failure before an incorrect decision is made. This review examines self supervised failure prediction as a theoretical and practical framework for answering that question under unseen distribution shifts. The discussion connects self supervised representation learning, uncertainty estimation, calibration, out of distribution detection, selective prediction, anomaly detection, test time training, test time adaptation and robust decision making. Self supervised objectives are particularly valuable because they can exploit unlabelled observations available during deployment and can provide auxiliary signals about reconstruction quality, transformation consistency, representation stability and temporal coherence. These signals may complement predictive confidence, which can remain misleading when neural networks are exposed to unfamiliar inputs. The review further examines the distinction between aleatoric uncertainty, epistemic uncertainty and distributional uncertainty, and explains why no single score should be treated as a complete failure predictor. A conceptual framework is proposed in which the model observes an incoming sample, evaluates representation and prediction stability, estimates uncertainty, detects evidence of distributional novelty and converts these signals into a calibrated failure risk. Important limitations include false alarms, adaptation induced error, feedback loops, computational cost, changing environments and the absence of reliable failure labels. Future research should therefore focus on jointly learning prediction and failure awareness, evaluating models across realistic unseen shifts, establishing standardized early warning metrics and developing safety oriented mechanisms that can abstain or request human review when failure risk becomes high.

Keywords: self supervised learning, failure prediction, distribution shift, uncertainty estimation, out of distribution detection, test time adaptation, trustworthy artificial intelligence.

INTRODUCTION

Modern artificial intelligence systems are commonly evaluated under an assumption that training data and future data are sufficiently similar. In mathematical terms, a conventional learning problem often assumes that observations used for training and observations used for evaluation arise from closely related probability distributions. Real deployments rarely satisfy this assumption for long periods. Sensors change, environments evolve, users behave differently, acquisition devices are replaced, operating conditions drift and rare events appear that were not represented during model development. Under these circumstances, a model may retain a high

numerical confidence while its actual probability of being correct decreases. The central reliability problem is therefore not only how to improve predictive accuracy, but also how to recognize when the model is entering a region in which its learned knowledge is inadequate [1, 2]. The question posed by this review is whether an artificial intelligence model can predict its own failure before that failure becomes visible. Here, failure prediction refers to estimating the probability that a future prediction, decision or action will be incorrect or unsafe. This is different from ordinary classification because the model must reason about the reliability of its own output. A useful failure predictor must therefore combine evidence about the input, the learned representation, the

Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

prediction and the changing data environment [3,4]. Self supervised learning offers an important route toward this objective because it creates learning signals from the input itself rather than requiring a human label for every observation. A model can be trained to reconstruct masked information, predict transformations, maintain agreement between different views, recover temporal order or identify whether two representations should be consistent. These auxiliary tasks can make internal representations sensitive to structure that is not directly encoded by the final task label. At deployment time, the same principles can be used to measure whether a new observation behaves like the examples that shaped the learned representation. Test time training showed that an unlabeled test sample can itself be converted into a self supervised learning problem, providing a mechanism for adaptation under distribution shift [5]. The connection between self supervision and failure prediction becomes especially

important when distribution shifts are unseen. Research has shown that neural networks can become overconfident on unfamiliar inputs, demonstrating why raw softmax confidence cannot be treated as a universal reliability signal [6]. When the input distribution changes, the learned approximation may become poorly supported by observed training data. Epistemic uncertainty can increase when the model lacks knowledge about the relevant region of the input space, while aleatoric uncertainty represents irreducible variability in observations. Distributional uncertainty adds another layer by asking whether the current environment itself is outside the range for which the model was validated [7]. This perspective is relevant to medical imaging, autonomous systems, industrial monitoring, language models, robotics, cybersecurity and other applications where a silent error may be more costly than a visible loss of confidence [8].

