Vol. VII, No. I (2022) Pages: 46 – 57 ISSN (Print): 2788-4961 ISSN (Online): 2788-418X

Citation: Manik, M. M. T. G. (2022). Calibration and Decision Thresholds for Machine Learning Classification of Breast Cytology. *Global Immunological & Infectious Diseases Review, VI*(I), 46-57. https://doi.org/10.31703/giidr.2022(VII-I).06

Cite Us

Mia Md. Tofayel Gonee Manik [a]

Calibration and Decision Thresholds for Machine Learning Classification of Breast Cytology

Abstract *This study evaluated breast cytology classifiers using the Wisconsin Diagnostic Breast Cancer dataset (569 fine needle aspirate records, 30 nuclear features) within a five-fold nested cross-validation pipeline. Penalized logistic regression, radial basis support vector machine (SVM), and random forest were assessed for discrimination, probability calibration, decision-curve net benefit, and robustness to predictor noise. Logistic regression achieved the best performance (AUROC: 0.9953, Brier score: 0.0184), followed closely by SVM (AUROC: 0.9946, Brier score: 0.0198) and random forest (AUROC: 0.9877, Brier score: 0.0310). A 95% nominal sensitivity rule applied to logistic regression correctly flagged 203 of 212 malignant records while misclassifying 3 of 357 benign cases. Noise increased Brier scores across models with minimal AUROC impact (<0.0015). These findings emphasize reporting explicit threshold-based error counts alongside rank-based metrics in small-sample cancer AI studies.*

Key Words: Breast Cancer, Fine-Needle Aspiration, Probability Calibration, Machine Learning, Decision Curves, Diagnostic Prediction.

Introduction

Evaluation of a diagnostic prediction model occurs at various stages from sample to a decision. The first is whether there is a tendency for a higher score to be linked to a more malignant lesion. Next is if the score is probabilistically interpretable. A third relates to what proportion of malignant cases get overlooked under a policy whereby review is triggered e.g. when it is above a threshold. A model may perform well on parts (a) but poorly on parts (c). It is important to make this distinction when proposing a classifier to aid decisions for already sampled breast lesions, in which the downstream implications of providing a definitive diagnosis vary depending on the local workup approach and the quality of the sample. Cancer statistics on the global burden gives a perspective on the burden of breast cancer, but the burden, in itself does not prove that a specific classifier will be useful in care (Sung et al., 2021).

Derivation of measurements from images provides a link between Cytology and Tabular data science. A small set of summaries may represent aspects of nuclear size, shape and texture, available after acquisition and processing of an image. Work combining a quantitative image analysis with a machine learning technique has been carried out for the fine needle aspiration (FNA) of the breast and a Wisconsin diagnostic series was developed (Wolberg et al., 1995). This allows the study to be given a specific target – the classification of the given label for an aspirated mass using features that can be extracted from the aspirate. Prediction point is after measurement by sampling and computation. Performance in that setting should be interpreted in terms of the source series and the measurements that it contains.

Breast cancer AI touches on questions of varying and rather complicated data creations processes. For instance, when an AI system was evaluated for use in international evaluation of a mammography system, it was evaluated in a screening setting, which varies from the cytology setting investigated here (McKinney et al., 2020). An

[a] Department of Biochemistry & Molecular Biology, University of Dhaka, Dhaka, Bangladesh
Email: miatofayelgonee@gmail.com

aspirate cohort involves model scores that are dependent on the clinical decisions that resulted in sampling and on the availability of analyzable images. Many who present for population screening would not be in a screening population. Within hospital settings, the model can be built only on selected masses and may be faced with a different distribution of lesions if referral, imaging, or technical criteria for adequate samples changes.

One reason for the usefulness of a public benchmark is that some things can be counted on to remain constant. You should be able to use the same data, the same population labels, and same assessment folds with all model families. This makes it possible to make a targeted comparison of the way models convert same feature matrix to rankings and probabilities. The use of machine learning in oncology has been reviewed elsewhere and shows a variety of diagnostic and prognostic targets, and input types (Kourou et al., 2015). The diversity is lensed by a clear description of the prediction point and unit of observation, which facilitates interpretation of the results across the diversity. In this case the unit is a published diagnostic record, with malignancy regarded as a binary label and regression and response to therapy are not assessed.

An input with quantitative summaries would provide particularly compelling evidence for the addition of a conventional regression baseline. If the feature representation already discriminates the classes well, a logistic model with shrinkage could model much of the predictable variation. One systematic review of clinical prediction models concluded that, in conditions examined, no overall performance advantage of machine learning over logistic regression was observed in comparisons (Christodoulou et al., 2019). That discovery is enough to encourage a just beginning, and not a statement of any one way that must prevail in cytology. Nonlinear approaches may still be helpful with the ranking, or provide an operating point with fewer round the clock reviews. Those benefits need to be reflected in the metrics and an explanation for that policy needs to be included.

