Comparing Automated Search Strategies for Discovering Uncertainty-Collapse Failures in Medical Imaging Classifiers

Chase Lambuth
Johns Hopkins University

Abstract

This paper investigates which automated search methods are most effective at finding stress-test perturbations that are likely to cause medical imaging classifiers to fail. This is important because these perturbations represent the types of distribution shifts that can occur in messy real-world data, which allows potential failures to be identified before a model is deployed. Previous research has shown that high accuracy alone is not enough evidence that a model is safe for deployment. This makes pre-deployment stress testing an important part of building reliable machine learning systems, especially in safety critical domains such as medical imaging. This research compares random search, greedy search, Bayesian optimization, evolutionary algorithms, and a novel reinforcement learning approach proposed in this work. A refined uncertainty-collapse reward that considers multiple aspects of model failure is used to guide the search process, which allows each method to identify perturbations that produce the most severe prediction failures. The results show that Bayesian optimization tied for the strongest individual discovery reward, achieved the strongest validation reward, found the largest number of high-quality policies, and achieved the highest average top-k discovery reward. Evolutionary search also performed competitively across these metrics, while PPO identified competitive policies with relatively few evaluations. Random and greedy search demonstrated that simpler search strategies can still identify strong failure scenarios. Ensemble approaches were also compared in the paper by combining multiple search methods, which showed that different combinations provided trade-offs between policy diversity and coverage while retaining the strongest validation reward. The paper also investigates whether certain search methods discover perturbation policies that other methods fail to identify. The results show differences in the types of perturbations discovered across search strategies. Gaussian noise appears most consistently among high-reward policies by a wide margin. Overall, the results suggest that Bayesian optimization provides the strongest standalone strategy for discovering uncertainty failures, while combining complementary search methods may provide the most comprehensive approach to automated stress testing.

Keywords: Medical Imaging, Uncertainty Estimation, Stress Testing, Distribution Shift, Reinforcement Learning, Calibration

1. INTRODUCTION

In deep learning systems, model safety has become increasingly important as models are deployed into real-world applications where failures can have major consequences. Deep learning systems are now deployed across a multitude of domains, and many of these domains have strict regulatory requirements regarding safety. Medical imaging is one of the most safety-critical domains because an incorrect classification can lead to an incorrect diagnosis or a missed diagnosis due to a false negative [1], [2]. Although a domain expert such as a radiologist is typically involved in the diagnostic process, errors can still occur because the model may identify subtle patterns that indicate a particular diagnosis which the radiologist could overlook. This ability to detect complex patterns is one of the key advantages of machine learning models in medical imaging.

One of the most important ways to improve the safety of these systems is to identify mistakes during pre-deployment before they occur in real-world settings, allowing the model to be refined and more safely released. This can make models more trustworthy before they are deployed, reducing the likelihood of errors in post-deployment. One of the main methods for identifying mistakes before they happen is pre-deployment stress testing [3], [4]. This involves manipulating the data to determine whether it will ”break” the model and harm its performance. Considering how messy real-world data is, this is essential because data can become corrupted or difficult to interpret, and it is important that the model is still able to make accurate predictions under these conditions [1], [5].

There are many different ways stress testing can be performed, and the specific approach often depends on the domain. For large language models, this typically

involves techniques such as jailbreaking. In medical imaging systems, which this paper focuses on, stress testing is typically performed by manipulating images. These corruptions commonly involve introducing noise, blur, or compression artifacts [1].

There are also many different ways medical imaging stress testing can be conducted. In terms of search methods, a search algorithm can be used to identify the perturbations that degrade model performance the most [3], [4]. For example, a search algorithm may discover that a particular magnitude of blur consistently causes the model to fail. Search methods that can potentially accomplish this include random search, greedy search, local search, reinforcement learning, evolutionary algorithms, and Bayesian optimization. This raises the question of which search methods are most effective for stress testing, while also asking which types of perturbation, such as blur or compression artifacts, have the greatest impact on model performance. Answering these questions can help determine which corruptions should be prioritized during stress testing and which search methods are most effective at discovering model failures.