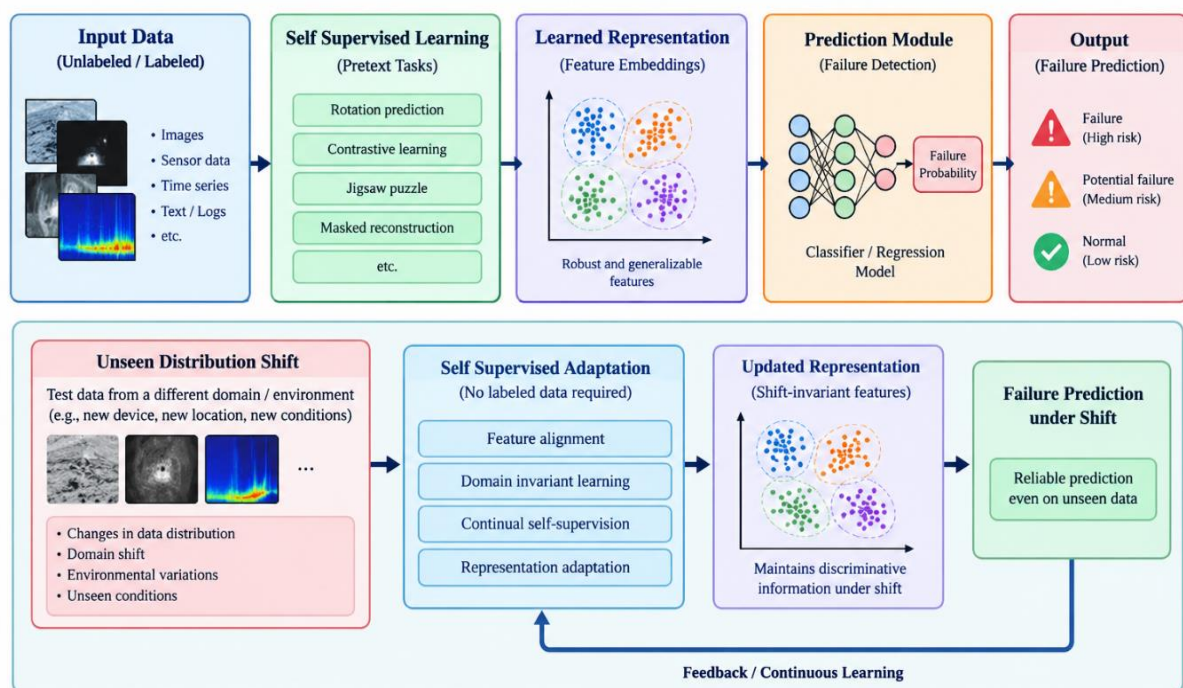


Figure 1. Conceptual framework for self supervised failure prediction under unseen distribution shifts.

2. Meaning of Model Failure and Unseen Distribution Shift

Model failure should be defined operationally rather than philosophically. In a classification system, failure may mean an incorrect class assignment. In regression, it may mean an error above a clinically or operationally meaningful threshold. In control, failure

may mean that an action violates a safety constraint. Instead, the target should be a risk variable that can be mapped to the application specific outcome of concern. This distinction is essential because a model can have excellent average accuracy while still producing rare but consequential failures. Label shift concerns changes in the outcome frequencies.

Concept or conditional shift occurs when the relationship between input and outcome itself changes. In practice these categories can coexist. For example, a medical imaging system may encounter a new scanner, a different patient population and a changed disease prevalence at the same time. A failure predictor that monitors only input statistics may therefore miss a shift in the underlying relationship between observations and outcomes [9, 10]. Unseen shifts are more difficult than familiar shifts because the system cannot rely on a predefined target domain. If all possible future environments were known, one could collect representative data, adapt the model and evaluate the adaptation. Under an unseen shift, the model must identify that its operating conditions have changed even when the precise cause is unknown. This makes failure prediction partly an open world problem. Out of distribution detection provides one solution by estimating whether an input is unlike the training distribution, but the boundary between in distribution and out of distribution data is not always aligned with predictive correctness [11]. A useful theoretical distinction is therefore between novelty and unreliability. Novelty is evidence that an observation differs from known data. Unreliability is evidence that the model is likely to make an error.

Novelty can increase unreliability, but the relationship is probabilistic rather than deterministic. A robust model may correctly classify a novel observation, while a highly familiar observation may still be misclassified because of label ambiguity, corrupted features or systematic bias. Self supervised signals are valuable because they can provide additional information about whether the model has internally coherent evidence for its decision [12].

3. Self Supervised Learning as a Source of Failure Signals

Self supervised learning constructs targets from the data itself. Instead of asking a human to provide a label, the learning system creates a pretext objective such as reconstructing a masked region, predicting a transformation, matching two augmented views or predicting temporal relationships. Contrastive learning demonstrated that representations can be learned by encouraging agreement between related views and separation from unrelated examples. Momentum based contrastive learning and bootstrap methods further showed that strong representations can be learned without conventional labels [13].

