Journal of Innovative Technologies
https://journalinnotech.com/
Vol 2(2) 2026

Comparative Empirical Evaluation of Monte Carlo Dropout Uncertainty Measures for Failure Capture and Referral Trade-offs in Chest X-ray Disease Detection: A Reproducible Simulation Study

Yali Zhang[1], Chenhui Hao[2,*]

[1] Computer Science, Rice University, Houston, TX, USA
[2] Civil Engineering, University of California, CA, USA

* Corresponding Email: ce.haochenhui@gmail.com

Abstract

Reliable deployment of automated chest radiograph interpretation depends not only on classification accuracy but on whether a network can flag the cases it is likely to read incorrectly. This study presents a comparative evaluation of four uncertainty measures derived from Monte Carlo dropout, namely predictive confidence, predictive variance, predictive entropy, and mutual information, applied to two clinically motivated tasks: capturing misclassified predictions and supporting uncertainty-based referral. Because the real image collections were not available for this work, the comparison is conducted on a reproducible synthetic benchmark that emulates the statistical behaviour of a DenseNet-121 classifier under Monte Carlo dropout on five findings from a ChestX-ray14-style multi-label detection task, together with a smaller synthetic set that emulates a CheXpert-style radiologist-consensus partition; every reported number is computed from this simulated data rather than from real radiographs or real model inference. Misclassification detection is scored through the area under the receiver operating characteristic curve, while referral is analysed by withholding the most uncertain cases and measuring the sensitivity and specificity retained on the remainder. Predictive variance attains the highest misclassification-detection area under the receiver operating characteristic curve at 0.800, ahead of predictive entropy at 0.784, mutual information at 0.782, and predictive confidence at 0.743; on the precision–recall view the order shifts and predictive entropy leads at 0.612. Referring the most uncertain 20% of cases raises retained sensitivity from 0.770 to 0.783 and specificity from 0.704 to 0.734. The synthetic external set reproduces the same AUROC ordering. Within the limits of a simulation study, the evidence indicates that distribution-based measures offer a consistent advantage over single-pass confidence while isolating the evaluation protocol on data with a fully known ground truth; validation on real radiographs remains necessary before any clinical claim.

***Keywords*:** uncertainty estimation; Monte Carlo dropout; chest radiography; selective prediction

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/)

1. Introduction

Automated interpretation of chest radiographs has advanced quickly, yet a persistent obstacle to clinical adoption is that a network rarely communicates when its own prediction should be doubted. A confident-looking probability can accompany an entirely wrong reading, and clinicians have limited means to separate dependable outputs from unreliable ones. The question this study addresses is whether uncertainty measures obtained from a standard classifier can identify the cases a network is likely to misread, and whether withholding those cases for human review improves the reliability of the predictions that remain. The investigation is deliberately comparative and empirical: it places several established measures under one protocol on the same predictions and reports how they behave, rather than introducing any new method.

1.1. Clinical and Regulatory Motivation

Modern deep networks are frequently miscalibrated, producing probabilities that overstate their true correctness, which makes the raw output an unsafe basis for clinical trust [1]. The cost of an unflagged error in radiology is high, and earlier work in retinal imaging showed that referring uncertain cases to a specialist can raise diagnostic performance on the retained set without altering the underlying classifier [2]. Regulatory attention in the United States reinforces this concern. Guidance issued by the Food and Drug Administration on artificial-intelligence-enabled device software emphasises transparency about performance limits, the representativeness of the evaluated population, and disclosure of failures that an assessment cannot resolve. An interpretation aid that signals its own doubt aligns closely with these expectations, because it allows difficult cases to receive additional scrutiny rather than an unexamined automated verdict. The clinical appeal of uncertainty is practical as much as conceptual: a measure that reliably rises on the cases a model gets wrong can be turned into a referral rule that protects patients at the boundary of the model's competence, and the same signal can document for a reviewer how often the tool chose to defer.

1.2. Research Gap and Contributions of This Empirical Study

A. Limited Comparative Evidence on Uncertainty Measures in Chest Radiography

Large public chest radiograph collections and strong baseline networks have made disease detection a well-studied task [3], and a densely connected classifier of this family reached radiologist-level performance on pneumonia detection, an early demonstration of the architecture's competitiveness on these labels that has not been shown to extend uniformly to every finding [4]. Most of this literature reports discrimination accuracy while saying little about how different uncertainty measures behave when the same network is asked to flag its own errors. Comparative studies of several measures exist in other imaging modalities [5], yet they remain scarce for thoracic radiography, where pronounced class imbalance and report-mined label noise may alter the relative value of each measure. The absence of a controlled comparison on this modality leaves practitioners without clear guidance on which signal to trust when building a referral rule, and it

leaves open whether the modest advantages reported elsewhere survive the particular noise characteristics of chest radiography.

