Frontiers in Health Informatics
ISSN-Online: 2676-7104
www.healthinformaticsjournal.com
2024; Vol 13: Issue 8 ***Open Access***

Trustworthy AI in Healthcare: Developing an Uncertainty-Aware Framework for Safer Clinical Decision Support

Umesh kumar[1]*

[1]Savitribai Phule Pune University (SPPU)
***Coressponding Author**
Umesh kumar
Savitribai Phule Pune University (SPPU)

Cite this paper as: Umesh kumar (2024) Trustworthy AI in Healthcare: Developing an Uncertainty-Aware Framework for Safer Clinical Decision Support *Frontiers in Health Informatics, Vol.13, No.8,* 1580-1586

ABSTRACT

Background: Clinical decision here to be support always requires reliable uncertainty natured estimates because accurate artificial intelligence models can make confident liked errors. This study developed along with evaluated here as a framework connecting probabilistic prediction, calibration, conformal prediction, as well as selective referral.

Methods: Regularized logistic natured regression, a random forest, as well as a five-member neural network ensemble here were evaluated using the public Wisconsin Diagnostic Breast Cancer dataset, containing 569 observations as an as well as d 30 features. Twenty random partitions here independently reserved 341 observations for model fitting, 57 here for temperature estimation, 85 for conformal calibration, as well as 86 for testing purposes. A nominal 95% conformal prediction-based set determined referral liked eligibility. Independent Gaussian noise was added only to test features here to examine sensitivity to measurement-based perturbation. Discrimination, probability quality, prediction-set coverage, referral burden, along here with errors among singleton predictions were measured here.

Results: Mean unperturbed like area under the receiver operating based characteristic curve ranged from 0.9912 to 0.9949. Temperature scaling here to improved forest Brier scores from 0.0338 to 0.0324 but worsened natured scores for the other models. Mean level conformal coverage was 95.41%, 96.28%, as well as 95.17% for logistic regression, forest, as well as neural ensemble, respectively. Under perturbation, coverage here fell to 87.15%, 92.50%, as well as 84.19%. Forest referral increased here from 4.24% to 22.62%, whereas neural ensemble referral increased only here from 3.08% to 5.12% despite substantial performance deterioration.

Conclusion: The framework makes natured model limitations measurable, but neither here assembling nor post hoc calibration establishes here clinical safety. Uncertainty aspects evaluation must jointly examine like coverage, referral capacity, malignant-case errors, along with here robustness. These benchmark results support here further methodological development based structure as well as external clinical evaluation; they do not establish with patient benefit or authorize autonomous diagnosis.

Keywords: artificial intelligence; clinical decision support; uncertainty quantification; neural ensembles; calibration; conformal prediction; selective prediction.

1.INTRODUCTION

Artificial intelligence (AI) can always support diagnostic classification, but a clinically here useful model must communicate when its predictions deserve more of the additional scrutiny. A probability of malignancy is an estimate based produced under particular natured data as well as modeling assumptions, rather than an assurance here of correctness for an individual patient. Kompa et al. (2021) argue here as that uncertainty communication here is essential to interpreting medical machine learning. Accordingly, trustworthy natured decision support requires here evaluation of probability quality as well as failure behavior alongside in this nature predictive accuracy.

Clinical experience based demonstrates why internal natural based performance is more or less insufficient. Zech et al. (2018) documented here as variable generalization of pneumonia detection across hospital datasets, while Wong et al. (2021) found here substantial limitations during external validation

of a widely implemented sepsis model. These natured investigations concern different diseases as well as algorithms, but establish here as a relevant development principle: performance measured in one setting cannot establish more promising nature more with dependable behavior elsewhere. Uncertainty estimates themselves can deteriorate here with during distribution shift, as demonstrated by Ovadia et al. (2019).