In stress testing, it is also important that the perturbations discovered are diverse rather than repeatedly identifying the same failure scenarios [6]. Oftentimes, the most unique perturbations cause the greatest problems because they represent edge cases, which can be difficult to identify during pre-deployment testing. In systems with many users, even rare edge cases are likely to occur eventually, making them important to discover before deployment. This is especially important for ensemble-based stress testing research, where combining the strongest search method with one that discovers diverse failure modes may provide more comprehensive coverage of model vulnerabilities.

This work makes two primary contributions. First, it provides a systematic comparison of automated search strategies for discovering uncertainty-collapse failures in medical imaging classifiers. Second, it investigates whether combining complementary search strategies can improve pre-deployment stress testing by providing broader coverage of potential model failures.

2. RELATED WORK

As machine learning systems continue to be deployed in safety-critical applications, making sure that they remain reliable under unexpected conditions has become an increasingly important research problem. One approach that has emerged to address this challenge is stress testing, which evaluates model behavior under challenging scenarios that may not be represented in the training or testing data. Although research in this area is still relatively limited, several studies have demonstrated that stress testing can identify failure scenarios that traditional evaluation methods may overlook.

Prior research has also shown that automated search techniques, including reinforcement learning, can discover failure scenarios that are difficult to identify using manually designed tests [3]. In this paper, Lee et al. explored how a reinforcement learning agent could be used as an adaptive stress testing technique for automated systems such as robotics. Their results demonstrate that a learning-based stress testing technique can be used to identify problems in a model before deployment.

There is also research regarding how stress testing can be used in medical imaging specifically. Islam et al. [1] argue that achieving high metric scores on the test set is not enough evidence to conclude that a medical imaging model is safe for deployment. The authors point out that test sets are often drawn from the same distribution as the training data, meaning that messy real-world data may produce much lower accuracy because it comes from a different distribution. It is also worth mentioning that the training data may come from a different part of the world than where the model is deployed. For example, in areas with high air pollution, lung scans may follow a different distribution than those from areas with low air pollution. Likewise, certain demographic groups may be underrepresented in the training and testing data, which could become problematic in medical imaging applications. Instead of using an adaptive stress testing approach, the authors evaluated robustness using a fixed set of perturbations, such as blur with severity levels ranging from 1 to 6.

Additionally, there has been some research on uncertainty estimation in medical imaging. Lambert et al. [2] argue that uncertainty estimation is an important component of trustworthy clinical AI because it allows models to indicate when they are uncertain instead of making highly confident incorrect predictions. The authors review techniques such as Monte Carlo Dropout, deep ensembles, and calibration methods, while also discussing the challenges of evaluating uncertainty in real clinical environments. This demonstrates that uncertainty estimation plays an important role in improving the safety and reliability of medical imaging systems before deployment.

In stress testing, it is also important that the perturbations discovered are diverse rather than

repeatedly identifying the same failure scenarios. Previous work has shown that generating diverse test cases improves the ability of automated stress testing methods to uncover corner-case behaviors and latent model defects that repetitive transformations may fail to expose [4]. This is especially important for ensemble-based stress testing research, where combining the strongest search method with one that discovers diverse failure modes may provide more comprehensive coverage of model vulnerabilities.

Although related work has shown that automated search methods can be used to identify failure scenarios and that stress testing can improve the safety of medical imaging systems before deployment, relatively little research has systematically compared different search strategies for discovering uncertainty-collapse failures. There is also limited research investigating whether different search methods identify specific perturbation types, or broader stress-test coverage when used together. This project compares reinforcement learning, random search, greedy search, Bayesian optimization, and an evolutionary algorithm as automated stress testing methods to determine which approaches most effectively discover uncertainty-collapse failures and whether different search strategies provide complementary coverage of potential model vulnerabilities.

3. RESEARCH PROBLEM

This study addresses a central problem in deep learning systems, which is how to properly stress test a model to determine whether it is safe to deploy. This is particularly relevant for medical imaging systems considering how safety-critical these domains are. The problem itself centers around the idea that medical imaging models can be highly confident when they are incorrect, which may make them less likely to trigger human review or uncertainty-based safeguards. Under distribution shift, this problem can become even more common, as model uncertainty and calibration can become less reliable when the input distribution differs from the training distribution [5].

