ISSN 2349-4506
Impact Factor: 3.799

Global Journal of Engineering Science and Research Management

LEAKAGE AND EVALUATION BIAS IN MACHINE LEARNING FOR BREAST CYTOLOGY DIAGNOSIS: A REVIEW

Deokar Pooja Balasaheb*[1], Arpit Solanki*[2]
[1]Dr. A.P.J. Abdul Kalam University, Indore, M.P., India
[2]Dr. A.P.J. Abdul Kalam University, Indore, M.P., India

KEYWORDS: Breast Cancer Diagnosis; Fine-Needle Aspiration; Wisconsin Diagnostic Breast Cancer Data; Machine Learning; Feature Selection; Stacked Generalization; Cross-Validation; Selection Bias.

ABSTRACT
Breast cancer remains the most frequently diagnosed cancer among women, and fine-needle aspiration (FNA) cytology offers a minimally invasive route to separating benign from malignant masses. Over several decades, machine-learning classifiers trained on nuclear features taken from digitised cytology images have reported very high accuracy on public benchmarks, above all the Wisconsin Diagnostic Breast Cancer (WDBC) data set. This paper reviews the literature on benchmark data, classification algorithms, feature selection, ensemble learning and evaluation methodology, and compares the surveyed studies in terms of scope, strengths and limitations. The comparison shows that headline accuracies are hard to compare, because studies differ in how data are split, in whether pre-processing and feature selection were fitted inside or outside the validation loop, and in the metrics that are reported. Further problems are the reliance on a single split, little attention to false negatives and decision thresholds, a weak separation between the effect of feature selection and the effect of the classifier, and uncertainty about whether stacked ensembles help on small clinical data sets. Based on these problems, the paper outlines a future direction in which every fitted step is learned only from training data and models are judged by repeated resampling and operating-point analysis.

INTRODUCTION
Breast cancer is the cancer most often diagnosed in women across the world; global estimates for 2020 recorded about 2.3 million new cases and about 685,000 deaths [1]. The chance of a good outcome depends on how early the disease is detected and how dependably a lesion is labelled as benign or malignant. FNA is among the least invasive ways of sampling a breast mass: cells are withdrawn with a thin needle and a pathologist inspects the nuclei under the microscope. Because this inspection is a human judgement, it carries some observer-to-observer variation, and a computer-aided second opinion that is accurate, compact and well validated could make it more consistent.

Machine learning has been applied to cancer prediction and prognosis for several decades [2], [3]. When nuclear measurements are extracted from digitised FNA images, as in the WDBC data set [4], [5], even simple classifiers separate the two classes very well, and published accuracies are often close to ceiling. This apparent success hides a methodological problem. Papers differ in how the data are partitioned, in whether scaling, feature selection or parameter tuning was carried out before or after the partition, and in the performance measures they quote [6], [7]. As a result, two reported accuracies are seldom directly comparable, and a small numerical lead may reflect the evaluation protocol rather than a better model.

The purpose of this paper is to review the literature around this task and to identify, from the evidence, the problems that remain open. The review is thematic and selective and rests on 32 sources covering benchmark data, algorithms, feature selection, ensembles and evaluation methodology. Section II surveys this literature, Section III compares the surveyed work and formulates the problems, and Section IV concludes and, within the future scope, points to a possible solution.

LITERATURE SURVEY
2.1 Benchmark data and early diagnostic models
The benchmark on which most studies rely goes back to Street, Wolberg and Mangasarian, who described a semi-automatic procedure in which nuclei in digitised FNA slides are outlined and ten shape and texture measurements are computed for each nucleus [4]. Summarising these measurements over all nuclei of a sample produces the WDBC data set, which is distributed through the UCI repository [5]. In a companion line of work, Mangasarian

ISSN 2349-4506
Impact Factor: 3.799



Global Journal of Engineering Science and Research Management

et al. used linear programming to construct separating planes for both diagnosis and prognosis and reported accuracy in the high-90 % range [8]. Together, these papers established the data and the expectation that a handful of nuclear features can discriminate the two classes.