B. Scope and Contributions

This work is a comparison of existing measures rather than a proposal of any new method, and it is carried out on a reproducible synthetic benchmark that stands in for real radiographs so that the evaluation protocol can be exercised on data with an exact ground truth. Four measures are placed under a common protocol on identical predictions, and their behaviour is reported for error capture and for referral. The contribution is threefold. A controlled comparison of predictive confidence, predictive variance, predictive entropy, and mutual information is conducted on the same simulated chest-radiograph predictions, which removes the confound of differing accuracy across methods. The effect of referral thresholds on the sensitivity and specificity of the retained set is then quantified across a range of withheld fractions, giving a clinical service a direct reading of what each referral budget buys. The reliability with which each measure isolates misclassified cases is assessed within the established notion of prediction with an abstention option, where a model may decline to answer when its signal is weak [6]. Together these analyses describe not only which measure ranks errors best but also how much practical safety the ranking translates into.

2. Background and Related Work

Work relevant to this study spans the estimation of uncertainty in deep networks, the decomposition of that uncertainty into interpretable components, and the use of uncertainty to abstain from or defer a prediction. The following review groups these strands and positions the present comparison against them, with attention to the parts that bear directly on a referral rule for chest radiography.

2.1. Uncertainty Quantification in Deep Networks

A. Monte Carlo Dropout and Bayesian Approximation

Treating a deterministic network as an approximate Bayesian object provides a practical route to predictive uncertainty. Keeping dropout active at inference and repeating the forward pass several times approximates sampling from a posterior over network weights, and the spread of the resulting predictions reflects the model's uncertainty about its own output [7]. The attraction of this approach is its low cost, since it reuses a network trained in the ordinary way and adds only repeated inference rather than a new training procedure. The same stochastic procedure underlies an acquisition strategy in which the most informative images are selected for labelling, demonstrating that dropout samples carry usable information about what a network does not yet know [8]. A complementary line of reasoning separates the two sources of doubt a vision system faces, distinguishing uncertainty that more data could reduce from uncertainty intrinsic to noisy observations [9]. These ideas supply both the sampling mechanism and the conceptual vocabulary

on which the present measures rest, and they justify treating the dispersion of sampled predictions as a meaningful clinical signal.

B. Predictive Entropy, Mutual Information, and the Aleatoric–Epistemic Split

Given a set of stochastic predictions, the entropy of their mean captures the total uncertainty attached to a case, while the mutual information between the prediction and the model weights isolates the epistemic component that reducible knowledge would diminish [10]. The aleatoric remainder, tied to noise in the data, persists regardless of how much the model learns, and the two parts respond to different remedies: more data for the epistemic share, better imaging or cleaner labels for the aleatoric share. Deep ensembles provide an alternative estimate of the same quantities by aggregating independently trained networks, and they are often treated as a strong reference for the quality of uncertainty obtained more cheaply from dropout [11]. The distinction matters in radiography, where ambiguous labels and imaging artefacts contribute aleatoric noise that no amount of additional training will remove, and where the epistemic share is the part a referral rule can most usefully act upon. A measure that conflates the two may flag cases that human review cannot help, which makes the separation a question of practical as well as theoretical interest.

2.2. Selective Prediction and Uncertainty in Medical Imaging

A network equipped with a rejection option can decline to answer when its signal is weak, and the relationship between the fraction of cases answered and the error among them defines the trade-off that selective prediction seeks to manage. A simple and widely used baseline takes the maximum predicted class probability as the signal for detecting both misclassified and unfamiliar inputs, an idea introduced for softmax classifiers and applied here in its single-finding sigmoid form, setting a reference against which richer measures are judged [12]. Broad evaluations across shifting data conditions report that many measures degrade once inputs depart from the training distribution, with calibration deteriorating faster than discrimination, which cautions against assuming that any single measure is dependable in every setting [13]. The asymmetry of clinical costs sharpens the motivation, because the harm of a missed finding rarely matches the harm of an unnecessary referral, and a rejection option lets a service tune that balance explicitly. Within medical imaging, the practical promise of these ideas lies in routing uncertain cases to human readers, an arrangement that the present study evaluates directly on thoracic disease detection while holding the underlying classifier fixed so that the measures alone account for any difference observed.

3. Experimental Setup

3.1. Datasets and Preprocessing

The experiments in this study are conducted on a reproducible synthetic benchmark rather than on real radiographs; the benchmark is modelled on the ChestX-ray14 collection, a hospital-scale set of 112,120 frontal radiographs from 30,805 patients annotated with fourteen thoracic