Signal	What it measures	Potential failure indication	Main limitation
Predictive confidence	Probability assigned to the selected output	Low confidence may indicate ambiguity	Neural networks can be confidently wrong [14]
Representation novelty	Distance from learned representation space	Large novelty can indicate unfamiliar input	Novelty does not always mean failure [15]
Self consistency	Agreement across	Large disagreement suggests	Sensitive to choice of [16]
	transformations or views	instability	Transformations [17]
Uncertainty	Epistemic and aleatoric uncertainty	High uncertainty can signal weak knowledge	Uncertainty estimates can be miscalibrated [18]
Distribution score	Similarity between current and reference distributions	Large shift can indicate elevated risk	Small shift can still hide label changes [19]

Temporal instability	Change in predictions across nearby observations	Abrupt changes may precede failure	Can produce false alarms in dynamic settings [20]
Selective risk	Observed error among accepted predictions	Rising selective risk indicates unsafe confidence	Requires reliable validation and thresholds [21]

Table 1. Major signals that can support self supervised failure prediction

4. Uncertainty Estimation and the Idea of Knowing When Not to Trust a Prediction

Predictive uncertainty is central to failure prediction because the system needs a numerical or categorical representation of how much it should trust its output. Aleatoric uncertainty reflects variability or noise that is inherent to the observation process, whereas epistemic uncertainty reflects uncertainty about the model or the learned parameters [22]. In a failure prediction setting, epistemic uncertainty is especially important under unseen distribution shifts because a model may encounter regions where its training evidence is sparse. Aleatoric uncertainty remains relevant when the input itself is ambiguous or noisy. Bayesian approaches attempt to represent uncertainty over model parameters or functions. Approximate Bayesian methods, Monte Carlo dropout and deep ensembles have demonstrated practical approaches for estimating predictive uncertainty [23]. Deep ensembles are particularly useful because independently trained models can disagree when the input lies in a region where the learned functions differ. Such disagreement can be transformed into a risk signal. However, ensembles increase computational cost and can still fail to detect some forms of shift when all ensemble members share similar inductive biases. Calibration is a separate but related concept. A confidence value is calibrated when predictions assigned a particular probability are correct at approximately the corresponding frequency. Modern neural networks can be poorly calibrated even when their accuracy is high, and temperature scaling can improve calibration on validation data [24]. This means a calibration procedure learned under the original distribution cannot automatically be assumed valid after deployment. Uncertainty should therefore be treated as one component of a broader failure predictor. A

model with low uncertainty may still fail under a systematic label shift. The strongest design is likely to combine uncertainty with representation novelty, consistency, temporal stability and empirical risk. This is consistent with the observation that no single uncertainty method dominates across all types of distribution shift [25].

5. Out of Distribution Detection and Failure Awareness

Out of distribution detection asks whether an input is sufficiently different from the data used to train a model. Early work showed that maximum softmax probability can serve as a simple baseline for detecting misclassified and unfamiliar examples. Later methods improved this idea using energy scores, feature representations, input preprocessing and auxiliary outlier exposure. These approaches established an important principle: the model can be equipped with a second task that asks whether the current input lies inside a region where the classifier is expected to behave reliably [26]. Energy based detection uses the energy associated with model outputs rather than relying only on the largest class probability. This can produce a more informative separation between familiar and unfamiliar samples in some settings [27]. Other methods use feature distances or nearest neighbor structure. Deep nearest neighbor approaches measure how close a new representation is to known representations, creating an intuitive link between representation space and failure risk [28]. The benefit is interpretability: a system can report not only that risk is high, but also that the current representation is weakly supported by training examples. Consequently, outlier exposure can improve a model's ability to reject known types of novelty while still leaving blind spots. A self supervised approach may be more flexible because it does not require a complete list of future outlier

classes. The most important theoretical limitation is that out of distribution status is not identical to failure. A small distribution change can cause a major performance loss if it changes the label relationship, while a large input change may be harmless if the learned representation remains stable. A strong failure prediction framework should therefore evaluate OOD detection as an intermediate signal rather than the final target. This distinction should be reflected in evaluation metrics that measure actual future error prediction, not only OOD discrimination accuracy [30].