2.2 Comparative studies and classification algorithms

Comparative studies place several classifiers on the same data. Asri et al. evaluated support-vector machines, a decision tree, naive Bayes and k-nearest neighbours on the Wisconsin data and found the SVM to be the most accurate, at about 97 % [9]. A study of this type shows which algorithm scored highest on one split of one data set, but it says little about variability across resamples or about the cost of missed malignancies. Broader reviews of machine learning in oncology note that neural networks, SVMs and decision trees are in wide use, while the strength of validation differs markedly from paper to paper [2], [3].

The algorithms that recur in such comparisons are few. Logistic regression is a fast, well-calibrated linear model of the log-odds [10]; an SVM with a radial-basis kernel finds the widest-margin surface and can bend it smoothly around the data [11]; k-nearest neighbours votes among the closest training samples and is sensitive to feature scaling [12]; naive Bayes and decision trees are cheap and interpretable but strained by correlated features [13]; random forests average many decorrelated trees to cut variance [14]; and gradient boosting fits trees sequentially to residual errors, with a risk of over-fitting on small samples [15].

2.3 Feature selection

Guyon and Elisseeff divide selection methods into filters, wrappers and embedded approaches and stress that selection must be treated as part of learning when performance is estimated [16]. Recursive feature elimination with an SVM has worked well for gene selection in cancer classification [17], and mutual information offers a model-free filter score that captures non-linear dependence, with nearest-neighbour estimators available for continuous variables [18]. Ambroise and McLachlan showed that choosing features on the complete data set and cross-validating afterwards gives strongly optimistic error estimates [6], and Cawley and Talbot reached a similar conclusion for hyper-parameter tuning: whenever model selection touches the evaluation data, the estimate is biased [7].

2.4 Ensemble learning and stacking

Dietterich explained why ensembles can outperform single models: they reduce statistical variance, computational error caused by local optima, and representational limitations [19]. Zhou gives a book-length account of bagging, boosting and combination strategies [20]. Wolpert introduced stacked generalization, in which the predictions of first-level models, made on data they were not trained on, become the inputs of a second-level model [21]. Whether stacking pays off depends on the diversity and individual quality of its members; with a small data set and base models that are already strong, the gain can be small.

2.5 Evaluation methodology and metrics

Kohavi recommended stratified ten-fold

cross-validation for model selection on moderately sized data [22]. Dietterich cautioned that standard significance tests on resampled accuracies can mislead because training sets overlap [23], while Demšar recommended non-parametric tests such as the Wilcoxon signed-rank test for comparing classifiers [24]. Accuracy alone conceals class-specific errors; the Matthews correlation coefficient gives a balanced single value [25], precision–recall analysis is more informative when positives are rare [26], the ROC curve and its area remain the standard threshold-independent summary [27], [28], and the Brier score judges the predicted probabilities themselves [29]. For unequal class sizes, remedies include re-weighting, threshold shifting and synthetic over-sampling such as SMOTE [30]; applied before the split, however, over-sampling puts synthetic copies of one sample into both training and test sets, which is the same kind of leakage discussed above.

COMPARATIVE ANALYSIS

The surveyed studies address different parts of the diagnostic pipeline and were not evaluated under identical conditions, so their results should not be compared number by number. The comparison below therefore focuses on what each contribution offers and on the limitation that matters for this review.

ISSN 2349-4506
Impact Factor: 3.799
Global Journal of Engineering Science and Research Management

3.1 Comparison of the reviewed studies

Table 3.1 summarises the main studies and the limitation of each that is relevant to this review.