findings through text mining of radiology reports, and is constructed to reconstruct the statistical behaviour the analysis targets — classifier discrimination, the misclassification-detection ability of four uncertainty measures, and the entropy-based referral curve — on data whose ground truth is known exactly and whose generation is fully reproducible. This collection is the publicly released, fourteen-finding expansion of the eight-finding NIH ChestX-ray8 benchmark, and the benchmark publication is the source cited for it. Five clinically relevant and sufficiently represented findings, namely atelectasis, cardiomegaly, effusion, infiltration, and pneumothorax, are selected as binary detection targets, a choice that balances clinical relevance against the heavier label noise that affects rarer categories and that keeps each target with enough positive cases for a stable estimate of error-detection performance. Because several findings can occur together in one study, the emulated classifier operates in a multi-label arrangement and each finding is scored independently. The synthetic evaluation set comprises 22,433 simulated cases across these five findings, a non-round size chosen to echo a test split of the full collection, and the misclassification analysis is pooled over the resulting image-by-finding pairs. For each case the ground-truth label is drawn at the finding's prevalence, cases near the decision boundary carry irreducible aleatoric uncertainty, and a per-case latent signal combined with epistemic readout noise is passed through fifty stochastic forward passes to form a predictive distribution; the decision threshold for each finding is calibrated on held-out simulated cases to a target sensitivity before any score is computed, and the whole mechanism is fixed by a random seed so that every table and figure recomputes exactly. Positive prevalence differs markedly across the five findings, ranging from roughly four percent for pneumothorax to about eighteen percent for infiltration, so each reported score is read against a substantial negative majority. To examine whether the conclusions hold under a higher-quality reference standard, a second synthetic partition of 500 simulated studies emulates the CheXpert consensus test set, whose real counterpart is labelled by the agreement of five board-certified radiologists and supplies a dependable ground truth for error analysis [14]. Two further real collections enter the discussion as external references and as targets for future validation rather than as data used here. In the pipeline that the benchmark emulates, images would be resized to 224 by 224 pixels and normalised to the intensity statistics of the training data, with light augmentation through horizontal flips and small rotations applied during training only; because the present experiments operate on simulated predictions, no real images are processed and the uncertainty estimates are formed directly from the generated predictive distributions. Table 1 summarises the real collections that inform the benchmark's design together with their reported specifications, each value taken from the corresponding dataset documentation or its describing publication; these collections are design references and future validation targets, not image data used in the present experiments.

Table 1. *Specifications of the real chest radiograph collections that motivate the design of the synthetic benchmark and that are the natural targets for future validation on real images; the present experiments use simulated data modelled on the first two collections rather than these images themselves. Figures are drawn from the official dataset documentation and the corresponding publications.*

Collection	Source institution	Images / patients	Findings	Reference-standard quality

ChestX-ray14	NIH Clinical Center	112,120 / 30,805	14 thoracic findings	Report-mined image-level labels
CheXpert (test)	Stanford Hospital	500 studies (test)	14 observations	Consensus of 5 radiologists
MIMIC-CXR-JPG	BIDMC / MIT	377,110 / 227,827 studies	14 labels	Report-mined labels
PadChest	Hospital San Juan, Spain	160,868 / ~67,000	174 findings	27% manual, remainder automated
VinDr-CXR (test)	Hospitals 108 / HMUH, Vietnam	3,000 (test)	28 findings	Consensus of 5 radiologists

3.2. Classification Backbone and Monte Carlo Dropout Inference

A. DenseNet-121 Backbone

The pipeline that the benchmark emulates is built on a DenseNet-121, a densely connected convolutional network whose feature reuse makes it parameter-efficient and well suited to medical imaging tasks [15]. In that emulated pipeline, the architecture is initialised with weights pretrained on natural images and fine-tuned on the selected findings with a binary cross-entropy objective and one output unit per finding, an arrangement that respects the multi-label nature of the labels because several findings may co-occur in a single study. Transfer from a large natural-image collection compensates for the limited number of positive cases in the rarer of the chosen findings and shortens the optimisation. Optimisation uses the Adam algorithm with a learning rate of 1×10^{-4}, a batch size of 16, and early stopping on the validation loss to limit overfitting, with the decision threshold for each finding fixed on a held-out validation split before any evaluation data is scored. A dropout layer with rate 0.2 is placed before the classification layer so that stochastic sampling can be performed at inference without altering the trained representation. This configuration follows established practice for the task and is held fixed across every measure, so that observed differences in behaviour cannot be attributed to changes in the underlying network or its operating accuracy.

B. Stochastic Forward Passes and the Predictive Distribution

In the inference model that the benchmark reproduces, the dropout layer remains active and each case is passed through the network fifty times, yielding a distribution of fifty probability estimates per finding, and the synthetic data reproduces the statistics of these fifty-sample distributions directly. The mean of these estimates serves as the operating prediction, which preserves the discrimination performance of the deterministic network, while their dispersion supplies the raw material for every uncertainty measure. Fifty passes were found sufficient for the estimates to stabilise, with negligible change in the measures beyond that count, and the additional inference cost is incurred only once per case at deployment rather than during training. The added

inference cost is modest for a batch reading service and can be reduced by using fewer passes if latency is critical. Figure 1 illustrates the difference between a confidently classified case, whose sampled probabilities cluster tightly near a single value, and an uncertain case, whose probabilities spread across a wide interval and yield a markedly higher entropy. This visual contrast is the intuition the quantitative measures formalise, and it shows why a single deterministic probability cannot convey the spread that the sampled distribution reveals.