There is also the problem that there is currently limited research on which stress testing methods are the most effective. If someone wants to stress test a deep learning system, they may resort to using the most common method simply because it is well established. Previous research has shown that different search strategies can produce different results when searching for failure-revealing test cases, particularly in terms of the coverage and diversity of the failures they discover [6]. There is currently little guidance regarding which automated search strategy should be used for stress testing, and different search methods may produce different trade-offs between search efficiency, failure severity, and policy diversity. As a result, those developing these systems may need to spend additional time evaluating multiple search methods themselves, creating an unnecessary extra step that could potentially be avoided if there were prior evidence to guide which methods perform best for a particular application.

Additionally, there is little research regarding ensemble methods in automated stress testing. Existing research typically only evaluates individual search methods, even though ensembles could potentially improve both the severity and diversity of the discovered failure scenarios. Because ensemble-based stress testing has received little attention, it is not yet well understood how effective these approaches are compared to individual search methods.

This project studies whether automated search strategies can discover perturbation policies that expose these uncertainty-collapse failures. Several search strategies are compared to evaluate their effectiveness and efficiency. Specifically, search strategies are evaluated based on the severity of the failures they discover, how well those failures transfer to validation data, the number of evaluations required to identify strong policies, and the diversity of the perturbation policies they produce. This allows the study to examine not only which method discovers the strongest individual failure, but also which methods consistently discover larger collections of high-quality failure scenarios.

The project also examines whether different search strategies tend to discover different types of perturbations and whether combining multiple strategies provides broader stress-test coverage than relying on a single method. This is important because the strongest individual search strategy may not necessarily identify every type of failure that could affect a deployed classifier. By comparing individual and ensemble-based approaches, this study aims to provide clearer guidance regarding which automated search strategies may be most useful under different stress-testing objectives.

4. METHODS

This study evaluates whether automated pre-deployment stress testing can identify interpretable distribution shifts that cause uncertainty-collapse failures in medical imaging classifiers. The experimental workflow consists of

multiple stages, including preparing the CheXpert dataset, training a fixed baseline classifier, evaluating uncertainty calibration, searching for high-risk perturbation policies using multiple automated search strategies, and comparing the discovered failure scenarios in terms of severity, generalization, diversity, and search efficiency.

4.1. Dataset and Prediction Task

The experiments use the CheXpert chest X-ray dataset for binary classification of Cardiomegaly. This dataset was chosen because it has been widely validated and is commonly used for medical imaging tasks. Because it is widely used, it makes it easier to compare the results against existing benchmarks. CheXpert is also one of the largest publicly available medical imaging datasets, which means it is large enough for sizable training, validation, and testing sets. Small sample sizes can sometimes produce misleading results, especially when performance differences are not major.

In terms of preprocessing, uncertain Cardiomegaly labels were mapped to zero to create a binary target. A binary target was used to make the results easier to interpret and to reflect the common clinical framing of whether a finding is present or absent. Images were resized to 224×224 pixels and paired with their metadata-derived labels. To reduce data leakage, the dataset was split at the patient level rather than the image level, ensuring that studies from the same patient did not appear across multiple splits.

The prepared dataset contained 223,414 studies from 64,540 patients, with an overall positive Cardiomegaly rate of approximately 12.1 percent. The training split contained 157,122 studies, the validation split contained 32,800 studies, and the test split contained 33,492 studies, with similar class balance across all three splits. The baseline classifier was trained using the training split and selected using the validation split. Subsequent uncertainty estimation and standard corruption evaluation used the validation split.

For the automated perturbation-search experiments, fixed stratified subsets of the validation split were used to make repeated policy evaluation computationally feasible. The discovery subset contained 1,024 studies and was used to search for high-risk perturbation policies. A separate, non-overlapping search-validation subset contained 2,048 studies and was used to evaluate whether the discovered policies generalized beyond the discovery subset.

4.2. Baseline Classifier

A ResNet-18 classifier was trained as the fixed target model for all later stress-testing experiments. The model used a binary output head for Cardiomegaly prediction and was trained using binary cross-entropy with logits. Training used a batch size of 32, a learning rate of 10^{-4}, a dropout rate of 0.20, and up to 10 training epochs. The final checkpoint was selected based on validation AUROC because the dataset is imbalanced, which can make accuracy alone a misleading evaluation metric.