6. Selective Prediction, Abstention and Risk Based Decision Making

A model that can estimate failure risk does not necessarily need to improve every prediction. It can instead decide when not to answer. Selective prediction introduces a reject option so that the system produces predictions only when estimated risk is acceptable [31]. This changes the objective from maximizing overall accuracy to controlling the error rate among accepted prediction. A useful failure predictor should produce a favorable risk coverage curve, meaning that error decreases as the system becomes more selective. This gives a direct operational interpretation of failure prediction. If a model can correctly identify the observations most likely to fail, removing those observations should substantially reduce the remaining error [33,34]. The reject decision can be based on a combined score that includes predictive confidence, uncertainty, representation distance, selfconsistency and recent shift measures. A threshold can then be calibrated on a validation environment. However, threshold transfer is difficult under unseen shifts. A fixed threshold may become too conservative in one environment and too permissive in another. Adaptive thresholds based on recent unlabeled data may help, but adaptation can itself be unstable. This creates a central tradeoff between responsiveness and reliability. Human review can be incorporated as a formal part of the architecture rather than treated as an emergency fallback. A high risk prediction can trigger additional sensing, a second model, a rule based safety check or human inspection. In safety critical settings, this layered approach may be more realistic than expecting a single neural network to recognize every possible failure. The role of self supervised failure

prediction is then to decide when additional protection should be activated [35].

7. Test Time Training and Test Time Adaptation

Test time training uses unlabeled test observations to optimize a self supervised objective before producing the task prediction. This is attractive because the target environment is available at deployment even when labels are not. The approach can improve robustness to changes that affect the input distribution while preserving the original task objective. Related test time adaptation methods use entropy minimization, normalization statistics, consistency objectives or parameter updates to reduce the mismatch between source and target conditions [36]. Entropy based adaptation assumes that confident predictions are generally desirable. Yet the same assumption that motivates adaptation can become dangerous under severe shift because a model can confidently optimize toward an incorrect solution. Methods that constrain adaptation, maintain teacher models or use selective updates attempt to reduce this risk. TeST uses a student teacher self training strategy under distribution shift, illustrating how unlabeled target data can improve robustness without requiring target labels [37]. From a failure prediction perspective, adaptation should be monitored as a dynamic process. Let the model parameters before adaptation be θ_0 and after adaptation be θ_1 . The parameter displacement, change in self supervised loss, change in predictive entropy and change in agreement across augmented views can all be monitored. A sudden large parameter displacement combined with increasing prediction disagreement should be treated as a warning. The system should be able to decide that adaptation is unsafe and preserve a stable reference model. Continual test time adaptation also introduces memory and forgetting issues. A model may encounter a temporary shift and later return to its original environment. If adaptation permanently changes the parameters, performance on the original distribution may deteriorate. Therefore a failure aware architecture may benefit from multiple timescales: a stable base model, a rapidly adapting auxiliary representation and a slow memory of recent environments. This creates a control problem in which adaptation itself becomes a variable that must be monitored [41,42].

8. Theoretical Framework for Predicting Failure Before It Occurs

Let X denote an input, Y the unknown target outcome, M the model and R the event that the model prediction is wrong according to an application specific criterion. The central quantity is the conditional failure probability $P(R \text{ given } X, M, C)$, where C represents contextual information such as recent observations, environment state and model history. The difficulty is that R is not observed before the prediction is made. A failure predictor therefore learns indirect signals Z that are available before or at decision time. The goal is to estimate $P(R \text{ given } Z)$ accurately enough to support a decision. A useful signal vector may be written conceptually as Z equal to confidence, uncertainty, representation novelty, self consistency, temporal instability and distribution shift evidence. A failure score F can then be represented as a learned or calibrated function of Z . The function should ideally satisfy monotonicity properties for selected components, meaning that stronger evidence of uncertainty or instability should not arbitrarily reduce predicted risk. It should also be calibrated so that among cases assigned a risk near a given value, the empirical failure frequency is approximately similar [43,44]. The theoretical distinction between prediction and failure prediction can be expressed through two conditional distributions. The task model estimates $P(Y \text{ given } X)$, while the failure model estimates $P(R \text{ given } Z)$. The first asks what the answer is; the second asks whether the answer should be trusted. This separation explains why improving task accuracy alone may not improve failure awareness. A model can become more accurate on average while becoming more overconfident on rare cases. Conversely, a slightly less accurate model may be safer if it reliably abstains on difficult cases [46,47]. Distribution shift changes the joint distribution of X and Y . If the relationship between the signals Z and failure event R changes after deployment, a failure predictor can also become miscalibrated. This creates a second order problem: the system must detect not only failure risk in the task model but also possible failure of the failure predictor itself. Monitoring the calibration of risk estimates, using multiple signals and retaining a reference distribution are therefore important theoretical requirements. A mature system should report

uncertainty about its own risk estimate when possible [50].

