Evaluation of Performance Metrics for Histopathological Image Classifier Optimization

Nishant Zachariah-*IEEE Student Member*, Sonal Kothari, Senthil Ramamurthy, Adeboye O. Osunkoya, and May D. Wang, *IEEE Member*

***Abstract*—Clinical decision support systems use image processing and machine learning methods to objectively predict cancer in histopathological images. Integral to the development of machine learning classifiers is the ability to generalize from training data to unseen future data. A classification model's ability to accurately predict class label for new unseen data is measured by performance metrics, which also informs the classifier model selection process. Based on our research, commonly used metrics in literature (such as accuracy, ROC curve) do not accurately reflect the trained model's robustness. To the best of our knowledge, no research has been conducted to quantitatively compare performance metrics in the context of cancer prediction in histopathological images. In this paper, we evaluate various performance metrics and show that the Lift metric has the highest correlation between internal and external validation sets of a nested cross validation pipeline (R^2 = 0.57). Thus, we demonstrate that the Lift metric best generalizes classifier performance among the 23 metrics that were evaluated. Using the lift metric, we develop a classifier with a misclassification rate of 0.25 (4-class classifier) for data that the model was not trained on (external validation).**

I. Introduction

Machine learning techniques have made significant inroads into the arena of diagnostic medicine [1-3] with the availability of large public data sets such as The Cancer Genome Atlas [4]. Clinical decision support systems (CDSS) aim to leverage advances in machine learning to assist pathologists in making fast and objective decisions. Previous work has examined this problem in detail with special emphasis on the amelioration of batch effects that can lower prediction accuracy [5-7]. Thus far, work in the field of automated histopathological image classification has focused on the development of machine learning classifiers using a performance metric. To the best of our knowledge, no research has been conducted on systematic evaluation of the performance metrics that are used to select optimal classification models [2, 3]. Work in the literature has sought to innovate on two main fronts: classifier design and feature definition and extraction [2, 8, 9]. In doing so, the undergirding use of the performance metric which informs both classifier selection and feature usefulness has largely been unexplored.

In the machine learning community, work on metrics have focused on the efficient implementation of distance measures that form the core of classifiers such as k-nearest neighbors [10, 11]. These papers have sought to develop fast online distance measures to enable rapid computation of data classes. While this is important for large scale problems, these studies do not address the underlying performance metrics which play an integral role in the model selection of models formed by the distance measured developed in [10, 11]. In this regard, it can be argued that performance metrics supersede all machine learning paradigms short of classifier training. The metrics of choice for biologically applied machine learning work has been that of accuracy, error and ROC curve analysis [12, 13] with little focus on whether these metrics are indeed appropriate for robust model selection.

In this paper, we examine the ramifications that metrics have in the choice of a machine learning classifier. We systematically study 23 metrics in a multiclass classification problem and identify the metric that has the highest correlation between internal and external validation for renal histopathological images thereby making it ideally suited for classifier identification. Finally, using the optimal metric, we develop a robust parametric classifier by leveraging nonlinear dimensionality reduction on the extracted nonlinear features taken from the raw data. The robust classifier demonstrates very low external misclassification rate despite imbalances in overall distribution of each class.

II. Methods

A. Data

The data used in this study are digital micrographs of renal cell carcinomas acquired at the Georgia Institute of Technology, Atlanta, GA. Each image has dimension 2048x2048 pixels and has an overall Fuhrman grade annotation between 1 and 4. Each image was acquired using a Zeiss Axio Imager z2 microscope at 40x magnification. The images were acquired in 2012 from 18 subjects for a total of 160 labeled images. The images were originally acquired and stored using the portable network graphics format. The distribution of grades within the labeled data set was not balanced with 12.25% grade 1 samples, 39.38% grade 2 samples, 31.25% grade 3 samples and 16.98%

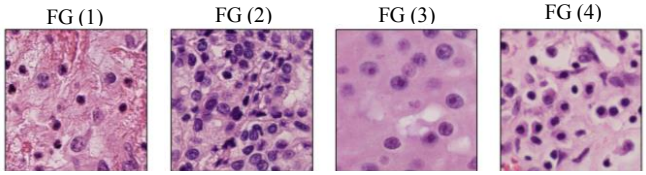

Figure 1: Fuhrman grade for renal cell carcinomas. FG: Fuhrman grade. From left to right: samples of progressive Fuhrman grades are shown.

Nishant Zachariah and Senthil Ramamurthy are with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332. (email: nishant@gatech.edu)

Sonal Kothari, PhD is with the Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA.

Adeboye O. Osunkoya is with the Department of Pathology, Emory University School of Medicine, Atlanta, GA 30322 USA.

May D. Wang, PhD is with the Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA (corresponding author: phone: 404-385-2954; fax: 404-385-0383; e-mail: maywang@bme.gatech.edu).

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

grade 4 samples. A cursory examination of Figure 1 reveals that the grade classification is indeed a challenging task. The differences between grades are based on nuclear, nucleoli and chromatin morphology along with image texture [14].

B. Feature Extraction

We extracted 572 features from each image. These features consisted of wavelet based texture features, color coherence vectors and eccentricity based morphological features. The features were scaled to ensure cross feature consistency and redundant features were eliminated to ensure non-singularity of the feature matrix. The features extracted here and procedure used here are similar to the work performed by Bucheron, L in 2008 [15], Kothari et al [5-7]. Table 1 enumerates a list of nonlinear features that were extracted. In particular, wavelet coefficients are intrinsically sparse thereby making them ideally suited for the detection of texture based features in images. A comprehensive list of features that were extracted can be found at http://users.ece.gatech.edu/~nzachariah3/EMBC14/fea_list.rtf.