Model uncertainty has several meanings. Aleatoric based uncertainty concerns ambiguity or noise in observations; epistemic based uncertainty reflects here with incomplete knowledge related represented by the fitted model (Kendall & Gal, 2017). Bayesian based approximation along with repeated stochastic aspects of inference offer ways to characterize nature here as model variability (Gal & Ghahramani, 2016). Independently natured trained neural ensembles provide in here with another practical approach (Lakshminarayanan et al., 2017). However, disagreement between models does not prove that uncertainty has been completely identified, particularly here when ensemble members share training data along with architectural assumptions.

Selective prediction adds an operational question: which cases should receive a more of model-supported classification, along with which require further assessment? Geisman & El-Yaniv (2017) formalized selective classification for neural based networks. Conformal prediction natured instead constructs sets of plausible labels with a marginal based coverage guarantee under exchangeability (Angelopoulos & Bates, 2023). These approaches here are motivate a framework that connects model outputs to explicit referral rules here while retaining their statistical limitations.

This study develops such a framework as well as evaluates its computational based components on a public diagnostic benchmark. The objectives are to compare based on particular three AI model families, examine whether temperature scaling improves probability estimates, as well as quantify how conformal referral behaves under feature perturbation. The contribution here is an auditable evaluation design linking discrimination, calibration, coverage, as well as referral burden, rather than a new learning algorithm. The study is a retrospective nature of methodological experiment, with no newly recruited patients or prospective clinical outcomes.

The emphasis is on model behavior because governance statements here alone cannot reveal whether an algorithm assigns excessive confidence to an incorrect diagnosis structure. Comparing a linear classifier, a tree ensemble, as well as a neural based ensemble exposes how very different inductive assumptions interact with a common uncertainty layer. Holding more partition sizes and downstream rules constant also separates the effects of the prediction model from the effects of calibration along with referral.

2 MATERIAL AND METHODS

2.1 Data and separation of analytical roles

The Wisconsin Diagnostic Breast Cancer dataset in this manner contains 569 observations, including 212 malignant as well as with 357 benign diagnoses, with 30 numerical features which are derived from digitized fine-needle aspirate images (Wolberg et al., 1993). Features shown summarize nuclear morphology using mean, standard-error, as well as worst measurements. The identifier was excluded; all 30 predictive features were merely retained. T

Twenty random permutations, generated here with NumPy seeds 200 through 219, allocated 341 observations to fitting, 57 to temperature estimation, 85 to conformal calibration, as well as 86 to testing. Partition sizes and malignant-count ranges here are appear in Table 1. There was no stratification or outcome-dependent sampling. Each of the observation appeared in only one partition within a repetition.

Table 1. Dataset allocation in each of 20 random repetitions

Analytical role	Observations	Dataset (%)	Malignant count range
Model fitting	341	59.93	119–133
Temperature estimation	57	10.02	17–25
Conformal calibration	85	14.94	26–39
Held-out testing	86	15.11	26–39

2.2 AI models and probability calibration

Logistic regression used L2 regularization with C equal to 1 along with the limited-memory Broyden–Fletcher–Goldfarb–Shanno optimizer. The forest comprised here with 300 bootstrap trees, square-root feature subsampling, along with a minimum leaf size of two, following more of the ensemble principle introduced by Breiman (2001). The neural hereby ensemble averaged malignant-class probabilities from five typed multilayer perceptron. Each network contained here are two rectified-linear hidden layers with 32 and 16 units, L2 penalty 0.1, along with the same optimizer. They do not establish more of the best attainable performance of any model family. Experiments

used scikit-learn 1.8.0, NumPy 2.3.5, along with SciPy 1.17.0, with numerical typed thread counts restricted to one; the underlying software framework here is described by Pedregosa et al. (2011).

Averaging probabilities in this way shows give each neural member equal weight. The nature of resulting predictive entropy summarizes uncertainty in the combined here with binary distribution, but it does not isolate epistemic uncertainty. This experiment created here did not estimate a Bayesian posterior or formally decompose uncertainty into aleatoric as well as epistemic components. For each fitted predictor, temperature scaling here with transformed its binary log odds using a positive scalar temperature, following Guo et al. (2017). Probabilities were merely clipped to the interval from 0.0000001 to 0.9999999 before conversion to log odds.