Study	Primary information / approach	Main contribution	Important consideration
Street et al. [4]; Dua and Graff [5]	Nuclear features from FNA images	Feature-extraction method; WDBC benchmark	Data paper; no comparison of modern learners or evaluation protocols
Mangasarian et al. [8]	Linear-programming separating planes	Diagnosis and prognosis with accuracy in the high-90 % range	Single model family; limited analysis of error types
Asri et al. [9]	Classifier comparison	SVM, decision tree, naive Bayes and k-NN compared; SVM best at about 97 %	Single split; no variability estimate or false-negative analysis
Cruz and Wishart [2]; Kourou et al. [3]	Reviews of machine learning in oncology	Overview of neural networks, SVMs and trees	Note uneven validation but do not quantify its effect
Guyon and Elisseeff [16]; Ambroise and McLachlan [6]	Feature selection	Taxonomy of methods; evidence of selection bias	General guidance; not applied side by side to a stacked ensemble
Wolpert [21]; Zhou [20]	Ensemble learning	Stacked generalization; ensemble theory	Gain depends on the data; needs an empirical check on small clinical sets

Table 3.1: Comparative Analysis of Existing Studies

3.2 Problems identified

The comparison points to several open problems. The first concerns validation and leakage. Reviews of the field note uneven validation [2], [3], and a typical comparative study reports one split of one data set [9], so a difference of a fraction of a percentage point cannot be interpreted. Selecting features on the full data set before cross-validation [6] and tuning hyper-parameters on evaluation data [7] both inflate the reported score, and the principle that selection belongs inside the learning procedure [16] is widely accepted but not always followed in applied papers.

The second concerns what is measured and compared. Accuracy hides class-specific errors [25], yet in diagnosis a missed malignancy generally has more serious consequences than an unnecessary follow-up test; the comparative work reviewed here [8], [9] gives limited information on error types, and threshold-dependent behaviour is rarely examined although ROC and precision–recall tools are available [26]–[28]. Many WDBC descriptors are also near-copies of one another, such as the radius, perimeter and area of the same nucleus [4], and the surveyed studies seldom separate how much performance comes from the feature subset and how much from the classifier [16], [18]. Finally, standard significance tests on resampled accuracies can mislead [23], and the recommended non-parametric alternatives [24] are not routinely applied.

The third concerns generality. Ensemble theory promises gains [19], [20], but their size depends on the diversity and quality of the members [21], and empirical checks on small clinical sets are scarce. Most work also rests on one historical collection of pre-extracted features [4], [5], so results inherit any error in nucleus segmentation and may not transfer to other populations or imaging protocols.

CONCLUSION AND FUTURE SCOPE

This paper reviewed the literature on machine-learning diagnosis of breast masses from FNA-derived nuclear features. Benchmark data [4], [5], a wide range of classifiers [8]–[15], feature-selection methods [16]–[18], ensembles [19]–[21] and evaluation tools [22]–[30] are all well established, and reported accuracies are high. The comparative analysis, however, shows that the evidence is less solid than the headline numbers suggest: validation

is uneven, leakage through pre-processing, selection and tuning is common, accuracy is emphasised over false negatives, and the value of stacking on small data sets is unclear. Progress on this task therefore depends less on adding another classifier than on a transparent, leak-free and statistically careful evaluation.

Future scope. Future work can address the problems identified above in several ways. A leakage-free stacked-ensemble pipeline, in which standardisation, correlation filtering, mutual-information selection and the stack are fitted only on training folds and evaluated by repeated cross-validation and operating-point analysis, is one possible solution to the validation, leakage and stacking problems. Hyper-parameter tuning inside nested cross-validation [7], using systematic strategies such as random search [31], would remove the assumption of fixed settings. Probability calibration and cost-sensitive operating points chosen with clinicians, supported by the Brier score [29] and ROC-based analysis [27], [28], would address the false-negative concern, while Shapley-value explanations [32] would show clinicians why a sample was flagged. Non-parametric statistical comparison [24], applied with the caution advised in [23], should become standard practice. Finally, testing on independent multi-centre data and extending the work to learning directly from images, including nucleus segmentation [4], would widen the data basis.