9. Evaluation Strategies for Self Supervised Failure Prediction

Evaluation should measure whether the system identifies future errors before or at the moment they occur. Conventional classification accuracy is insufficient because it ignores the quality of the failure signal. The appropriate metric depends on the cost of false alarms and missed failures. A realistic benchmark should contain several types of shift rather than a single synthetic corruption. Common corruptions, changes in acquisition conditions, new domains, temporal drift, class prevalence changes and semantic novelty should be represented separately. Benchmarking work has shown that performance can degrade under common corruptions even when models perform strongly on conventional test sets [51]. Realistic deployment data are particularly important because synthetic shifts can overestimate the usefulness of a method. Failure prediction should also be evaluated across shift severity. If a mild shift produces a gradual increase in risk, the system may have useful early warning behavior. If risk increases only after accuracy has already collapsed, the system is detecting failure rather than predicting it. A valuable metric is therefore warning lead time, defined as the interval between a meaningful increase in predicted failure risk and the subsequent occurrence of an actual error or unsafe state. In temporal systems, repeated early warnings can be aggregated into event level measures [52]. Ablation studies are necessary to determine whether self supervised signals provide information beyond ordinary confidence. A rigorous experiment should compare confidence alone, uncertainty alone, OOD score alone, self consistency alone and combined models. It should also test whether the failure predictor transfers to shifts that were not present during its development. This directly addresses the central claim of generalization under unseen distribution shifts [55].

10. Practical Applications and Safety Considerations

In medical imaging, a failure aware model could identify scans that differ from the development population, contain unusual artifacts or produce

unstable predictions across transformations. Such a system could route high risk cases for specialist review rather than silently returning a diagnosis. However, demographic or clinical shifts can alter the relationship between image patterns and outcomes, so OOD detection alone is insufficient. The risk model must be evaluated across institutions, devices and patient populations before clinical use. In autonomous systems, failure prediction can be linked to perception uncertainty, sensor disagreement and temporal instability. A vehicle or robot can use these signals to slow down, increase sensing, switch to a safer controller or request intervention. Research on autonomous safety has emphasized the importance of uncertainty because rare failures can have severe consequences [56,57]. The key requirement is that the warning mechanism must remain reliable under conditions that were not represented during development. In industrial monitoring, self supervised objectives can exploit large volumes of unlabeled sensor data. A model may learn normal temporal relationships and then detect reconstruction or prediction errors when machine behavior changes. In addition, the warning mechanism should be independent enough from the main prediction pathway to provide genuine protection. If the same flawed representation controls both prediction and failure scoring, correlated errors can pass through both channels [60].

11. Research Gaps and Future Perspective

A major research gap is the absence of a universal definition of self predicted failure. Existing work often studies related problems separately, including uncertainty estimation, OOD detection, test time adaptation and selective prediction. A unified benchmark should define failure events and measure whether auxiliary signals predict them across multiple tasks. These are valuable, but they do not fully represent the emergence of new concepts, new classes or new causal relationships. Future work should evaluate systems on genuinely prospective data where the shift type is not disclosed during development. This would provide a stronger test of whether self supervised signals can generalize beyond known failure categories [62,63]. Future systems should combine self supervised representation learning with calibrated risk estimation. Causal reasoning is another theoretical opportunity. A distribution shift may affect

a superficial feature, a genuine causal mechanism or both. If a model can distinguish stable causal structure from changing nuisance structure, it may become better at predicting when a shift is likely to cause failure. Combining self supervised objectives with causal representation learning could therefore move failure prediction from statistical novelty toward mechanism aware reliability assessment [64]. There is also a need for formal guarantees. Most current approaches provide empirical evidence that uncertainty or OOD scores correlate with error, but correlation is not the same as a safety guarantee. Future work should investigate conformal prediction, distributionally robust learning, certified uncertainty and formal abstention rules under explicit assumptions. Such guarantees will be especially important in high consequence applications. Finally, failure prediction should be treated as a continuously monitored safety function rather than a one time model property. This layered design provides a practical interpretation of the question posed by the title: an AI model cannot literally know the future, but it can learn measurable evidence that its current prediction is becoming less trustworthy [67,68].