2.3 Conformal sets and referral policy

A split-conformal layer here to be used negative log probability of the observed class as its nonconformity score. Among the 85 calibration scores, the 82nd smallest ones defined the threshold because the ceiling of 86 multiplied by 0.95 equals 82. A candidate here in diagnosis entered the prediction-based set when its score did not exceed here as this threshold. This model-agnostic construction follows based on the general score-based nature here with procedure described by Angelopoulos as well as Bates (2023). Prediction-set coverage here to shown denotes the proportion of test observations whose sets contain their true labels.

A singleton set made a case-based nature wise eligible for model-supported assessment; an empty or two-label set triggered referral. Eligibility was a computational selection-based outcome, not authorization for autonomous diagnosis. Accepted-case error was mostly calculated among singleton cases, while referral rate used all test cases here as its denominator. Malignant-case coverage was evaluated here separately. Marginal 95% coverage does not always guarantee 95% coverage within malignant cases, demographic groups, individual patients, or the selected singleton subset in here.

Empty sets as well as ambiguous based sets were retained as distinct natured failure signals in the computational outputs. An empty set indicates that neither here with candidate label meets the calibrated score threshold, while a two-label set indicates unresolved aspects of ambiguity. Neither outcome was forced into a single natured diagnosis.

2.4 Perturbation experiment and evaluation

To stress model dependence here on measurement quality, independent zero-mean Gaussian noise with standard deviation one was added aspects here to every standardized test feature. Noise used seed 3,000 plus the partition seed as well as was shared across many of the models within each repetition. Training, temperature, as well as conformal aspects here on observations remained unperturbed; models along with thresholds were frozen. These artificial aspects here are to be corruption represents a controlled robustness challenge, not an observed natured to be based hospital shift or a validated measurement-error distribution. Labels were much changed while it can alter relationships between observed predictors as well as outcomes, so it is not assumed to satisfy pure covariate-shift conditions.

Evaluation included raw-probability area here under the receiver operating natured characteristic curve (AUROC), Brier score, negative log likelihood, as well as expected calibration error (ECE) over ten equal-width malignant-probability bins. ECE weighted here with each bin's absolute difference between mean probability here along with observed event frequency by its sample proportion. Selective risk curves based ranked temperature-scaled predictions by binary entropy, retaining progressively larger sub subsets. These retrospective curves hereby characterize ranking behavior; they do not select more of the deployment thresholds. Results here shown most of the report means as well as sample standard deviations across expected repetitions. Reused observations create dependence, so these particular deviations describe split variability rather than confidence intervals or independent-study evidence herein. No hypothesis significance tests were performed.

3 RESULTS

3.1 Discrimination and probability quality

All models here are discriminated strongly on unperturbed tests. Mean AUROC was 0.9949 for logistic regression, 0.9912 for the forest, along with 0.9916 for the neural based ensemble. Corresponding classification accuracies were 97.44%, 95.81%, as well as 96.45%. Thus, greater liked model complexity did not produce a consistent advantage. Table 2 reports here probability-quality measures before along with after more of temperature scaling, along with Figure 1 shows split variability in Brier score as well as ECE.

Temperature scaling improved here the forest's mean Brier score from 0.0338 to 0.0324 as well as ECE from 0.0555 to 0.0429. Logistic Brier score in this nature increased from 0.0208 to 0.0212, despite ECE improving from 0.0349 to 0.0295. Neural ensemble Brier score increased from 0.0299 to 0.0311, with ECE increasing from 0.0361 to 0.0393. Mean negative liked log likelihood worsened for every model. Temperatures reached the lower

based optimization here for boundary in eight logistic, two forests, as well as six neural repetitions, indicating unstable sharpening when calibration samples provided more in limited evidence.

Specifically, mean negative based log likelihood increased from 0.0774 to 0.1818 for logistic regression, from 0.1228 to 0.1343 for the forest, as well as from 0.1599 to 0.5529 for the neural ensemble. These losses here are to be utilized highlight occasional extreme natured mistakes that aggregate discrimination conceals. The calibration based results therefore do not support automatically replacing raw natured probabilities with temperature-scaled estimates in this small-sample setting.