REFERENCES

1. H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, “Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” CA: A Cancer Journal for Clinicians, vol. 71, no. 3, pp. 209–249, 2021.
2. J. A. Cruz and D. S. Wishart, “Applications of machine learning in cancer prediction and prognosis,” Cancer Informatics, vol. 2, pp. 59–77, 2006.
3. K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, “Machine learning applications in cancer prognosis and prediction,” Computational and Structural Biotechnology Journal, vol. 13, pp. 8–17, 2015.
4. W. N. Street, W. H. Wolberg, and O. L. Mangasarian, “Nuclear feature extraction for breast tumor diagnosis,” in Proc. IS&T/SPIE Int. Symp. Electronic Imaging: Science and Technology, vol. 1905, 1993, pp. 861–870.
5. D. Dua and C. Graff, “UCI Machine Learning Repository,” School of Information and Computer Sciences, University of California, Irvine, 2019. [Online]. Available: http://archive.ics.uci.edu/ml
6. C. Ambroise and G. J. McLachlan, “Selection bias in gene extraction on the basis of microarray gene-expression data,” Proc. Natl. Acad. Sci. USA, vol. 99, no. 10, pp. 6562–6566, 2002.
7. G. C. Cawley and N. L. C. Talbot, “On over-fitting in model selection and subsequent selection bias in performance evaluation,” Journal of Machine Learning Research, vol. 11, pp. 2079–2107, 2010.
8. O. L. Mangasarian, W. N. Street, and W. H. Wolberg, “Breast cancer diagnosis and prognosis via linear programming,” Operations Research, vol. 43, no. 4, pp. 570–577, 1995.
9. H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, “Using machine learning algorithms for breast cancer risk prediction and diagnosis,” Procedia Computer Science, vol. 83, pp. 1064–1069, 2016.
10. D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed. Hoboken, NJ, USA: Wiley, 2013.
11. C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
12. T. Cover and P. Hart, “Nearest neighbor pattern classification,” IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, 1967.
13. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY, USA: Springer, 2009.
14. L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
15. J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
16. I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” Journal of Machine Learning Research, vol. 3, pp. 1157–1182, 2003.
17. I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines,” Machine Learning, vol. 46, no. 1–3, pp. 389–422, 2002.
18. A. Kraskov, H. Stögbauer, and P. Grassberger, “Estimating mutual information,” Physical Review E, vol. 69, no. 6, Art. no. 066138, 2004.

ISSN 2349-4506
Impact Factor: 3.799



19. T. G. Dietterich, “Ensemble methods in machine learning,” in Multiple Classifier Systems, Lecture Notes in Computer Science, vol. 1857. Berlin, Germany: Springer, 2000, pp. 1–15.
20. Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms. Boca Raton, FL, USA: CRC Press, 2012.
21. D. H. Wolpert, “Stacked generalization,” Neural Networks, vol. 5, no. 2, pp. 241–259, 1992.
22. R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in Proc. 14th Int. Joint Conf. Artificial Intelligence (IJCAI), 1995, pp. 1137–1143.
23. T. G. Dietterich, “Approximate statistical tests for comparing supervised classification learning algorithms,” Neural Computation, vol. 10, no. 7, pp. 1895–1923, 1998.
24. J. Demšar, “Statistical comparisons of classifiers over multiple data sets,” Journal of Machine Learning Research, vol. 7, pp. 1–30, 2006.
25. D. Chicco and G. Jurman, “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation,” BMC Genomics, vol. 21, Art. no. 6, 2020.
26. T. Saito and M. Rehmsmeier, “The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets,” PLoS ONE, vol. 10, no. 3, Art. no. e0118432, 2015.
27. T. Fawcett, “An introduction to ROC analysis,” Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.
28. J. A. Hanley and B. J. McNeil, “The meaning and use of the area under a receiver operating characteristic (ROC) curve,” Radiology, vol. 143, no. 1, pp. 29–36, 1982.
29. G. W. Brier, “Verification of forecasts expressed in terms of probability,” Monthly Weather Review, vol. 78, no. 1, pp. 1–3, 1950.
30. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
31. J. Bergstra and Y. Bengio, “Random search for hyper-parameter optimization,” Journal of Machine Learning Research, vol. 13, pp. 281–305, 2012.
32. S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Advances in Neural Information Processing Systems 30 (NIPS 2017), 2017, pp. 4765–4774.