CONCLUSION

Self supervised failure prediction provides a promising framework for making artificial intelligence systems more aware of the conditions under which their own predictions may become unreliable. The central insight is that future failure cannot usually be observed before it occurs, but several measurable signals can provide indirect evidence of rising risk. These signals include predictive confidence, epistemic and aleatoric uncertainty, representation novelty, self consistency, temporal instability and evidence of distribution shift. Self supervised learning is particularly valuable because these signals can be extracted from unlabeled deployment data, reducing dependence on costly failure labels. Test time training and adaptation further show that a model can use incoming observations to improve robustness, although adaptation must itself be monitored because it can create new failure modes. Out of distribution detection is useful but should not be treated as synonymous with failure because novelty and incorrectness are related rather than identical concepts. Similarly, confidence alone is insufficient

because neural networks can remain highly confident on unfamiliar or incorrect inputs. A practical failure aware architecture should therefore combine several complementary signals and convert them into a calibrated risk estimate that supports selective prediction, abstention or human review. Future progress depends on realistic evaluation under genuinely unseen shifts, better definitions of failure, standardized early warning metrics, continual learning with verified feedback, causal reasoning and stronger safety guarantees. The most useful interpretation of model self prediction is therefore not artificial consciousness but measurable reliability estimation. An AI system may not know exactly when it will fail, yet it can be designed to recognize patterns associated with weakening evidence and to respond before an error becomes an unsafe decision.

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. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In International conference on machine learning 2017 Jul 17 (pp. 1321-1330). PMLR.
3. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*. 2017;30.
4. Gal Y, Ghahramani Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In international conference on machine learning 2016 Jun 11 (pp. 1050-1059). PMLR.
5. Kendall A, Gal Y. What uncertainties do we need in bayesian deep learning for computer vision?. *Advances in neural information processing systems*. 2017;30.
6. Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, Dillon J, Lakshminarayanan B, Snoek J. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*. 2019;32.
7. Gawlikowski J, Tassi CR, Ali M, Lee J, Humt M, Feng J, Kruspe A, Triebel R, Jung P, Roscher R, Shahzad M. A survey of uncertainty in deep neural networks: J. Gawlikowski et al. *Artificial intelligence review*. 2023 Oct;56(Suppl 1):1513-89.
8. Malinin A, Gales M. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*. 2018;31.
9. 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.
10. Lee K, Lee K, Lee H, Shin J. A simple unified framework for detecting out of distribution samples and adversarial attacks. *Advances in neural information processing systems*. 2018;31.
11. Liu W, Wang X, Owens J, Li Y. Energy based out of distribution detection. *Advances in neural information processing systems*. 2020;33:21464-75.
12. Sun Y, Ming Y, Zhu X, Li Y. Out of distribution detection with deep nearest neighbors. In International conference on machine learning 2022 Jun 28 (pp. 20827-20840). PMLR.
13. Hendrycks D, Mazeika M, Dietterich T. Deep anomaly detection with outlier exposure. arXiv preprint arXiv:1812.04606. 2018 Dec 11.
14. Sun Y, Wang X, Liu Z, Miller J, Efros A, Hardt M. Test time training with self supervision for generalization under distribution shifts. In International conference on machine learning 2020 Nov 21 (pp. 9229-9248). PMLR.
15. Wang D, Shelhamer E, Liu S, Olshausen B, Darrell T. Tent: Fully test time adaptation by entropy minimization. arXiv preprint arXiv:2006.10726. 2020 Jun 18.
16. Niu S, Wu J, Zhang Y, Chen Y, Zheng S, Zhao P, Tan M. Efficient test time model adaptation without forgetting. In International conference on machine learning 2022 Jun 28 (pp. 16888-16905). PMLR.
17. Zhang M, Levine S, Finn C. Memo: Test time robustness via adaptation and augmentation. *Advances in neural information processing systems*. 2022 Dec 6;35:38629-42.
18. Schneider S, Rusak E, Eck L, Bringmann O, Brendel W, Bethge M. Improving robustness against common corruptions by covariate shift adaptation. *Advances in neural information processing systems*. 2020;33:11539-51.