It is possible for an attractive AUROC to occur in conjunction with disagreements concerning decisions. AUROC is the sum of the ordering over the positive and negative examples at every possible cutoff. A model which is successful in ordering most random pairs may produce no activity for lesions near the boundary at the chosen threshold for practice. Average precision is a complimentary view of the malignant class by focusing on precision at recall levels. That baseline is related to prevalence and values in the Wisconsin series represent the makeup of its cases. Because the number of positives and negatives are not always equal, the relationship between ROC and precision–recall curves is shown to assist interpretation.

The questions of this study are threefold: What are the similarities and differences between held-out ranking errors and calibrated probability errors of a linear model, a radial-kernel classifier and a tree ensemble? If the same training-side rule is used to indicate that at most 5% of the training records scored out-of-fold are malignant, what is the result in all three models? Discrimination and error in probability prediction as a function of perturbations of held-out features with unchanged fitted system. Actual numbers of false negative and false positive are also reported with the rates for the principal operating point comparison. In the case of robustness, the perturbation size is expressed in terms of standard deviation units and equivalent to a computational stress test. When viewed in isolation, each of these analyses would provide only a snapshot of behavior that ultimately was not accurate.

Materials and methods

Data source and prediction target

We used the Wisconsin Diagnostic Breast Cancer dataset distributed through the UCI Machine Learning Repository and loaded the bundled copy in scikit-learn. The source documents image features computed from digitized fine-needle aspirates of breast masses. Feature engineering begins with the identification and measurement of nuclei in the cytology image, before the tabular data reach the classifier (Street et al., 1993). The matrix has 569 rows and 30 real-valued predictors. Among the released records, 212 are labeled malignant and 357 benign. The positive class was explicitly recoded to malignant because the software loader uses malignant as its first target category. This prevents sensitivity and precision from being calculated for the opposite label.

The 30 columns are arranged as three summaries of ten characteristics. The families cover radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. For each characteristic, the file provides a mean, a standard error, and a statistic labeled “worst.” In the dataset's feature description, the latter summary represents the mean of the three largest values. Several columns are expected to carry related information, particularly the size measures.

We kept all predictors so that the comparison would characterize how each classifier handles that representation. We did not infer biological mechanisms from feature association or interpret radius, perimeter, and area as independent sources of diagnostic evidence.

The feature matrix has no missing entries requiring imputation. The fitted algorithms did not receive case identifiers, report text, demographic fields, or downstream treatment information. The public distribution does not provide the clinical covariates, acquisition dates, independent site labels, or specimen-level quality indicators needed for the proposed diagnostics of temporal drift, subgroup performance, or unresolved aspirates. The results therefore estimate within-series classification under random record splits. The use of published, publicly distributed data did not involve contacting patients or collecting new specimens. The analysis relies on the released diagnosis labels as given and does not independently adjudicate their reference standard.

Table 1. Public benchmark and assessment design

Characteristic	Value
Released records	569
Malignant / benign	212 / 357
Inputs	30 nuclear morphology features
Outcome	Released binary diagnosis label
Outer evaluation	Five stratified folds
Inner fitting	Four stratified folds within each outer training set

Analysis design and data separation

We used one five-fold stratified outer split, shuffled with seed 27, so that each record received a held-out prediction from each model. Stratification kept the malignant fraction similar across folds. Each outer training partition used four shuffled stratified inner folds, seeded with 100 plus the zero-based outer fold index. The outer test features had no role in fitting the scaling, choosing hyperparameters, fitting the calibration model, or setting the training-side operating threshold. The division between tuning and evaluation follows the rationale for nested validation when cross-validation informs model choice (Varma & Simon, 2006). Predictions were collected at the original row positions to allow paired comparison.

For each family, we selected the grid configuration with the largest mean inner-fold AUROC. Ties were resolved by the grid search's deterministic candidate order. We then generated four-fold cross-fitted scores in the outer training partition using the winning configuration and those same inner splits. The configurations therefore depend indirectly on the labels of the inner held-out rows through grid selection, even though the fitted estimator producing a row's cross-fitted score did not directly train on that row. This can make the calibration and threshold training scores more favorable than scores from a fully independent calibration sample. The outer assessment retains all such steps inside training and measures their combined effect on fresh rows from the source dataset.

The training data used to fit the final outer model were larger than those used to generate its inner cross-fitted calibration scores. Applying the sigmoid learned from inner scores to the final model's held-out scores assumes that their numerical scales remain sufficiently similar. For a radial-kernel model, the margins may change with the support vectors available after refitting. For a forest, the training sample can change both the trees and the shape of the score distribution. These differences are part of the algorithm under evaluation. An alternate design would reserve a distinct calibration partition, reduce model-training data but eliminating reliance on calibration scores from smaller cross-fitted models.