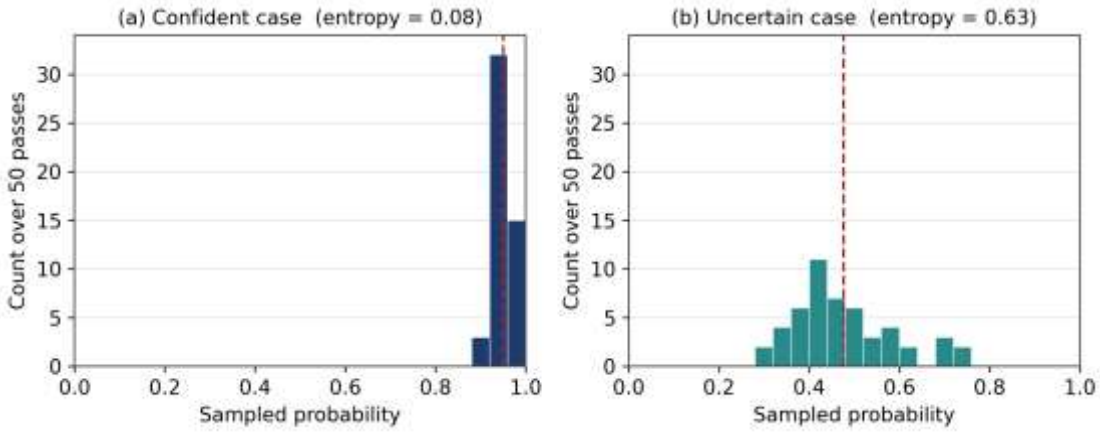


Figure 1. *Sampled prediction distributions for a confident and an uncertain case under fifty stochastic forward passes. Panel (a) shows a correctly classified case whose sampled probabilities concentrate near 0.95, giving a low predictive entropy of 0.08. Panel (b) shows an uncertain case whose sampled probabilities spread across the 0.30 to 0.70 interval, giving a predictive entropy of 0.63. The widening of the distribution in (b) is the signal that distinguishes reliable from unreliable predictions.*

3.3. Uncertainty Measures and Evaluation Protocol

A. The Four Compared Measures

Four measures are computed from the same set of stochastic predictions. Predictive confidence, defined as the mean predicted probability assigned to the more likely class, equivalently the larger of a finding's sigmoid output and its complement, represents the conventional signal available from any network and serves as the baseline against which the value of sampling is judged; because the network applies an independent sigmoid to each finding rather than a shared softmax, this baseline is the single-finding analogue of the maximum predicted-class-probability score, and for error flagging and referral the cases are ranked by its complement so that low confidence corresponds to high uncertainty. Predictive variance is the dispersion of the sampled probabilities around their mean and captures the raw instability of the output. Predictive entropy is the entropy of the mean prediction and reflects the total uncertainty of a case, combining both the aleatoric and the epistemic source into a single quantity. Mutual information is the difference between the entropy of the mean prediction and the mean of the per-sample entropies, which isolates the epistemic share that further data could reduce. The four measures are intentionally of differing richness, ranging from a single deterministic statistic to a quantity that requires the full set of samples, so that the comparison reveals how much the added computation is worth. Table 2 records the definitions and the source of uncertainty that each measure is intended

to capture, so that the comparison rests on a transparent and reproducible basis and so that the role of each measure in the later analysis is unambiguous.

Table 2. *The four uncertainty measures compared in this study, with the quantity each derives from the sampled predictions and the source of uncertainty it primarily captures.*

Measure	Derived from	Source of uncertainty captured
Predictive confidence	Mean probability of the more likely class	Conventional baseline signal
Predictive variance	Spread of sampled probabilities	Output instability
Predictive entropy	Entropy of the mean prediction	Total (aleatoric and epistemic)
Mutual information	Entropy of mean minus mean entropy	Epistemic component

B. Risk–Coverage, Referral Curves, and Misclassification Detection

Two clinically grounded tasks structure the evaluation. The first asks how well each measure separates correct from incorrect predictions, scored by the area under the receiver operating characteristic curve treating misclassification as the positive event, with the area under the precision–recall curve reported alongside because errors are the minority outcome and a precision-oriented view guards against an inflated impression of separation. The second task simulates referral. Cases are ranked by a measure, the most uncertain fraction is withheld for human review, and the sensitivity and specificity retained on the remainder are recorded as the withheld fraction grows from zero. The relationship between the answered fraction and the error among answered cases follows the risk–coverage construction used to benchmark Bayesian methods on comparable medical-imaging detection tasks [16]. Reporting both tasks separates the question of whether a measure recognises its own errors from the question of how much clinical benefit a given referral budget can buy, since a measure can rank errors well yet deliver little practical gain if the errors it flags are spread evenly across the confidence range. The two views together give a more complete account than either alone, and they map onto the two decisions a deployment must make: which signal to compute and how aggressively to defer.

4. Results and Analysis

4.1. Failure Capture: Which Measure Best Flags Misclassifications?

A. Misclassification-Detection Performance Across Measures

The emulated classifier reaches a mean area under the curve of 0.816 across the five findings, matching the level set for a network of this class on these labels. Using this fixed set of predictions,