19. Croce F, Andriushchenko M, Schwag V, Debenedetti E, Flammarion N, Chiang M, Mittal P, Hein M. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*. 2020 Oct 19.
20. Geirhos R, Jacobsen JH, Michaelis C, Zemel R, Brendel W, Bethge M, Wichmann FA. Shortcut learning in deep neural networks. *Nature Machine Intelligence*. 2020 Nov;2(11):665-73.
21. Recht B, Roelofs R, Schmidt L, Shankar V. Do imagenet classifiers generalize to imagenet?. In *International conference on machine learning* 2019 May 24 (pp. 5389-5400). PMLR.
22. Torralba A, Efros AA. Unbiased look at dataset bias. In *CVPR 2011* 2011 Jun 20 (pp. 1521-1528). IEEE.
23. Moreno Torres JG, Raeder T, Alaiz Rodríguez R, Chawla NV, Herrera F. A unifying view on dataset shift in classification. *Pattern recognition*. 2012 Jan 1;45(1):521-30.
24. Ben David S, Blitzer J, Crammer K, Pereira F. Analysis of representations for domain adaptation. *Advances in neural information processing systems*. 2006;19.
25. Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, March M, Lempitsky V. Domain adversarial training of neural networks. *Journal of machine learning research*. 2016;17(59):1-35.
26. Long M, Cao Y, Wang J, Jordan M. Learning transferable features with deep adaptation networks. In *International conference on machine learning* 2015 Jun 1 (pp. 97-105). PMLR.
27. Chadha A, Andreopoulos Y. Improved techniques for adversarial discriminative domain adaptation. *IEEE Transactions on Image Processing*. 2019 Nov 6;29:2622-37.
28. Sun B, Saenko K. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision* 2016 Oct 8 (pp. 443-450). Cham: Springer International Publishing.
29. Wilson G, Cook DJ. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*. 2020 Jul 5;11(5):1-46.
30. Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In *International conference on machine learning* 2017 Jul 17 (pp. 214-223). Pmlr.
31. Goodfellow IJ, Pouget Abadie J, Mirza M, Xu B, Warde Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. *Advances in neural information processing systems*. 2014;27.
32. Kingma DP, Welling M. Auto encoding variational bayes. *arXiv preprint arXiv:1312.6114*. 2013 Dec 20.
33. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* 2016 (pp. 770-778).
34. Dosovitskiy A, Springenberg JT, Riedmiller M, Brox T. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*. 2014;27.
35. Gidaris S, Singh P, Komodakis N. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*. 2018 Mar 21.
36. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning* 2020 Nov 21 (pp. 1597-1607). Pmlr.
37. He K, Fan H, Wu Y, Xie S, Girshick R. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR) 2020 Jun 13* (pp. 9726-9735). IEEE.
38. Grill JB, Strub F, Altché F, Tallec C, Richemond P, Buchatskaya E, Doersch C, Avila Pires B, Guo Z, Gheshlaghi Azar M, Piot B. Bootstrap your own latent a new approach to self supervised learning. *Advances in neural information processing systems*. 2020;33:21271-84.
39. Chen X, Xie S, He K. An empirical study of training self supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV) 2021 Oct 10* (pp. 9620-9629). IEEE.
40. Zbontar J, Jing L, Misra I, LeCun Y, Deny S. Barlow twins: Self supervised learning via redundancy reduction. In *International conference on machine learning* 2021 Jul 1 (pp. 12310-12320). PMLR.