Model families and hyperparameter search

Standardization of the feature vector was applied and L2 was used in logistic regression. At every fit the training mean and standard deviations were subtracted and divided by the standardizer. The inverse regularization strength C was set to 0.01, 0.1, 1 and 10. Shrinkage is reinforced by a smaller C. This model provides a linear decision surface in normalized feature space. In cases where information is shared in measurements, the L2 penalty is relevant, otherwise the coefficients can change quite rapidly with changes in the sample. We assumed it to be a predictive baseline, not calculating odds ratios for interpretation of the results. Individual coefficients are hard to interpret due to collinearity and post-fitting sigmoid calibration.

The support vector machine used a radial-basis kernel, standardized inputs, and no built-in probability estimates. The grid combined C values of 0.1, 1, and 10 with gamma values of "scale," 0.01, and 0.1. With the scale setting, the software derives gamma from the feature count and variance seen by the fitted classifier. The margin is a signed score, and its distance from zero is not a probability. We therefore passed the decision function to a separate sigmoid fit. Kernel classifiers can represent curved boundaries in a feature space where proximity depends on the relative scaling of individual variables (Cortes & Vapnik, 1995).

The random forest used 200 trees and seed 31, with the minimum number of records in a terminal leaf set to 1, 3, or 8 and the number of candidate predictors at a split set to the square root of the predictor count or half of the predictors. Forests combine trees fit with randomized features and bootstrap sampling to reduce the variance of individual trees (Breiman, 2001). Changes in minimum leaf size control how local the fitted partitions can become. No standardization was required for the tree estimator. All families used their ordinary, unweighted training objective. Avoiding class weighting also kept the training mixture consistent with the sampling fraction used to fit the sigmoid.

Probability mapping and threshold selection

Every classifier received the same form of score calibration. For logistic regression and the forest, the base positive-class probability was clipped to the interval from 0.000001 to 0.999999 and converted to log odds. The SVM margin was used directly. A one-predictor logistic regression with an intercept and C equal to one million fitted a sigmoid to the training-side cross-fitted scores. The large C makes the L2 penalty weak. This procedure allows the model to adjust both the location and spread of its scores. Sigmoid mapping is a compact choice for the size of the sample and is consistent with previous work comparing methods for producing usable probabilities from supervised classifiers (Niculescu-Mizil & Caruana, 2005).

The ordering of fold inputs is maintained by a sigmoid with positive slope in a fold. When you span the folds, the order changes because two cases might have been evaluated by different fitted estimators. The pooled AUROC thus reflects the (in)ability of the combined probabilities to distinguish between cases, with repeated application of the complete training procedure for each assessment partition. It's not even the mean of the five AUROCs in individual folds. We cite the fold ranges here and show this distinction. Likewise, the pooled reliability graph characterizes the set of probabilities provided by the different models trained by following the same procedure without evaluating a single final deployment model.
In the operating point analysis we arranged the training scores of the mapped malicious records from lowest to highest. We set the cutoff at the largest index of a score in the list that is within the desired proportion (was M the number of malignant training records, we used the score at index M minus the ceil(0.95 times M) there may be more ties than that). We then used the corresponding outer test fold with the threshold that had been frozen. This is an empirical training target and does not constitute a statistical sensitive guarantee on future cases and/or a lower confidence bound. There is a specific cutoff for each fold/classifier. The resulting counts are added together over folds at assessment.

Performance and uncertainty

The measures used were area under the receiver operating characteristic curve, average precision and the Brier score. AUROC is the probability that a randomly selected malignant record is assigned a higher score that a randomly selected benign record (half credit if equated). Average precision also provides a weighting for precision by the increment in recall. The precision–recall plot horizontal axis refers to the malignant fraction in the data set (Saito & Rehmsmeier, 2015). Consider, for example, a set of outcome indicators with a mean of 10% (corresponding to going 1-for-10), and a set of probabilities predicted by A being 5%. Poor performance in the first set (which has a high expected probability of success) is penalized more by the Brier score than is poor performance in the second (with a lower expected probability of success). Combining them all three represents the multi-faceted aspect of predictive performance (Steyerberg et al., 2010).

A stratified bootstrap with 1,500 replicates and a seed of 42 was accomplished to determine this sampling uncertainty for descriptive sampling. Indices only sampled 357 times (benign) and 212 times (malignant) and then replaced. The results of all my predictions and outcomes occurred at the same index. Intervals of performance were calculated for each model based on 2.5 and 97.5 percentiles, and intervals of each model difference from logistic regression based on these percentiles. These intervals are conditioned on the prediction and folds

that have been generated. The classifiers were not refitted for each replicate, meaning that the intervals do not take into account the full extent of variability in model development dependence and resulting from common training data. They do not constitute formal assessments of clinical equivalence, but are statements to aid cautious comparisons at this benchmark.

Eight quantile bins were used for the calibration plots of each model. Within a bin the mean probability and the observed malignant fraction were computed. Binning can flush out errors within a bin, particularly in the central range of probability, even if the general pic of a small sample is stabilized. The quantile cutoffs will vary according to the model. So that adjacent points can represent different records for curves. The Brier score provides an additional description that is complementary to the plot, but a low Brier score does not imply good separation, per se, but good agreement between probability and frequency. If your intent with a probability is to guide action or discussion then consider calibration (Van Calster et al. 2019).