4.3. Uncertainty Estimation and Calibration

The classifier's uncertainty behavior was evaluated before distribution shifts were introduced. Three uncertainty estimation approaches were compared, which included raw sigmoid confidence, temperature scaling, and Monte Carlo Dropout. Temperature scaling fit a scalar temperature on the validation data and applied it to the model logits before probability estimation. Monte Carlo Dropout used repeated stochastic inference passes to estimate predictive uncertainty. This evaluation is important because the main goal of this paper is to determine which search methods can find perturbations that cause the model to make confidently incorrect predictions, and those predictions are measured using uncertainty estimates. Since there are multiple ways to estimate uncertainty, this experiment determined which approach provided the best calibrated confidence estimates for the baseline model on this dataset. Evaluation metrics included accuracy, AUROC, Expected Calibration Error, Brier Score, high-confidence error rate, mean confidence, and mean confidence on incorrect predictions. A high-confidence error was defined as an incorrect prediction with a confidence of at least 0.90.

4.4. Perturbation Search Space

Distribution shifts were represented as interpretable perturbation policies applied to chest X-ray images before inference. Each policy contained up to three perturbation actions. The search space included Gaussian noise, Gaussian blur, brightness adjustment, contrast adjustment, JPEG compression, and affine translation/rotation. Each perturbation had five discrete severity levels.

This design allowed search methods to test both single corruptions and compound perturbation policies. Compound policies were important because severe uncertainty collapse often appeared when multiple realistic image degradations were combined,

especially Gaussian noise, JPEG compression, and affine transformation.

4.5. Uncertainty-Collapse Reward

Each perturbation policy was evaluated relative to clean-image performance. The reward was designed to prioritize failures where the classifier became confidently incorrect. The selected reward was:

$$R = 2\Delta\mathrm{HCER} + \Delta\mathrm{ECE} + 0.5\Delta\mathrm{WrongConfidence}.$$

ΔHCER is the increase in high-confidence error rate, ΔECE is the increase in Expected Calibration Error, and ΔWrongConfidence is the increase in mean confidence on incorrect predictions. This reward function directly targets uncertainty collapse, which is the primary focus of this project. High-confidence error rate received the largest weight because confidently incorrect predictions represent a major safety risk in medical imaging. Expected Calibration Error received a moderate weight because it reflects the reliability of the model's confidence estimates, while wrong confidence received a smaller weight because it serves as a supporting indicator of uncertainty collapse. Metrics such as accuracy drop, AUROC drop, and Brier Score were still measured as supporting evaluation metrics; however, they were not included in the final reward. As a result, the reward prioritizes uncertainty collapse over overall performance degradation.

4.6. Automated Search Strategies

Multiple automated search strategies were compared over the same perturbation space. Random search sampled perturbation policies without using feedback. Greedy search constructed policies step-by-step by selecting the locally strongest perturbation at each step. PPO reinforcement learning treated policy construction as a sequential decision problem, where the agent selected perturbation type, severity, and stopping behavior. Evolutionary search used an evolutionary algorithm to iteratively evolve perturbation policies toward higher reward values. Bayesian optimization used a probabilistic surrogate model to systematically select and refine perturbation policies that were most likely to break the model while minimizing the number of evaluations.

4.7. Validation and Comparison

Candidate perturbation policies were first evaluated on a fixed discovery subset. Top policies were then evaluated on a larger validation subset to test whether high-reward stress-test policies continued to produce uncertainty-collapse behavior beyond the discovery set. Search methods were compared using best reward, top-k reward, validation reward, search efficiency, perturbation-family frequency, policy diversity, overlap among discovered policies, and the number of high-quality policies discovered. High-quality policies were defined as policies that achieved a discovery reward of at least 0.388985, corresponding to the 90th-percentile reward threshold across unique discovered policies.