41. Caron M, Touvron H, Misra I, Jégou H, Mairal J, Bojanowski P, Joulin A. Emerging properties in self supervised vision transformers. In 2021 IEEE/CVF international conference on computer vision (ICCV) 2021 Oct 10 (pp. 9630-9640). IEEE.
42. Jing L, Tian Y. Self supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*. 2020 May 4;43(11):4037-58.
43. Jaiswal A, Babu AR, Zadeh MZ, Banerjee D, Makedon F. A survey on contrastive self-supervised learning. *Technologies*. 2020 Dec 28;9(1):2.
44. Angelopoulos AN, Bates S. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*. 2023 Mar 27;16(4):494-591.
45. Romano Y, Patterson E, Candes E. Conformalized quantile regression. *Advances in neural information processing systems*. 2019;32.
46. Cauchois M, Gupta S, Duchi JC. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *Journal of machine learning research*. 2021;22(81):1-42.
47. Geifman Y, El Yaniv R. Selective classification for deep neural networks. *Advances in neural information processing systems*. 2017;30.
48. Chow C. On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory*. 1970 Jan 31;16(1):41-6.
49. Corbière C, Thome N, Bar Hen A, Cord M, Pérez P. Addressing failure prediction by learning model confidence. *Advances in neural information processing systems*. 2019;32.
50. Jiang H, Kim B, Guan M, Gupta M. To trust or not to trust a classifier. *Advances in neural information processing systems*. 2018;31.
51. Ren J, Liu PJ, Fertig E, Snoek J, Poplin R, Depristo M, Dillon J, Lakshminarayanan B. Likelihood ratios for out of distribution detection. *Advances in neural information processing systems*. 2019;32.
52. Choi H, Jang E, Alemi AA. Waic, but why? generative ensembles for robust anomaly detection. *arXiv preprint arXiv:1810.01392*. 2018 Oct 2.
53. Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, Fieguth P, Cao X, Khosravi A, Acharya UR, Makarenkov V. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*. 2021 Dec 1;76:243-97.
54. 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.
55. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* 2016 Aug 13 (pp. 1135-1144).
56. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Advances in neural information processing systems*. 2017;30.
57. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*. 2019 May;1(5):206-15.
58. Doshi Velez F, Kim B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*. 2017 Feb 28.
59. Sinha S, Gehler P, Locatello F, Schiele B. Test: Test time selftraining under distribution shift. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 2023 Jan 2* (pp. 2758-2768). IEEE.
60. Nado Z, Band N, Collier M, Djolonga J, Dusenberry MW, Farquhar S, Feng Q, Filos A, Havasi M, Jenatton R, Jerfel G. Uncertainty baselines: Benchmarks for uncertainty & robustness in deep learning. *arXiv preprint arXiv:2106.04015*. 2021 Jun 7.
61. Mummadi CK, Huttmacher R, Rambach K, Levinkov E, Brox T, Metzen JH. Test-time adaptation to distribution shift by confidence maximization and input transformation. *arXiv preprint arXiv:2106.14999*. 2021 Jun 28.
62. Zhou Z, Guo LZ, Jia LH, Zhang D, Li YF. Ods: Test time adaptation in the presence of open world data shift. In *International Conference on Machine Learning* 2023 Jul 3 (pp. 42574-42588). PMLR.

63. Alfarrar M, Itani H, Pardo A, Alhuwaider S, Ramazanov M, Pérez JC, Cai Z, Müller M, Ghanem B. Evaluation of test time adaptation under computational time constraints. arXiv preprint arXiv:2304.04795. 2023 Apr 10.
64. Liang J, He R, Tan T. A comprehensive survey on test time adaptation under distribution shifts. *International Journal of Computer Vision*. 2025 Jan;133(1):31-64.
65. Ma W, Chen C, Zheng S, Qin J, Zhang H, Dou Q. Test time adaptation with calibration of medical image classification nets for label distribution shift. In *International Conference on Medical Image Computing and Computer Assisted Intervention* 2022 Sep 16 (pp. 313-323). Cham: Springer Nature Switzerland.
66. Choi S, Yang S, Choi S, Yun S. Improving test time adaptation via shift agnostic weight regularization and nearest source prototypes. In *European Conference on Computer Vision* 2022 Oct 23 (pp. 440-458). Cham: Springer Nature Switzerland.
67. DeVries T, Taylor GW. Learning confidence for out of distribution detection in neural networks. arXiv preprint arXiv:1802.04865. 2018 Feb 13.
68. Geifman Y, El Yaniv R. Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning* 2019 May 24 (pp. 2151-2159). PMLR.

HOW TO CITE: Shivika Singh*, A Comprehensive Review On Self Supervised Failure Prediction For AI Models Under Unseen Distribution Shifts, *Int. J. Sci. R. Tech.*, 2026, 3 (10), 481-491. <https://doi.org/10.5281/zenodo.23211369>