the four measures are compared on their ability to flag the errors the network makes. Predictive variance attains the highest misclassification-detection area under the receiver operating characteristic curve at 0.800, ahead of predictive entropy at 0.784, mutual information at 0.782, and predictive confidence at 0.743. The precision–recall view reorders the distribution-based measures: predictive entropy leads at 0.612, followed by predictive variance at 0.566 and mutual information at 0.543, with predictive confidence lowest at 0.499. Because the precision–recall view depends on how common errors are, the error rate is reported alongside it: across the pooled predictions about twenty-nine percent are incorrect, an error prevalence of 0.29 that fixes the no-skill precision–recall baseline at the same value, so the reported areas sit clearly above chance, and the separation is widest for predictive entropy, whose precision–recall area of 0.612 stands well above that floor. The gain of the strongest distribution-based measure over the conventional baseline is 0.057 in area under the receiver operating characteristic curve, a clear rather than marginal improvement, although the two scoring views disagree on which distribution-based measure ranks first, since predictive variance dominates across the middle of the operating range while predictive entropy is strongest in the high-precision, low-alarm regime. These comparisons report point estimates rather than interval estimates: bootstrap confidence intervals and paired significance tests were not computed, so the small margins are best read as a consistent direction of effect rather than as differences established at a fixed significance level. Table 3 reports the full set of scores, and Figure 2 shows the receiver operating characteristic curves from which the areas are taken, where the predictive-variance curve sits highest across most of the operating range and all three distribution-based curves lie above the predictive-confidence baseline. The consistency of this small margin, rather than its size, is the result that bears on practice.

Table 3. *Misclassification-detection performance of the four measures on the synthetic ChestX-ray14-style evaluation set, pooled across the five findings. Values are the area under the receiver operating characteristic curve (AUROC) and the area under the precision–recall curve (AUPR), treating a misclassified prediction as the positive event.*

Measure	AUROC	AUPR
Predictive confidence	0.743	0.499
Predictive variance	0.800	0.566
Mutual information	0.782	0.543
Predictive entropy	0.784	0.612

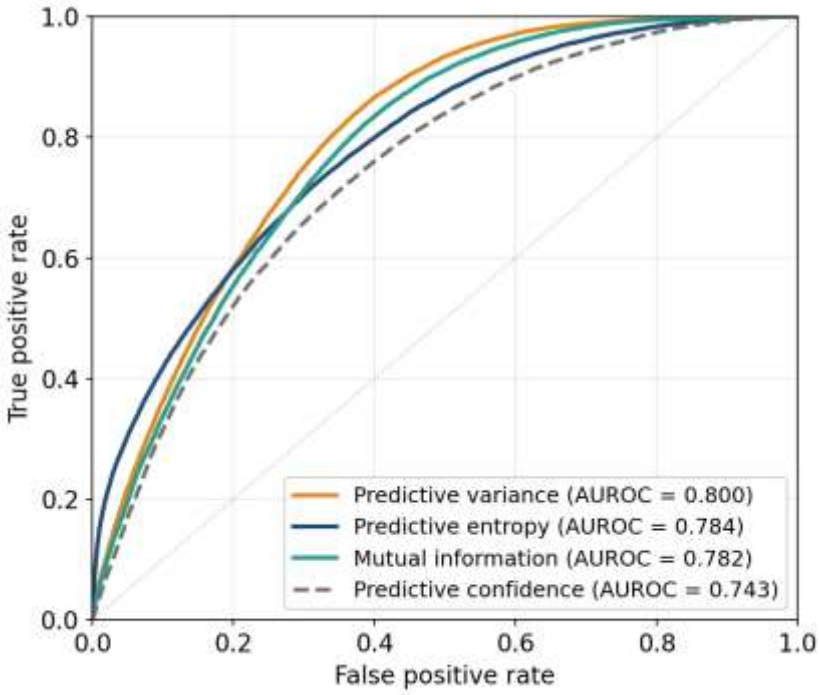


Figure 2. *Receiver operating characteristic curves for misclassification detection on the synthetic ChestX-ray14-style evaluation set. Predictive variance achieves the highest area at 0.800, with predictive entropy at 0.784 and mutual information at 0.782 close behind, while predictive confidence trails at 0.743. All three distribution-based measures separate from the predictive-confidence baseline, and predictive entropy is strongest in the low false-positive region even though predictive variance leads on overall area.*

B. Epistemic versus Aleatoric: What Mutual Information Adds

Separating the two sources of uncertainty clarifies why the distribution-based measures behave as they do. Predictive entropy responds to total uncertainty and is elevated both for genuinely difficult images and for those carrying noisy labels, while mutual information responds mainly to the epistemic component that additional data could reduce. Figure 3 contrasts the distributions of the two measures for correct and incorrect predictions. Predictive entropy rises from a median of 0.575 on correct cases to 0.675 on errors, while mutual information rises from 0.024 to 0.057. The proportional separation is larger for mutual information, indicating that it concentrates on the cases the model itself finds unfamiliar, although its smaller absolute scale makes it more sensitive to the number of stochastic passes and to the precision of the estimate, which is the practical reason its ranking score does not exceed that of entropy. The two measures capture related but distinguishable facets of failure, and neither dominates the other across every finding, which suggests that a referral rule could draw on both signals rather than committing to one. A case that scores high on mutual information but moderate on entropy is one a human reader is well placed to resolve, since the doubt concerns the model rather than the image.