4.8. Ensemble and Diversity Analysis

The final stage of the analysis examined whether combining search methods produced broader stress-test coverage than relying on a single method alone. This ensemble analysis was conducted post hoc using the policies already discovered by the individual search strategies. For each method, the highest-reward unique policies were selected, and candidate three-method ensembles were formed from those policy sets. Three ensemble strategies were compared: a Best-3 Ensemble, which selected the individually strongest methods; a Most-Diverse-3 Ensemble, which selected the combination whose policies were most different from one another; and a Hybrid-Score Ensemble, which balanced reward, held-out performance, high-risk policy count, perturbation coverage, severity-pattern coverage, predictiveness, and diversity. Diversity was measured using Jaccard distance over the perturbation-and-severity tokens in each policy, so ensembles received higher diversity scores when their selected policies used different transformation types or severity combinations. This analysis tested whether ensemble-based stress testing could identify complementary uncertainty-collapse failures that might be missed by a single search strategy.

5. RESULTS

The automated search strategies were evaluated using discovery reward, validation (held-out) reward, search efficiency, policy uniqueness, ensemble diversity, and classifier degradation. Tables 1–4 summarize the primary results.

5.1. Search Method Performance

Table 1 compares the search strategies based on the severity, held-out transfer, efficiency, and uniqueness of the uncertainty-collapse policies they discovered.

Table 1. Comparison of the automated search strategies based on discovery reward, validation (held-out) reward, method diversity, high-quality policy yield, policy uniqueness, and computational efficiency. Arrows indicate the desired direction of each metric.

Metric	Random	Bayes.	Evol.	Greedy	PPO
Best Disc. Reward ↑	0.393	0.397	0.397	0.386	0.390
Top-k Disc. Reward ↑	0.380	0.396	0.393	0.347	0.266
Top-k Val. Reward ↑	0.385	0.398	0.395	0.361	0.286
Best Validation Reward ↑	0.395	0.400	0.399	0.389	0.392
Method Diversity ↑	0.870	0.719	0.568	0.500	0.763
High-Quality Policies ↑	3	44	11	0	2
Unique Policies (%) ↑	100.0	84.0	100.0	100.0	100.0
Evaluations ↓	200	200	200	3	20
Eval. to Best ↓	37	190	72	3	1

Bayesian optimization achieved the highest top-k discovery reward and top-k validation reward, while also producing the largest number of high-quality policies. Bayesian optimization and evolutionary search achieved the highest best discovery reward, with Bayesian optimization producing the highest best validation reward. Random search achieved the highest method diversity, while PPO found its strongest policy using the fewest evaluations.

5.2. Classifier Performance

Table 2 reports classifier performance on the clean validation reference, the strongest standard corruption benchmark, and the strongest perturbation policy identified by each search method.

Table 2. Classifier performance on clean images, the strongest standard corruption, and the strongest perturbation policy discovered by each search strategy using reward, calibration, discrimination, and confidence-degradation metrics. Arrows indicate the direction corresponding to greater classifier degradation.

Metric	Clean	Std. Corr.	Random	Bayes.	Evol.	Greedy	PPO
Accuracy ↓	0.883	0.881	0.879	0.879	0.879	0.879	0.879
AUROC ↓	0.834	0.813	0.636	0.591	0.644	0.689	0.603
ECE ↑	0.041	0.061	0.113	0.116	0.116	0.114	0.111
Brier ↑	0.087	0.097	0.119	0.119	0.119	0.118	0.118
HCER ↑	0.024	0.064	0.121	0.121	0.121	0.118	0.121
Wrong Conf. ↑	0.731	–	0.991	0.995	0.994	0.987	0.989

Although accuracy remained relatively stable across the strongest perturbation policies, the other performance and uncertainty metrics degraded severely. AUROC decreased from 0.834 on clean images to between 0.591 and 0.689 across the discovered policies, while HCER increased from 0.024 to as high as 0.121. Confidence on incorrect predictions also increased substantially, reaching 0.995 for the policy discovered by Bayesian optimization.

5.3. Ensemble Performance

Table 3 compares the three search-method ensembles to determine whether combining complementary strategies improves failure severity, policy coverage, and policy diversity.

Table 3. Comparison of the Best-3, Most-Diverse-3, and Hybrid-Score search-method ensembles based on reward, held-out transfer, diversity, and policy coverage.