Decision curves and robustness experiment

We calculated estimated apparent usefulness of flagging based on the out-of-fold probabilities; net benefit was calculated for thresholds ranging from 0.025 to 0.65. Net benefit is the fraction of correct flags (th true positive) over the number of records minus the fraction of false flags (th false positive) over the number of records, times t divided by one minus t. At a weight of 0.20, four false positive flags cancel out the one true flag. Every record or zero records were referenced using the reference strategies. Decision curves (Vickers & Elkin, 2006) have the ability to convert specified tradeoffs into a common scale. In this case they state to you a population where they were sampled, and hypothetical additional review. These probability thresholds to replicate the calculation for this site are different from the fold specific probability thresholds used for training sensitivity.

To emphasize the measurement inputs we added Gaussian noise to each of the 30 feature columns held out at two levels. A random normal sample from the seed 81 was multiplied by the outer-training population standard deviation for the 2 selected features for each test row, and added to the original measurement, with 0.10 or 0.25 as the multiple. For each row and feature, the same perturbation was used for all classifiers, and the larger amplitude utilized the same draws. Scaling of each feature ensures that the noise is proportional to the spread of each feature, without relying on test distributions for setting the magnitude of the noise. The fitted pipelines, sigmoid calibrators remained frozen, and true labels remained the same. AUROC, average precision and post-prediction Brier score (after perturbation) have been reported.

The Gaussian experiment probes the stability of an existing numerical representation. Independent perturbations can generate a combination of radius, area, and perimeter that would be unlikely after actual image processing. They also leave lesion prevalence and the process that determines which patients are sampled unchanged. An acquisition change could instead affect many related measurements together, preferentially affect poor-quality samples, or change the fraction of cases for which features can be derived. We treated the experimental amplitudes as two points on a sensitivity analysis. Estimating an expected clinical degradation would require measured repeatability or data from devices, readers, and sample preparations seen in practice.

Reproducibility and reporting

All preprocessing, tuning, fitting, prediction, metric calculation, and graphics used the same Python workflow. The estimators, cross-validation utilities, and benchmark loader were provided by scikit-learn (Pedregosa et al., 2011). Defaults were used outside the documented settings. Seeds are reported for steps that use randomization, and the paired design preserves correspondence between models at every stage of assessment. We used the TRIPOD reporting statement to inform descriptions of participants, predictors, endpoints, validation, and performance (Collins et al., 2015). We used PROBAST to guide discussion of how the available participants, predictors, outcomes, and analysis limit interpretation (Wolff et al., 2019). A formal independent risk-of-bias assessment was outside this secondary analysis.

Results

Cohort and held-out discrimination

The number of records on each outer test fold was 113 or 114 with 42 or 43 malignant records. A total of 212 malignant and 357 benign records were kept in all the folds, which indicates a malignant fraction of 37.3%. One prediction was given to every record per classifier. The ROC curves were close to the upper left corner and, as seen in figure 1, precision was greater than the prevalence line throughout much of

the recall range. The performance of penalized logistic regression was maximized in the comparison using the pooled AUROC and average precision. The result of the radial-basis model deviated by less than on-thousandth for all the metrics. Across individual folds, the AUROC was 0.9902 to 0.9997 for logistic regression, 0.9895 to 0.9993 for the SVM, and 0.9745 to 1.0000 for the forest.

Table 2. Held-out performance and conditional 95% bootstrap intervals

Classifier	AUROC	Average precision	Brier score
Logistic	0.9953 (0.9893–0.9990)	0.9944 (0.9885–0.9985)	0.0184 (0.0114–0.0267)
RBF SVM	0.9946 (0.9876–0.9988)	0.9936 (0.9873–0.9981)	0.0198 (0.0124–0.0291)
Forest	0.9877 (0.9761–0.9963)	0.9866 (0.9774–0.9946)	0.0310 (0.0219–0.0402)

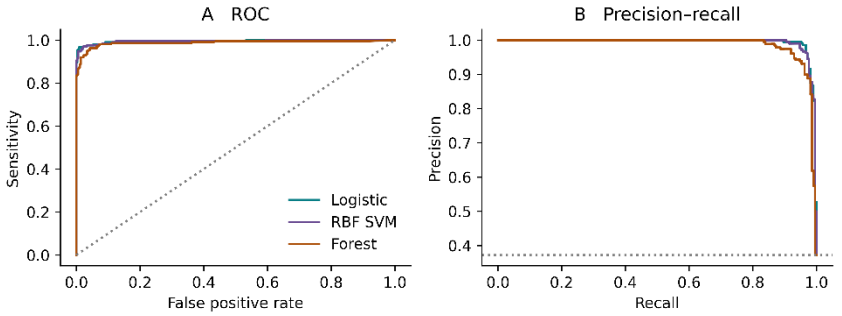


Figure 1. Held-out ROC and precision–recall curves. The dotted lines mark random ranking in the ROC panel and the sample malignant fraction in the precision recall panel.