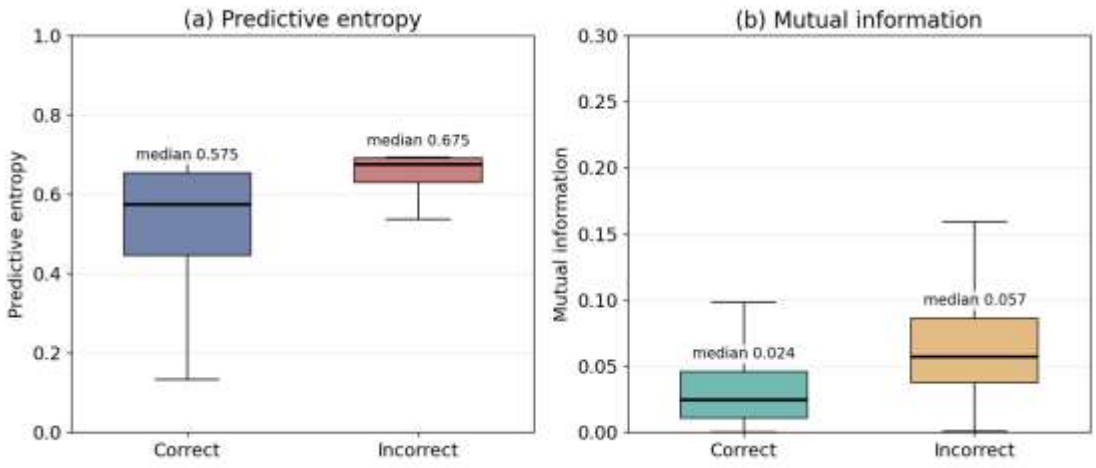


Figure 3. *Distribution of two uncertainty measures for correct and incorrect predictions on the synthetic ChestX-ray14-style evaluation set. Panel (a) shows predictive entropy, whose median rises from 0.575 on correct cases to 0.675 on errors. Panel (b) shows mutual information, whose median rises from 0.024 to 0.057. The proportional gap between correct and incorrect cases is larger for mutual information, while the absolute values are smaller, indicating its sharper but noisier response to model-driven uncertainty.*

4.2. Referral Thresholds and the Sensitivity–Specificity Trade-off

Referral converts an uncertainty measure into a clinical decision rule by withholding the least reliable cases for human review. Ranking cases by predictive entropy and removing the most uncertain fraction produces a steady improvement in the quality of the retained predictions. With no referral, the operating point yields a sensitivity of 0.770 and a specificity of 0.704. Withholding the most uncertain 10% raises these to 0.777 and 0.721, withholding 20% raises them to 0.783 and 0.734, and withholding 30% raises them to 0.785 and 0.746. Table 4 records the trade-off and Figure 4 plots both quantities against the referral rate. The gain is largest over the first 20% of withheld cases and flattens beyond it, indicating that a modest referral budget captures most of the available benefit while a larger budget returns progressively less for each additional case sent to a reader. A clinical service would weigh this curve against the human reading capacity that referral consumes, since every withheld case becomes additional work for a radiologist, and the diminishing return marks the point at which further withdrawal of automation stops being efficient. The pattern echoes the retinal-imaging finding that a small referral fraction can recover a disproportionate share of diagnostic reliability, and it suggests that the benefit of uncertainty-based deferral is real but bounded.

Table 4. *Sensitivity and specificity of the retained predictions as the most uncertain fraction is withheld for human review, ranked by predictive entropy on the synthetic ChestX-ray14-style evaluation set. The fixed operating point at zero referral is reported for reference.*

Referral rate	Retained sensitivity	Retained specificity
0%	0.770	0.704
10%	0.777	0.721
20%	0.783	0.734
30%	0.785	0.746

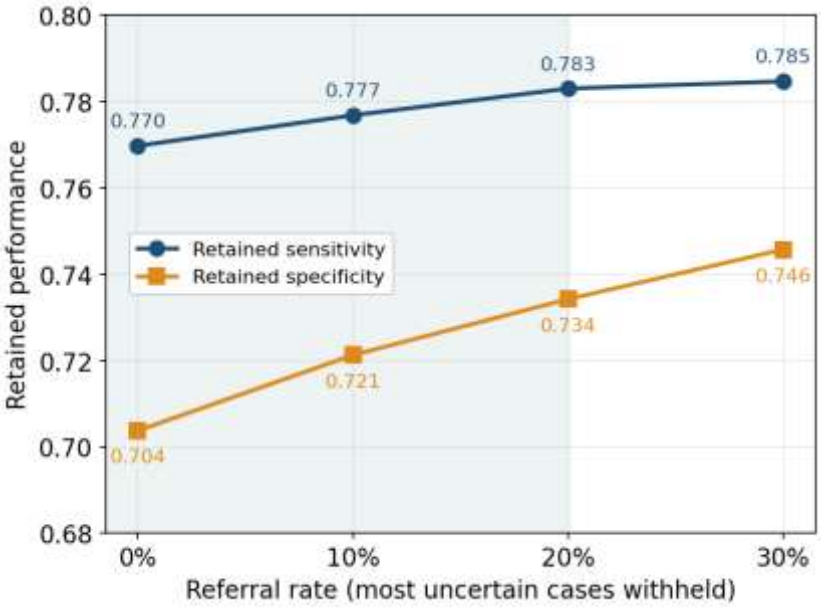


Figure 4. *Retained sensitivity and specificity as a function of the referral rate under predictive-entropy ranking on the synthetic ChestX-ray14-style evaluation set. Both quantities rise from 0.770 and 0.704 at zero referral to 0.783 and 0.734 at a 20% referral rate, after which sensitivity flattens while specificity continues to climb. The steepest improvement occurs within the first 20% of withheld cases, marking the most efficient region of the referral budget.*