Metric	Best-3	Div.-3	Hybrid
Best Reward ↑	0.397	0.397	0.397
Top-10 Reward ↑	0.396	0.396	0.396
Validation Reward ↑	0.400	0.400	0.400
Diversity Score ↑	0.791	0.807	0.791
Unique Policies ↑	28	21	28
High-Risk Policies ↑	22	13	22
Transform Coverage ↑	6	6	6
Severity Patterns ↑	22	16	22

All three ensembles achieved the same best reward, top-10 reward, validation reward, and transformation coverage. The Most-Diverse-3 ensemble achieved the highest diversity score of 0.807, while the Best-3 and Hybrid-Score ensembles contained more unique policies, high-risk policies, and severity patterns. These results show a trade-off between maximizing policy diversity and maintaining broader coverage of high-risk policies.

5.4. Perturbation-Family Analysis

Table 4 shows which perturbation families were most frequently selected among the highest-reward policies discovered by each automated search strategy.

Table 4. Frequency with which each perturbation family appeared among the highest-reward validated policies discovered by each search strategy. Counts are reported over the top 10 validated policies where available. Note that greedy search produced three validated policies.

Perturbation	Random	Bayes.	Evol.	Greedy	PPO
Gaussian Noise	10/10	10/10	10/10	3/3	10/10
Gaussian Blur	1/10	5/10	0/10	0/3	1/10
Brightness	4/10	2/10	1/10	0/3	2/10
Contrast	5/10	2/10	9/10	0/3	0/10
JPEG Compression	2/10	1/10	8/10	2/3	3/10
Affine Trans./Rot.	5/10	0/10	0/10	1/3	4/10

Gaussian noise was the most consistently selected perturbation, appearing in every highest-reward policy across all five search strategies. The remaining perturbation families differed more greatly between methods. Contrast and JPEG compression appeared frequently in the evolutionary policies, Gaussian blur appeared most frequently in the Bayesian optimization policies, and affine transformations and rotations

appeared more frequently in the random search and PPO policies.

6. DISCUSSION

6.1. Main Search Strategy Performance

The primary objective of this study was to determine which automated search strategies were most effective at identifying perturbation policies that expose classifier uncertainty failures. The search methods were compared based on their ability to discover high-severity policies, generalize to validation data, search efficiently, and expose diverse solutions. Although all search strategies identified policies that degraded calibration and confidence, clear differences emerged in their overall effectiveness, especially in search efficiency, solution quality, and policy diversity.

Bayesian optimization produced one of the strongest individual perturbation policies under the experimental configuration used in this study. It tied for the highest discovery reward and achieved the highest validation reward. It also achieved the highest average top-k discovery reward and the highest average top-k validation reward, which suggests that it was able to consistently find strong perturbation policies instead of only identifying a single strong solution. However, Bayesian optimization used a relatively high number of evaluations and did not find its strongest policy until later in the search. These results suggest that Bayesian optimization is a strong strategy for finding high-severity perturbations when a larger evaluation budget is available.

The evolutionary algorithm also produced a strong collection of high-quality policies. It tied with Bayesian optimization in terms of best discovery reward and only slightly trailed behind it in terms of best validation reward, while also achieving similar average top-k discovery and validation rewards. This suggests that the evolutionary algorithm was also consistent at finding strong perturbation policies across many solutions, although Bayesian optimization performed slightly better on the reward-based metrics.

The proposed PPO method performed well overall, although it did not perform as strongly as some of the other search strategies. While it used a relatively low number of evaluations, it achieved a lower top-k discovery reward of 0.266. However, its best discovery reward and best validation reward were comparable to those of the other top-performing methods. This suggests that PPO was capable of finding a strong perturbation policy efficiently but was less consistent at finding a larger collection of high-quality policies. Although PPO was not the strongest method in this study, the results demonstrate that a reinforcement learning agent can be used as a search method for discovering perturbations that expose uncertainty failures in medical imaging classifiers, an application for which there was limited prior evidence.

The remaining search strategies demonstrated different trade-offs. Random search performed surprisingly well given its simplicity, as it produced several competitive policies through broad exploration and had the highest method diversity among the individual search methods. This shows that the search strategy does not necessarily need to be complex to perform well. Lastly, greedy search located strong solutions almost immediately with only a few evaluations but did not perform as strongly as the other methods overall. This reflects its tendency to exploit local improvements instead of exploring alternative perturbation combinations. This method may be helpful for applications where the number of evaluations needs to be limited, at the cost of not achieving the strongest overall results. This could be particularly useful when each evaluation is computationally expensive or the available evaluation budget is limited.