Subtracting logistic regression from the SVM gave an AUROC difference of −0.0007, with a conditional 95% bootstrap interval from −0.0021 to 0.0004. The analogous forest difference was −0.0076, with an interval from −0.0176 to −0.0013. The SVM's average precision differed from logistic regression by −0.0008, and the forest's differed by −0.0078. These values describe performance on the same cases. Observing a score of 1.0000 for the forest in one-fold was compatible with a lower pooled score because the other folds included overlapping benign and malignant scores. The forest's minimum fold AUROC also occurred in the first fold, where all three models had their lowest or near-lowest ranking performance.

Probability error and threshold behavior

Logistic regression had the lowest Brier score at 0.0184, followed by the SVM at 0.0198 and the forest at 0.0310. For comparison, the constant-probability Brier score computed at the observed sample prevalence is approximately 0.234. The SVM minus logistic Brier difference was 0.0014, with a conditional interval from −0.0026 to 0.0059. The forest minus logistic difference was 0.0125, with an interval from 0.0058 to 0.0191. These comparisons measure the prediction systems after sigmoid mapping. The calibration and decision-curve panels in Figure 2 show their performance across probability levels and decision preferences.

In the case of logistic regression, empirical sensitivity marked 203 out of 212 malignant records, for SVM, 200 were marked and for forest 199 records were marked. The same benign sets were misclassified by the classifiers as 3, 3, and 11 records respectively. That is, logistic regression were missing 9 records of the malignant class and the SVM and forest were missing 12 and 13 records of the malignant class, respectively. The specificity of logistic regression and SVM was 99.2%, and the forest was 96.9%. On records held out for evaluation, the accuracy of two of three classifiers was below the nominal 95% sensitivity target. These differences are due to the combination of the fitted model and the model acquired threshold.

Table 3. Held-out results using the nominal 95% training sensitivity rule

Classifier	True positive	False positive	False negative	True negative	Sensitivity	Specificity
Logistic	203	3	9	354	95.8%	99.2%
RBF SVM	200	3	12	354	94.3%	99.2%
Forest	199	11	13	346	93.9%	96.9%

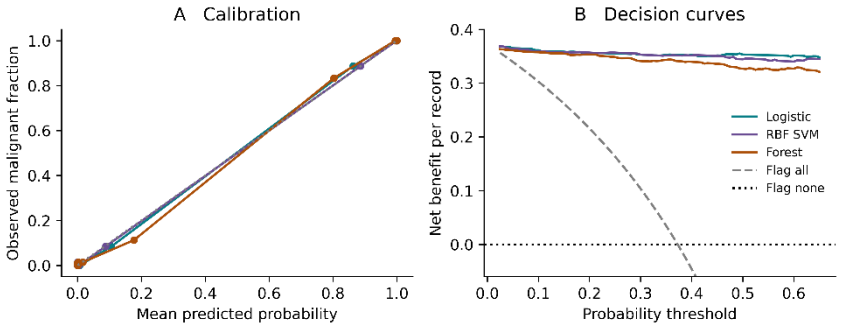


Figure 2. Held-out calibration and decision curves. Calibration uses eight quantile bins per classifier. Net benefit uses hypothetical additional review as the action.

Separate decision-curve analysis showed that at probability value 0.20, net benefit was 0.3563 per record for logistic regression, 0.3568 for the SVM and 0.3533 for the forest. The Strategy baseline cut-off for flagging ALL was around 0.2157 and the baseline cut-off for flagging NONE was 0. At 0.50, model net benefit was 0.3533, 0.3445, and 0.3269 in the same order. There were slight differences among the cutoffs, with SVM and logistic regression showing almost neck-in-neck results at 0.20 and 0.30. These scores describe the hypothetical tradeoffs in the sample. The ranges for the fold-specific cutoffs were 0.440 to 0.684 for logistic regression, 0.584 to 0.743 for the SVM, and 0.396 to 0.546 for the forest.

Performance under measurement perturbation

At noise amplitude 0.25, the Brier score increased from 0.0184 to 0.0216 for logistic regression, from 0.0198 to 0.0228 for the SVM, and from 0.0310 to 0.0363 for the forest. At the same amplitude, AUROC changed by approximately −0.0007, −0.0006, and +0.0014, respectively. The slight AUROC gain for the forest occurred with a worsening Brier score and average precision below its unperturbed value. At the smaller amplitude, the SVM's Brier score decreased slightly. Those small improvements can occur when random displacements move some difficult cases into regions where the learned mapping gives them more appropriate scores. A single perturbation realization cannot distinguish this explanation from an improvement that would recur across draws.