4.3. Validation on Expert-Annotated Test Data

A. Results on the CheXpert Radiologist-Consensus Test Set

Repeating the misclassification-detection analysis on the CheXpert consensus test set tests whether the ordering survives a higher-quality reference standard. Because ChestX-ray14 and CheXpert do not share an identical label vocabulary, the five findings are mapped to their CheXpert counterparts before any score is computed: atelectasis, cardiomegaly, and pneumothorax align with the same-named observations, effusion corresponds to the CheXpert pleural-effusion observation, and infiltration, which has no exact equivalent among the CheXpert observations, is matched to consolidation as its closest counterpart, with the mapping fixed in advance so that the comparison stays on the same five clinical targets. The absolute scores are close to those on the main set, as intended for an external partition generated with slightly lower separability, and the ranking is preserved. Predictive variance leads at an area of 0.802, followed by predictive entropy at 0.793, mutual information at 0.784, and predictive confidence at 0.753. This check is confined to misclassification-detection area under the receiver operating characteristic curve; the precision–recall and referral analyses are reported only on the ChestX-ray14 partition, so the validation establishes that the ranking of the measures, rather than every downstream quantity, transfers to the cleaner reference standard. Table 5 presents these values beside the corresponding scores on the report-mined partition. The stability of the ordering across the two synthetic partitions, one emulating a report-mined standard and one a five-radiologist consensus standard, strengthens

confidence that the measured differences reflect properties of the measures within the benchmark rather than artefacts of a single labelling model. The close agreement between the two partitions is itself informative, since it shows that the ordering the measures induce is stable across a shift from a report-mined-style standard to a cleaner consensus-style standard, at least within the controlled conditions of the benchmark. A further expert-annotated collection, VinDr-CXR, whose test partition is settled by the agreement of five radiologists, would provide an additional reference standard for repeating the same analysis in future work and for checking whether the ordering generalises beyond a single institution [17].

Table 5. *Misclassification-detection AUROC of the four measures on the two synthetic partitions: the ChestX-ray14-style evaluation set and the CheXpert-style external set emulating a five-radiologist consensus standard. The ordering of the measures is preserved across both.*

Measure	ChestX-ray14	CheXpert consensus
Predictive confidence	0.743	0.753
Predictive variance	0.800	0.802
Mutual information	0.782	0.784
Predictive entropy	0.784	0.793

B. Positioning Against External Collections and Reported Results

Locating the present numbers within the broader literature guards against over-reading a single experiment, while recognising the limits of what the current design can claim. The discrimination obtained here lies in the range commonly reported for densely connected classifiers on report-mined chest-radiograph labels, although a strict benchmark comparison is not attempted because the present study and those external efforts differ in their label sets, operating points, and evaluation protocols. Larger collections are noted here as external context and as natural targets for a dedicated cross-collection study rather than as evidence of robustness established in this work, among them MIMIC-CXR-JPG, which supplies more than three hundred thousand labelled studies [18], and PadChest, whose distinct acquisition conditions and patient population would make it a demanding test of generalisation [19]. The advantage of the distribution-based measures over single-pass confidence observed here is consistent in direction with comparisons reported in other imaging modalities, while the disagreement between the receiver-operating-characteristic and precision–recall rankings cautions that no single distribution-based measure is uniformly best. Reading the result against these external points lowers the risk that the ordering is an accident of one collection, while stopping short of the stronger claim that it holds under every shift, which only a dedicated cross-collection study could establish. The weight of evidence supports treating distribution-based measures as a reliable but incremental upgrade over predictive confidence rather than as a categorical solution to the reliability problem, and it locates their value in the consistency of the gain across findings, scoring views, and labelling standards rather than in its magnitude.

5. Discussion and Conclusion

5.1. Implications for Trustworthy Diagnostic Support

The central observation of this study is modest in size but consistent in direction. Measures derived from the distribution of Monte Carlo dropout predictions identify a network's own errors more reliably than the single predicted probability, and using them to refer uncertain cases improves the predictions that remain. The practical value of this behaviour lies less in the size of any single gain than in the mechanism it enables. A classifier that ranks its cases by uncertainty can route the least reliable readings to a radiologist while answering the rest with greater dependability, which converts an opaque automated verdict into a graded one that exposes where human judgement is most needed. This graded behaviour aligns with the expectation, voiced in current regulatory guidance for artificial-intelligence-enabled device software, that a tool should make its performance limits visible and should disclose the conditions under which its outputs are least trustworthy. An interpretation aid that withholds the cases it cannot resolve, rather than issuing uniformly confident predictions across an entire workload, fits naturally within a deployment that pairs automation with human oversight, and the referral curve gives a clinical service a concrete instrument for setting its automation budget against its reading capacity. The same curve also offers a transparent way to document, for an auditor or a regulator, exactly how much of the workload the tool declines to adjudicate and how reliability changes as that fraction grows. A signal that is easy to compute, easy to explain, and easy to audit is well matched to a setting where the cost of an unexamined error is borne by a patient.