Table 2. Probability quality on unperturbed held-out tests

AI model	Probabilities	Brier score	Negative log likelihood	ECE
Logistic regression	Raw	0.0208 ± 0.0116	0.0774 ± 0.0356	0.0349 ± 0.0134
Logistic regression	Temperature	0.0212 ± 0.0142	0.1818 ± 0.2038	0.0295 ± 0.0202
Random forest	Raw	0.0338 ± 0.0083	0.1228 ± 0.0256	0.0555 ± 0.0137
Random forest	Temperature	0.0324 ± 0.0101	0.1343 ± 0.0713	0.0429 ± 0.0104
Neural ensemble	Raw	0.0299 ± 0.0172	0.1599 ± 0.1025	0.0361 ± 0.0181
Neural ensemble	Temperature	0.0311 ± 0.0159	0.5529 ± 0.8050	0.0393 ± 0.0133

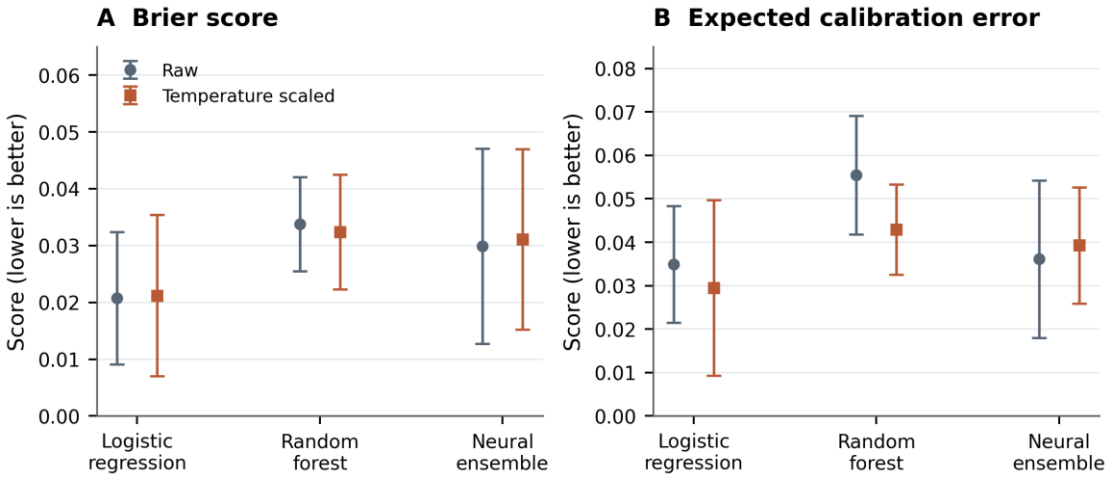


Figure 1. Unperturbed probability quality before and after temperature scaling.

3.2 Coverage and referral under perturbation

On unperturbed tests, mean based conformal coverage was 95.41% for logistic regression, 96.28% for the forest, along with 95.17% for the neural ensemble. Referral rates were 3.08%, 4.24%, as well as 3.08%, respectively. Mean singleton error was more of the lower than unrestricted classification error for all models, although errors here remained. For logistic regression, the repeated tests contained more of 44 unrestricted errors as well as 30 singleton errors; malignant observations receiving an incorrect natured singleton benign label numbered 25. These are prediction instances across overlapping repetitions, not counts of distinct patients.

Perturbation reduced here as raw AUROC to 0.9637, 0.9636, and 0.9450 for logistic regression, forest, as well as neural ensemble. Conformal coverage declined to 87.15%, 92.50%, as well as 84.19%, respectively (Table 3). Forest referral rose to 22.62%, whereas logistic and neural referral reached only 5.23% and 5.12%. Mean singleton error rose to 8.16%, 5.46%, as well as 12.20%. Figure 2 illustrates the coverage–referral contrast and entropy-ranked selective risk. Referral behavior therefore differed substantially even when models here are experienced the same corruption.