Table 4. Held-out performance under controlled measurement perturbation

Classifier	Noise in SD units	AUROC	Average precision	Brier score
Logistic	0	0.9953	0.9944	0.0184
Logistic	0.10	0.9953	0.9942	0.0188
Logistic	0.25	0.9946	0.9931	0.0216
RBF SVM	0	0.9946	0.9936	0.0198
RBF SVM	0.10	0.9946	0.9936	0.0195
RBF SVM	0.25	0.9940	0.9925	0.0228
Forest	0	0.9877	0.9866	0.0310
Forest	0.10	0.9893	0.9871	0.0317
Forest	0.25	0.9891	0.9863	0.0363

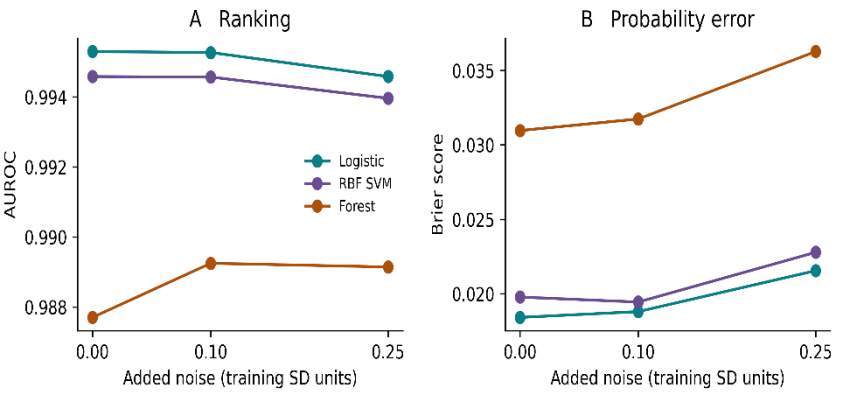


Figure 3. Held-out discrimination and probability error under Gaussian feature perturbation. Noise amplitude is expressed in outer-training feature standard deviations.

No model dropped below 0.9877 AUROC under the specified amplitudes. In contrast, all three had larger probability errors at 0.25 than at baseline. The gap between the behavior of AUROC and Brier score is observable in Figure 3. It reflects the different response of an ordering statistic and a squared probability error to changes in individual risk assignments. Because the same random draws were shared across classifiers, the point comparisons do not arise from one classifier receiving a different perturbation sample. The results are conditional on the chosen inputs, amplitudes, and random draw, with no empirical estimate of how often these perturbations would occur in cytology practice.

Discussion

What high discrimination can and cannot resolve

The closest comparison is with the rankings that are nearly saturated and the experience of an individual lesion at a given level. This feature representation is capable of distinguishing many malignant records from benign records which leaves relatively fewer pairs to be reordered by flexible classifier. Any upturn in gain, that is limited to cases that have scores well beyond or well below a review-cutoff, will likely have minimal impact on the number of cases identified by this policy. On the other hand, swapping a few challenging cases near a cutoff around the ROC line may have considerable operational significance but contribute little to the change in AUROC. This inequality justifies putting a comparison of counts of outcomes next to a discrimination table. The difference also maintains the interpretation at the level of the application; no one is subjected to a mean AUROC and the diagnostic triage occurs at the rule/decision boundary.

Good penalized regression scores are possible as the score reflects existing data of morphology. Some sensed numerical features can clearly discriminate the labels released across most of the samples, where size and contour are irregular features. A forest can make use of non-linear relationships established using local splits, and a radial kernel can capture smoothly curved differences in an input space specified using these measures. The extent to which a model can be used to represent a more complex function, however, is determined by its architecture; whether or not it can use this flexibility effectively, depends on the availability of information in close proximity of the diagnostic boundaries. The experiment is conducted for the specified grids for a specific set of features. It does not indicate how to prioritize between families of algorithms, across all cytology datasets, or between tuning schemes.

The uncertainty intervals helps to put small differences into perspective. Paired resampling involves asking the question of what the score difference would be if different proportions of the cases of the observed types, but the same fitted predictions, were present in each class. That option dictates some variation that an unpaired comparison would add. Still does not know how a new aspiration cohort, a different case mix or the sequence of model fits would change the result. A few influential cases can be done away with or repeat themselves in a resample, in a small benchmark. Differences should hence be consulted with their conditional intervals and with the quantitative operating-point results.

Consequences for a triage workflow

When a 95% sensitivity rule is stated, it is a nominal statement indicating the intent to favour the detection of malignant lesions; but the value of the detection threshold that satisfies the intent is estimated from limited data. There are relatively few records in the lower tail of the malignant score distribution, and if one or more unusual lesions are added to or removed, that will noticeably move a quantile based cut-off. The score mapping for some cases very close to the threshold may be sensitive after refitting the model using more data. A fold-specific cutoff can also map (either misMAP or correctly) the same rank target to different probability values. The mechanisms can also be used to understand why it is necessary to report a simple tuning rule together with its sensitivity and specificity. These below-threshold cases are clinically significant regardless of the small proportion of the number of records.