6.2. Ensemble Performance

The ensemble analysis suggests that combining complementary search methods can provide broader coverage of the perturbation space without sacrificing the quality of the strongest discovered policies. Both the Best-3 and Hybrid-Score ensembles retained the highest validation reward while including a larger collection of high-risk perturbations. The Most-Diverse-3 ensemble achieved the highest diversity while retaining the same best validation reward, although it covered fewer high-risk policies than the Best-3 and Hybrid-Score ensembles. Combining multiple search strategies may therefore provide a more comprehensive stress-testing framework than relying on a single search method.

6.3. Most Common Perturbations

Understanding which perturbations appear most frequently in policies that cause uncertainty failures could be particularly important for medical imaging classifiers because different image-quality problems can affect the diagnostic quality of an image. Gaussian noise was the most common perturbation among the highest-reward policies. It appeared in all of the top policies discovered by random search, Bayesian optimization, evolutionary search, PPO, and

greedy search. Contrast was also relatively common, particularly for evolutionary search, where it appeared in 9 of the top 10 policies. JPEG compression appeared in 8 of the top 10 evolutionary search policies, while Gaussian blur appeared in 5 of the top 10 Bayesian optimization policies. These results suggest that Gaussian noise was the perturbation most consistently associated with high-reward policies in this experiment, while the frequency of the other perturbations differed between search strategies.

The search strategies also differed in the types of perturbations included in their highest-reward policies. Random search and PPO more frequently discovered policies containing affine transformations and rotations, while Bayesian optimization more frequently discovered policies containing Gaussian blur and evolutionary search more frequently discovered policies containing contrast and JPEG compression. These results suggest that the best search strategy may also depend on the application. For example, in robotics, a search method that is more effective at discovering failures caused by blur may be more useful because motion blur can occur when cameras are affected by movement or vibration. In this case, a search strategy that frequently discovers blur-related failures could potentially provide more relevant stress tests for the types of distribution shifts that are likely to occur in that application.

6.4. Overall Recommendation

Based on the results of this study, Bayesian optimization produced the strongest overall results under the experimental configuration and evaluation budget used in this study. It tied for the strongest individual discovery reward while also achieving the highest validation reward and the highest average top-k discovery and validation rewards. Evolutionary search performed similarly across these metrics, only slightly trailing Bayesian optimization. These results suggest that Bayesian optimization may be particularly useful when the goal is not only to identify a single weakness, but to consistently discover strong perturbations that expose uncertainty failures.

However, the results also suggest that there may not be a single best search strategy for every application. Evolutionary search remained highly competitive with Bayesian optimization, while greedy search may be useful when the number of evaluations is heavily limited. When computational resources allow, combining complementary search strategies may provide a more comprehensive approach because the ensemble results showed that increased coverage and diversity could be achieved without reducing the quality of the strongest discovered policies. Overall, Bayesian optimization produced the strongest reward-based performance under the evaluated configuration, while an ensemble of complementary search methods provided broader stress-test coverage.

6.5. Limitations

There are several limitations to this study that should be considered when interpreting the results. First, the experiments were conducted using a single medical imaging dataset and a single classifier architecture. Although the results demonstrate differences between the search strategies for this application, the same results may not generalize to other medical imaging datasets, prediction tasks, or model architectures. Future research should evaluate these search methods across multiple datasets and classifiers to determine whether the same performance patterns remain consistent.

Another limitation is that the perturbation search space was restricted to six predefined transformation families with five severity levels. While these perturbations represent different types of image degradation, they cannot represent every type of distribution shift that may occur in real-world medical imaging data. Real-world shifts may result from differences in imaging equipment, acquisition protocols, patient populations, or other factors that are more difficult to reproduce using synthetic perturbations.

Additionally, the search methods were not extensively hyperparameter-tuned, meaning that the selected configurations may not represent the best possible performance of each method. More extensive hyperparameter tuning could potentially improve the performance of some search strategies or change their relative performance. As a result, the results should be interpreted as a comparison of the search strategies under the experimental configurations and budgets used in this study rather than a definitive ranking under every possible configuration or computational budget.