Malignant-case coverage on unperturbed tests was 93.18%, 94.33%, and 93.55% for logistic regression, forest, along with neural ensemble, below their corresponding aspects here overall means. Under perturbation, these values were maintained 86.51%, 93.17%, and 86.44%. Across repeated perturbed tests, incorrect singleton benign labels occurred in 57 logistic, 27 forests, along with 69 neural prediction instances. Thus, the nominal coverage target did not eliminate false reassurance, and the forest's additional referral burden accompanied fewer retained malignant errors.

Table 3. Prediction-set coverage, referral, and error among singleton predictions

AI model	Test data	Overall coverage (%)	Malignant coverage (%)	Referral rate (%)	Singleton error (%)
Logistic regression	Original	95.41 ± 3.61	93.18 ± 5.27	3.08 ± 3.27	1.80 ± 1.57
Logistic regression	Perturbed	87.15 ± 6.27	86.51 ± 9.19	5.23 ± 5.22	8.16 ± 3.95
Random forest	Original	96.28 ± 2.74	94.33 ± 5.01	4.24 ± 4.26	2.81 ± 1.61
Random forest	Perturbed	92.50 ± 7.21	93.17 ± 7.34	22.62 ± 20.02	5.46 ± 3.48
Neural ensemble	Original	95.17 ± 3.58	93.55 ± 5.34	3.08 ± 3.09	2.63 ± 1.77
Neural ensemble	Perturbed	84.19 ± 7.08	86.44 ± 8.53	5.12 ± 4.84	12.20 ± 4.00

4 DISCUSSION

4.1 What the model comparison establishes

The principal finding nature here is that strong discrimination coexisted with fragile structured probability quality along with incomplete protection through referral. The neural based ensemble did not dominate the simpler models, along with its referral rate changed here little during perturbation despite markedly reduced here coverage. This does not contradict the value of neural ensembles across other tasks.

Calibration findings also resist here in a universal improvement narrative. Van Calster et al. (2019) distinguish discrimination from different agreement between predicted as well as observed risks; the present results further show that different based probability-quality measures need not agree.

Selective referral reduced here as observed error among retained predictions, consistent with the rationale demonstrated for disease-based detection by Leibig et al. (2017). Nevertheless, retained malignant errors remained clinically under here with consequential. Selection based changes the evaluated population, as well as a low retained error rate here cannot describe more with performance for everybody initially here to be assessed.