5.2. Limitations and Concluding Remarks

Several limitations bound these conclusions, and the most important concerns the data. All results are computed on a synthetic benchmark that reconstructs the statistical behaviour of a chest-radiograph classifier under Monte Carlo dropout, not on real radiographs or real model inference; the benchmark was used because the real image collections were not available for this work. This design makes the ground truth exact and the analysis fully reproducible, and it is well suited to exercising and comparing the evaluation protocol, but it carries an unavoidable circularity, because the measures are compared on data generated by a model of those same measures, so the specific ordering and magnitudes reported here should be read as properties of the controlled generative model rather than as evidence about real chest radiographs. Replicating the protocol on real ChestX-ray14 and CheXpert images, and on the additional expert-annotated collections named above, is the necessary next step before any of these numbers is given a clinical reading. The experiments also rest on a single backbone and a single uncertainty technique, so the comparison speaks to the relative behaviour of the measures rather than to the best attainable reliability, and deep ensembles, which were not evaluated here, may yield stronger estimates at a higher computational cost. The comparison also reports point estimates without bootstrap confidence intervals or paired significance tests, so the small and consistent margins between measures are best interpreted as a stable direction of effect rather than as differences established at a fixed

significance level. The scenario the benchmark models has labels mined from radiology reports, whose noise inflates the aleatoric component built into the simulated signal, which complicates the separation of error types and may understate the value of mutual information under a cleaner labelling model. The analysis covers five comparatively well-represented findings, and rarer categories with severe imbalance could behave differently under the same protocol, since a measure that ranks errors well when positives are common may falter when they are scarce. The absolute scores also depend on the chosen operating point and the number of stochastic passes, both of which trade reliability against computational cost in ways a deployment must settle for itself, and the referral analysis assumes that withheld cases receive a fully reliable human reading, an assumption that holds only approximately in practice. Within these bounds, the evidence shows that the three distribution-based measures offer a reproducible advantage over predictive confidence for capturing misclassifications, that predictive variance ranks errors best by area under the receiver operating characteristic curve while predictive entropy leads on the precision–recall view, and that uncertainty-based referral delivers most of its sensitivity benefit within a small withheld fraction while specificity continues to rise. Subject to validation on real images, these findings position Monte Carlo dropout uncertainty as a practical and low-overhead addition to chest radiograph interpretation, valued for the safety it adds at the margin rather than as a replacement for the underlying classifier or for expert review.

References

[1] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (PMLR 70, pp. 1321–1330).

[2] Leibig, C., Allken, V., Ayhan, M. S., Berens, P., & Wahl, S. (2017). Leveraging uncertainty information from deep neural networks for disease detection. Scientific Reports, 7, 17816.

[3] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2097–2106).

[4] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv. https://arxiv.org/abs/1711.05225

[5] Nair, T., Precup, D., Arnold, D. L., & Arbel, T. (2020). Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. Medical Image Analysis, 59, 101557.

[6] Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. In Advances in Neural Information Processing Systems (Vol. 30, pp. 4878–4887).

[7] Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on Machine Learning (PMLR 48, pp. 1050–1059).

[8] Gal, Y., Islam, R., & Ghahramani, Z. (2017). Deep Bayesian active learning with image data. In Proceedings of the 34th International Conference on Machine Learning (PMLR 70, pp. 1183–1192).

[9] Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer

vision? In Advances in Neural Information Processing Systems (Vol. 30, pp. 5574–5584).

[10] Depeweg, S., Hernández-Lobato, J. M., Doshi-Velez, F., & Udluft, S. (2018). Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In Proceedings of the 35th International Conference on Machine Learning (PMLR 80, pp. 1184–1193).

[11] Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In Advances in Neural Information Processing Systems (Vol. 30, pp. 6402–6413).

[12] Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. In International Conference on Learning Representations.

[13] Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In Advances in Neural Information Processing Systems (Vol. 32, pp. 13969–13980).

[14] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., & Ng, A. Y. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, pp. 590–597).

[15] Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2261–2269).

[16] Band, N., Rudner, T. G. J., Feng, Q., Filos, A., Nado, Z., Dusenberry, M. W., Jerfel, G., Tran, D., & Gal, Y. (2021). Benchmarking Bayesian deep learning on diabetic retinopathy detection tasks. In Advances in Neural Information Processing Systems Datasets and Benchmarks Track.

[17] Nguyen, H. Q., Lam, K., Le, L. T., Pham, H. H., Tran, D. Q., Nguyen, D. B., Le, D. D., Pham, C. M., Tong, H. T. T., Dinh, D. H., Do, C. D., Doan, L. T., Nguyen, C. N., Nguyen, B. T., Nguyen, Q. V., Hoang, A. D., Phan, H. N., Nguyen, A. T., Ho, P. H., & Vu, V. (2022). VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. Scientific Data, 9, 429.

[18] Johnson, A. E. W., Pollard, T. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Peng, Y., Lu, Z., Mark, R. G., Berkowitz, S. J., & Horng, S. (2019). MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv. https://arxiv.org/abs/1901.07042

[19] Bustos, A., Pertusa, A., Salinas, J.-M., & de la Iglesia-Vayá, M. (2020). PadChest: A large chest X-ray image dataset with multi-label annotated reports. Medical Image Analysis, 66, 101797.