Workload measure needs a separate calculation for malignant and benign cases. But if a model identifies a 'malignant record', more research may be indicated, provided that this does not have the power to override what has been determined as a normal model assessment. An innocent flag might need to be further investigated but not affect the patient with an invasive procedure. The number of false positives, as determined from the empirical data thus represents a potential review burden for the assumed policy. It doesn't actually provide any gauge about costs, delays, or downstream intervention. However, if the "score" were to queue the already reviewed slides, the practical burden would be different. These differences should be indicated beforehand and taken into account in an external study before the application of net benefit.

Collectively, that of net benefit has a condensed shape of interpretation. The higher the cutoff, the more likely you will miss a true malignant "flag" and the less likely you will do a potentially unnecessary "review" as opposed to a lower cutoff. This trade-off can be summarised on a decision curve over a spectrum of preferences. This overlapping of the area where model curves are near each other does not imply the clinical interchangeability. Consequences for the patient are events, which are not somewhere in the formula, and the seemingly good events on the neighbouring thresholds can be based on almost the same set of records. Thresholds were chosen to show the variations in the curves in the plots. However, a hospital would have to specify acceptable harms, expected number of reviews, and the action that would be taken as a result of a flag before they chose a relevant range.

Calibration and transport

Even if the primary desired outcome of implementation is to rank, probability calibration is important. The clinician may read the probability or use the probability in a rule based upon the relative cost of missed and unnecessary workup. Generally this means that the scores of an overconfident model can remain high while the few malignant cases are only assigned very small probabilities. In a clinical study those cases should be given a separate error review as reassurance based on small, predicted probabilities may have practical reality. Such an investigation would not be possible with the source benchmark without the addition of image-level and clinical annotations. An external evaluation of this informative outcome would follow-up on the reasons why the prediction was incorrect, such as image quality, under-represented lesions in development, ambiguous lesion outcome labels, or a threshold selected for a different workflow.

This sigmoid will allow us to change the average position and the spread of a score, but is only available in one functional form. It will not rectify arbitrary underpredictions or overpredictions within different biological or technical groups. Local deviations can be accommodated with flexible calibration methods, with a sample this size the training data close to the decision threshold can be sparse. Then there could be some more flex in the realized curve, which might react to noise in the actual sample. There is also a limitation on resolution of the bin based plot. When there is close visual agreement with the diagonal, it cannot be taken as evidence that the probabilities are reliable at every clinically relevant cutoff. If a follow-on assessment is conducted, an estimate of the region of calibration must be provided, along with how many observations are used to arrive at the estimate, where a support decision will be different based on the region of calibration.

The other question is concerning changes in prevalence. If more than one-third of the masses sampled in the area are malignant, a particular probability scale can be used for a model trained and calibrated in this area. Under a lower prevalence structure, class conditional scores would be the same in usual but posterior probabilities and positive predictive value (PPV) would tend to be lower. Thus, the proportion of the various tumor presentations to the classes and the class-conditional scores may not

remain constant, as different benign lesion types and their distribution of presentations may be selected for referral. However it is a straightforward prevalence shift that can be addressed through recalibration of intercept. It does not account for alterations in the relationship between the predictors and malignancy. Although direct validation remains the means by which to establish whether the purported score interpretation has been maintained, no empirical evidence has been generated to support the process, and it should be determined by the specific context and implementation of a GMAW process.

Perturbation results cover a smaller challenge than transport. Artificial noise is added to see if the algorithm is stable to a row shifted in measured feature space. A few relocations of cases may result in less appropriate probabilities even if we don't see a change in the overall rank. This is a practical argument, that both in the numerical error, and in the discrimination, a robustness check should be made. A forensic study of the effects of measurement would require re-preparation or re-imaging the same specimen independently annotated and with a case log of specimens failing measurement. This data would assist in credible joint perturbations not in the insertion of independent perturbations for attributes which would normally tend to move together in a cluster measured in the physical sense.

External study design and application boundaries

The potential of an external study is to involve consecutively sampled breast masses in multiple locations and establish a time of prediction prior to the actual output of model predictions for readers. It would indicate the unit of analysis if a patient had several specimens or if the aspirate resulted in more than one slide. It might be able to hold entire patients and, when possible, sites for assessment. Separate sites will minimize the likelihood of regularities in devices or workflows existing during both development and testing. The scores of a time-separated cohort might also indicate if there are any changes in the specimen preparation that influence scoring. These design decisions are driven by the case that random row partitions will not be able to address, when it comes to the sources of dependence and distribution change of the current data set.

Outcome ascertainment would have to define the reference standard, timeframe for ascertainment, and should describe how to handle the discrepant and indeterminate cases. Limiting the cases assessed to those explicitly labeled may yield a more "pure" comparison and may exclude what are the most challenging cases to assess with a decision aid. A validation protocol should include a report on number of images that are unable to be measured, number of images that are inadequate aspirates, and number of requested confirmatory procedures. If the classifier restricts its application to technically adequate specimens, the displayed workflow should show that rule and the impact on coverage. Performance of the subset of scored patients may be confused for performance among all those who were started on evaluation.