This study also evaluates whether perturbations cause uncertainty and confidence failures in the classifier, but it does not evaluate the clinical consequences of these failures. The discovered perturbation policies should therefore be interpreted as stress-test conditions that expose model vulnerabilities rather than direct evidence of clinical risk.

6.6. Future Research

A potential extension of this work would be to use an image-quality assessment system to identify

the type or severity of degradation present in an image. Since certain perturbations were more likely to appear in policies that caused uncertainty failures, images containing these types of degradation could potentially be flagged for additional review before relying on the classifier output. Automated image-quality detection has already been proposed for identifying problematic medical images, although the use of perturbation-specific risk levels for deciding when to flag classifier predictions would need to be validated separately.

Future research could also evaluate these search methods across additional medical imaging datasets and classifier architectures. Since this study only focused on a single dataset and classifier, it is unclear whether the same search strategies would perform best in other applications. It would also be useful to compare the search methods using the same number of evaluations and with additional hyperparameter tuning.

7. CONCLUSION

This study investigated how automated search strategies can be used to stress test medical imaging classifiers and identify uncertainty-collapse failures before deployment. Stress testing is especially important in safety-critical domains, such as medical imaging, where strong performance on a traditional test set does not necessarily indicate that a model will remain reliable when tested on messy real-world data. The results show that automated search can identify perturbation policies that cause the model to become less reliable even when accuracy remains relatively stable.

This work demonstrates how automated search can be used to explore combinations of perturbations and identify perturbations that find weaknesses in medical imaging classifiers. Bayesian optimization produced the strongest overall reward-based performance, achieving the highest validation reward, the highest average top-k discovery and validation rewards, and the largest number of high-quality policies. Evolutionary search performed similarly, tying Bayesian optimization in best discovery reward and only slightly trailing it in best validation reward. Random search achieved the highest method diversity, while PPO and greedy search showed that efficient methods can still find strong policies but may be less consistent overall.

The discovered perturbation policies caused clear degradation in model reliability. Although accuracy remained relatively stable, AUROC dropped severely under the strongest discovered policies, while expected calibration error, high-confidence error rate, and confidence on incorrect predictions increased. Gaussian noise appeared most consistently among the highest-reward policies, while other perturbations such as blur, contrast, JPEG compression, and affine transformations appeared differently across search methods. This suggests that different search strategies can expose different types of failure cases, which is important because relying on a single search method may leave certain vulnerabilities undiscovered.

This study also demonstrates the potential value of combining complementary search strategies. The Best-3 and Hybrid-Score ensembles retained the strongest validation reward while covering more high-risk policies and severity patterns than the Most-Diverse-3 ensemble. The Most-Diverse-3 ensemble had the highest diversity score, but it covered fewer high-risk policies. This shows that ensemble-based stress testing can provide broader coverage than relying on a single search method, especially when the goal is to balance severe failure discovery with perturbation diversity.

This work supports automated stress testing as a potentially useful addition to the pre-deployment evaluation of medical imaging classifiers. Bayesian optimization was the strongest standalone method under the evaluated configuration, while hybrid ensemble strategies provided broader stress-test coverage. Future research could extend this framework to additional medical imaging datasets, classifier architectures, and more realistic sources of distribution shift.

References

[1] M. T. Islam, M. H. Ahsan, T. Karmaker, M. M. Rahman, et al., "Robustness stress testing in medical image classification," *arXiv preprint arXiv:2308.06889*, 2023.

[2] B. Lambert et al., "Trustworthy clinical ai solutions: A unified review of uncertainty quantification in medical imaging," *Artificial Intelligence in Medicine*, vol. 154, p. 102 923, 2024.

[3] R. Lee, M. J. Kochenderfer, O. J. Mengshoel, G. P. Brat, and M. P. Owen, "Adaptive stress testing of autonomous systems," *Journal of Artificial Intelligence Research*, vol. 69, pp. 1165–1212, 2020.

[4] H. Ben Braiek and F. Khomh, "Deepevolution: A search-based testing approach for deep neural networks," *arXiv preprint arXiv:1909.02563*, 2019.

[5] Y. Ovadia et al., "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[6] A. Sorokin, R. Safin, and S. Nejati, "Can search-based testing with pareto optimization effectively cover failure-revealing test inputs?" *Empirical Software Engineering*, 2025.