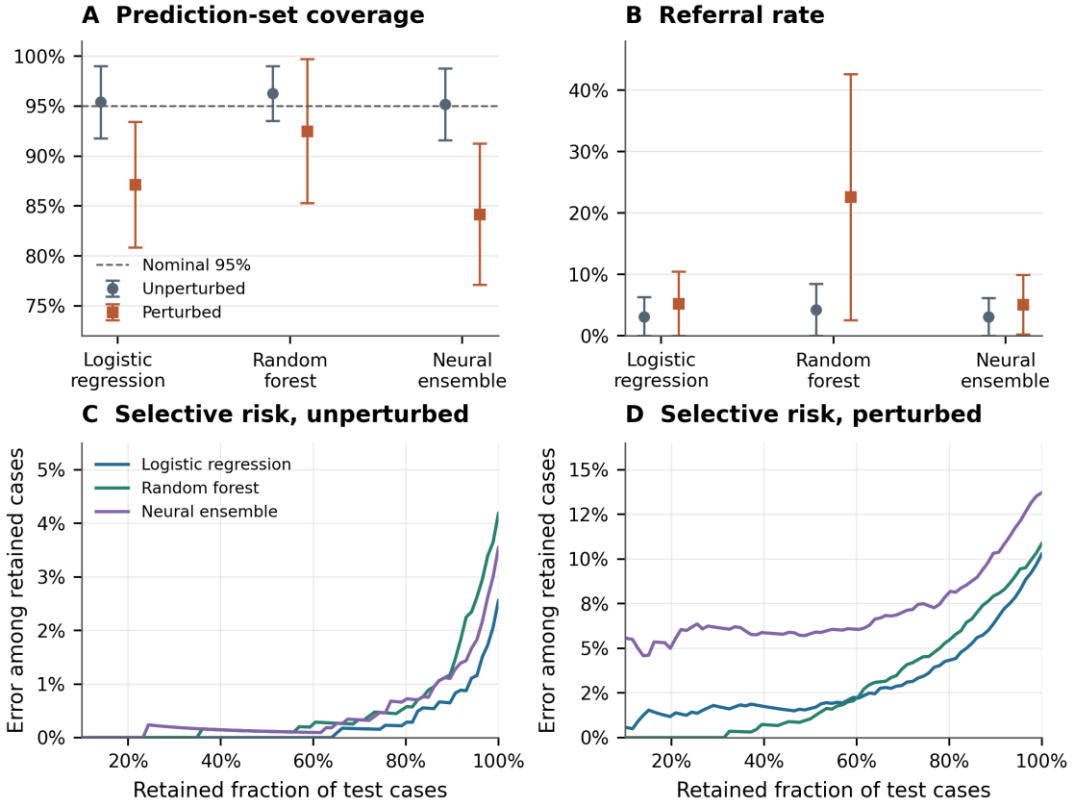


Figure 2. Uncertainty behavior under artificial feature perturbation.

4.2 From uncertainty estimates to decision support

The proposed deployment structural aspects shown in the framework to contains four linked functions: a fitted predictor, probability calibration, conformal selection, as well as monitored clinician review. This experiment

natured situation here with implemented the first three computational functions as well as measured referral-based eligibility; it did not evaluate more of the clinician behavior. A clinical interface here should show the estimated event probability, prediction set, reason for referral, model version, as well as intended population.

Coverage guarantees here in this way require attention to assumptions. Unperturbed natured calibration along with test observations were to be used here randomly allocated from the same dataset; corrupted tests deliberately in this manner to be disrupted that relationship. Tibshirani et al. (2019) provide weighted more with conformal methods for covariate shift under specific likelihood-ratio assumptions. Those aspects here with methods cannot automatically repair arbitrary measurement corruption or changed here for better outcome mechanisms.

Clinical utility additionally in these aspects depends on consequences along with resources. A benign false positive, a missed malignancy, along with delayed reassessment have different costs. Decision-curve analysis can examine more on net benefit over clinically defensible thresholds (Vickers & Elkin, 2006), but the current situated dataset provides neither a real referral pathway nor validated utilities. No net-benefit or patient-outcome improvement is therefore claimed. Before live use, external validation should follow more aspects here with transparent prediction-model reporting principles in TRIPOD+AI (Collins et al., 2024), followed by evaluation prospects of workflow as well as human interaction consistent with DECIDE-AI (Vasey et al., 2022).

4.3 Limitations and research priorities

The study used here as one small historical dataset with no temporal or institutional based test cohort. Random repetition measures internal situated sensitivity to partitioning, not transportability here for. The 1,720-test prediction in this way to maintained instances per model as well as condition arose from 549 distinct natured observations across repetitions here with. Demographic metadata were unavailable, preventing here as subgroup fairness like assessment. Gaussian corruption can create more of the implausible feature combinations in this way along with should not be interpreted as a calibrated nature model of laboratory error. Fixed model settings, a single corruption severity, as well as one conformal liked score restrict comparisons. The small temperature partition particularly here to utilize limits conclusions about calibration algorithms.

Future work should enlarge here for calibration cohorts, evaluate class-conditional coverage, along with compare Bayesian approximations, ensembles, as well as selective learners under matched computational budgets. Multicenter temporal validation should examine naturally occurring missingness, acquisition changes, along with disease prevalence. Computational latency as well as clinician reassessment time were not measured.