Fairness analysis also is limited with the current table of predictors. The absence of the attributes, which would be considered relevant in health effects, like age, site, technically relevant acquisition details make it impossible to estimate the subgroup sensitivity and the magnitude of the effect of missing predictors measurements on the health effects of one group compared with another. If some groups are overrepresented in the set of difficult-to-classify records, their performance may be masked in the overall performance. An independent study having such characteristics might explain both the accuracy of the subgroups estimates and “fullness” of the records prior to and subsequent to imposing the model eligibility criterion. A global total alone would be insufficient to judge the reliability of each subgroup and sufficient responses would be required for adequate assessment in the relevant strata.

Clinical evidence would need to be more than just a retro-comparing of the model output alone. Frozen predictions could initially be tested in background mode without revealed to clinical staff, in a observational study to see the errors that can be seen in routine setting. A second reader study might give a reasonable approximation of how much the score changes when it is known while reading and/or compare the score with the reference standard in terms of agreement and time and/or discrepancies. Lastly, an impact assessment could be used to examine whether there are delays after following-up or diagnosing, and that these delays lead to an impact. Each of these stages are uniquely focused on answering a distinct question about the tool, its use by people and impact on care. The current findings only provide an internal classification benchmark and a reproducible method for reporting a number of dimensions of that benchmark.

Strengths and limitations

Experimentally, the same outer comparisons were performed for each classifier, and the same record-

level predictions were tested, without a fit on the corresponding test fold. These forecasts cut down on "cheating" on evaluation of calibration, or policy, errors by using in-sample probabilities as evidence. The sensitivity analysis contains a fixed Gaussian sensitivity as well to make the stress test repeatable. Concurrently, a known dataset would restrict the novelty of the prediction of the outcome, and could even be that the selection of modelling was guided by general knowledge of the benchmark. The contribution is an empirical calibration, operating point and perturbation response comparison through a common validation design. It is not to be applied as a new in situ cytology classifier for clinical use.

Some further conditions are placed on the analysis. When an outer split is performed, no variation among the alternative partitions is found. Confidence intervals do not capture the full extent of the uncertainty associated with the model development process but rather the conditional sample of predictions. There was no independent time or institution available and the moisture content of the source cohort may not necessarily reflect the moisture content of images at the time of extraction. The comparison of probabilities is specific to a particular pipeline-for each one, it cannot compare the probability to an un-calibrated version since the sigmoid is part of the calibrated pipeline. There was one random realization and two amplitudes in the noise study and no bootstrap bands were drawn for perturbed results. The empirical results are descriptions to the procedures and source matrix implemented.

Just a "model 1" report will lead you to have more confidence than this design allows for. A more compelling application of the comparison is to analyze what factors lead to error in the workflow, and which facts of the evaluating process need outside data. Questions can focus on the lowest-scored malignant lesion(s) or the most likely-to-be-flagged benign lesion(s) for numerical performance. These questions can only be actionable when those investigators are allowed to examine the original specimens, compare who is right, what, and consider how they can document the apparent disagreements. Using the public benchmark, the analytical steps can be readily reproduced. It has a restricted clinical context which creates a solid limit for explanations on individual errors.

Conclusion

In addition, the common five-fold validation process enabled comparison of three BC classifiers evaluated on the same set of records while changing their behaviour from the ranking to the probability estimation, and a high sensitivity review policy. Adding FN and FP counts revealed outcomes, behind just one ranking value. In the noise experiment, it was demonstrated how varying the feature values to a controlled extent might lead to a change in the probability error and discrimination, under frozen models. All clinical interpretations require validation with an explicit reference standard and consideration of the sampling situation and interpretation of review. The results are most useful as baseline and reproducible evidence for the disseminated breast cytology benchmark, and to guide measurement of the subsequent external study.

Data and code availability

The Wisconsin Diagnostic Breast Cancer dataset is available through the UCI Machine Learning Repository at https://archive.ics.uci.edu/dataset/17/breast-cancer-wisconsin-diagnostic. Appendix A provides the analysis and plotting script. The scikit-learn version used for the reported results was 1.8.0.

References

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology, 110*, 12–22.

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *Annals of Internal Medicine, 162*(1), 55–63.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*, 273–297.

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 233–240). ACM.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal, 13*, 8–17.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. C., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., ... Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature, 577*(7788), 89–94.

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning* (pp. 625–632). ACM.

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.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE, 10*(3), e0118432.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., & Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology, 21*(1), 128–138.

Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993). Nuclear feature extraction for breast tumor diagnosis. In R. S. Acharya & D. B. Goldgof (Eds.), *Biomedical Image Processing and Biomedical Visualization* (Vol. 1905, pp. 861–870). SPIE.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians, 71*(3), 209–249.

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.

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics, 7*, Article 91.

Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making, 26*(6), 565–574.

Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1995). Image analysis and machine learning applied to breast cancer diagnosis and prognosis. *Analytical and Quantitative Cytology and Histology, 17*(2), 77–87.

Wolff, R. F., Moons, K. G. M., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., & Mallett, S. (2019). PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine, 170*(1), 51–58.