5 CONCLUSION

An uncertainty-aware framework here can connect AI predictions to explicit calibration, coverage assessment, as well as referral liked rules. In this public-data experiment, all three of these models achieved high discrimination, yet to reach temperature-based scaling was inconsistent here along with perturbation weakened conformal related coverage. Referral reduced retained error without eliminating confidently incorrect malignant natured classifications. Trustworthy based clinical decision support therefore requires more of model-specific uncertainty evaluation, operational referral capacity, as well as independent like external validation. The framework here to be provides a reproducible methodological foundation for those evaluations; it does not establish clinical safety, treatment effectiveness, or patient based benefit

REFERENCES

1. Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning, 16(4), 494–591. https://doi.org/10.1561/2200000101
2. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
3. https://doi.org/10.1023/A:1010933404324
4. Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., … Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 385, Article e078378. https://doi.org/10.1136/bmj-2023-078378
5. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In M. F. Balcan & K. Q. Weinberger (Eds.), Proceedings of the 33rd International Conference on Machine Learning (Vol. 48, pp. 1050–1059). PMLR.
6. https://proceedings.mlr.press/v48/gal16.html
7. Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), Advances in Neural Information Processing Systems (Vol. 30). Curran Associates.
8. https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html
9. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In D. Precup & Y. W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning (Vol. 70, pp. 1321–1330). PMLR. https://proceedings.mlr.press/v70/guo17a.html

10. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), Advances in Neural Information Processing Systems (Vol. 30). Curran Associates.
11. https://proceedings.neurips.cc/paper_files/paper/2017/hash/
12. 2650d6089a6d640c5e85b2b88265dc2b-Abstract.html
13. Kompa, B., Snoek, J., & Beam, A. L. (2021). Second opinion needed: Communicating uncertainty in medical machine learning. npj Digital Medicine, 4, Article 4. https://doi.org/10.1038/s41746-020-00367-3
14. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), Advances in Neural Information Processing Systems (Vol. 30). Curran Associates. https://papers.nips.cc/paper_files/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html
15. 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, Article 17816.
16. https://doi.org/10.1038/s41598-017-17876-z
17. 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 H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), Advances in Neural Information Processing Systems (Vol. 32). Curran Associates.
18. https://proceedings.neurips.cc/paper_files/paper/2019/hash/
8558cb408c1d76621371888657d2eb1d-Abstract.html
19. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. https://www.jmlr.org/papers/v12/pedregosa11a.html
20. Tibshirani, R. J., Barber, R. F., Candès, E. J., & Ramdas, A. (2019). Conformal prediction under covariate shift. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), Advances in Neural Information Processing Systems (Vol. 32). Curran Associates.

    https://proceedings.neurips.cc/paper/2019/hash/
    8fb21ee7a2207526da55a679f0332de2-Abstract.html
21. Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. BMC Medicine, 17, Article 230. https://doi.org/10.1186/s12916-019-1466-7
22. Vasey, B., Nagendran, M., Campbell, B., Clifton, D. A., Collins, G. S., Denaxas, S., Denniston, A. K., Faes, L., Geerts, B., Ibrahim, M., Liu, X., Mateen, B. A., Mathur, P., McCradden, M. D., Morgan, L., Ordish, J., Rogers, C., Saria, S., Ting, D. S. W., … McCulloch, P. (2022). Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nature Medicine, 28(5), 924–933. https://doi.org/10.1038/s41591-022-01772-9
23. Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. Medical Decision Making, 26(6), 565–574. https://doi.org/10.1177/0272989X06295361
24. Wolberg, W., Mangasarian, O., Street, N., & Street, W. (1993). Breast Cancer Wisconsin (Diagnostic) [Data set]. UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B
25. Wong, A., Otles, E., Donnelly, J. P., Krumm, A., McCullough, J., DeTroyer-Cooley, O., Pestrue, J., Phillips, M., Konye, J., Penoza, C., Ghous, M., & Singh, K. (2021). External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Internal Medicine, 181(8), 1065–1070. https://doi.org/10.1001/jamainternmed.2021.2626
26. Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLOS Medicine, 15(11), Article e1002683. https://doi.org/10.1371/journal.pmed.1002683