Table 1: List of nonlinear features

Type	Representative Classes
Texture (442 features)	Wavelet energy, Fourier energy, Gray Level Co-occurrence, Moment Invariance, Gray Level Run Length, Contrast, Correlation, Energy, Homogeneity.
Color (91 features)	Color Coherence Vectors
Morphology (39 Features)	Area, Major Axis Length, Minor Axis Length, Orientation, Eccentricity, Convex Area, Euler Number, Filled Area, Entropy, Convex Deficiency, Solidity, Extent, Border

C. Feature Reduction and Classifier Optimization

Post feature extraction, the optimal parameters for both dimensionality reduction and classifier performance was determined. A nested cross validation schematic was implemented in order to empirically determine the optimal parameter choice. The parameters that were optimized include choice of PCA / kernel PCA for dimensionality reduction, level of dimensionality reduction (2-40 dimensions, step size : 1), variance of the Gaussian kPCA kernel (σ^2 = 2-10, step size: 1) and type of discriminant analysis (liner, nonlinear, pseudo-linear and pseudo-quadratic). The dimensionality reduction step ensures model generalizability given the small sample size of our datatset. Kernel PCA was implemented using the open source implementation found in the Matlab dimensionality toolbox [16]. Parameter optimization was conducted within the internal loop of the nested cross validation for each metric. Once the optimal parameter choice was determined, a model was developed in the external loop for validation. This pipeline for classifier selection is depicted below in Figure 2.

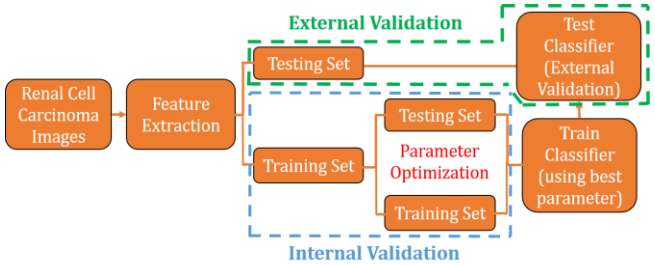

Figure 2: Schematic of classifier selection and nonlinear dimensionality reduction. The schematic show the nested cross validation implementation where the inner loop is used for parameter optimization while the outer loop is used for classifier validation

At each data splitting stage, a 4 fold 20 iteration cross validation was implemented. Thus, the testing data was never trained on to ensure generalizability. Given the double split, 40 untrained samples were used for the test set. This was done repeatedly in an iterative fashion with no contamination between training and test set. The structure of our problem is inherently a multiclass problem as such a one versus all (OVA) strategy was used for efficient classification.

The discriminant classifier (with a linear kernel) is a parametric classifier that can be written as follows

$$f(x) = \arg\min_k (x-\hat{\mu}_k)^T \hat{\Sigma}^{-1}(x-\hat{\mu}_k) - 2\log\hat{\pi}_k$$

$$\hat{\pi}_k = \frac{|\{i: y_i = k\}|}{n};\ \hat{\mu}_k = \frac{1}{|\{i: y_i = k\}|}\sum_{i:y_i=k} x_i;$$

$$\hat{\Sigma}_{linear} = \frac{1}{n}\sum_{k=0}^{K-1}\sum_{i:y_i=k}(x_i-\hat{\mu}_k)(x_i-\hat{\mu}_k)^T;$$

$$\hat{\Sigma}_{nonlinear,k} = \frac{1}{|\{i: y_i = k\}|}\sum_{i:y_i=k}(x_i-\hat{\mu}_k)(x_i-\hat{\mu}_k)^T;$$

$$\hat{\Sigma}_{pseudolinear} = \left(\hat{\Sigma}_{linear}^T\hat{\Sigma}_{linear}\right)^{-1}\hat{\Sigma}_{linear}^T;\ \text{and}$$

$$\hat{\Sigma}_{nonlinear,k} = \left(\hat{\Sigma}_{nonlinear,k}^T\hat{\Sigma}_{nonlinear,k}\right)^{-1}\hat{\Sigma}_{nonlinear,k}^T$$

Where $\hat{\pi}$ represents the apriori probability of each class, n represents the total number of samples, $\hat{\Sigma}$ are different covariance matrices, and $\hat{\mu}_k$ is the mean within each class. The notation $|\ |$ represents the cardinality of the set in question.

D. Classifier Performance Metrics

The optimal classifier is a function of the metric that is used to define the optimality. Since the grade classification problem is multiclass, we used the standard extension of each metric for multiclass problem as defined in [17]. The term 'macro' denotes averaging of a metric for each class across all classes while the term 'micro' denotes the pooled estimate of all classes for each numerator and denominator estimate of a metric. For clarity this form of macro/ micro is explicitly defined for the sensitivity metric below.

Sensitivity: Measures the ability to a classifier to positively identify a specific class among all existing classes.

$$Sensitivity = \frac{True\ Positive}{Total\ Number\ of\ Positives}$$

$$Sensitivity_{macro} = \frac{1}{4}\sum_{i=1}^{4}\frac{True\ Positive\ Class\ i}{Total\ Number\ of\ Positives\ Class\ i}$$

$$Sensitivity_{micro} = \frac{\sum_{i=1}^{4} True\ Positive\ Class\ i}{\sum_{i=1}^{4} Total\ Number\ of\ Positives\ Class\ i}$$

Specificity: Measures the ability of a classifier to accurately reject a class among all existing classes.

$$Specificity = \frac{True\ Negatives}{Total\ Number\ of\ Negatives}$$

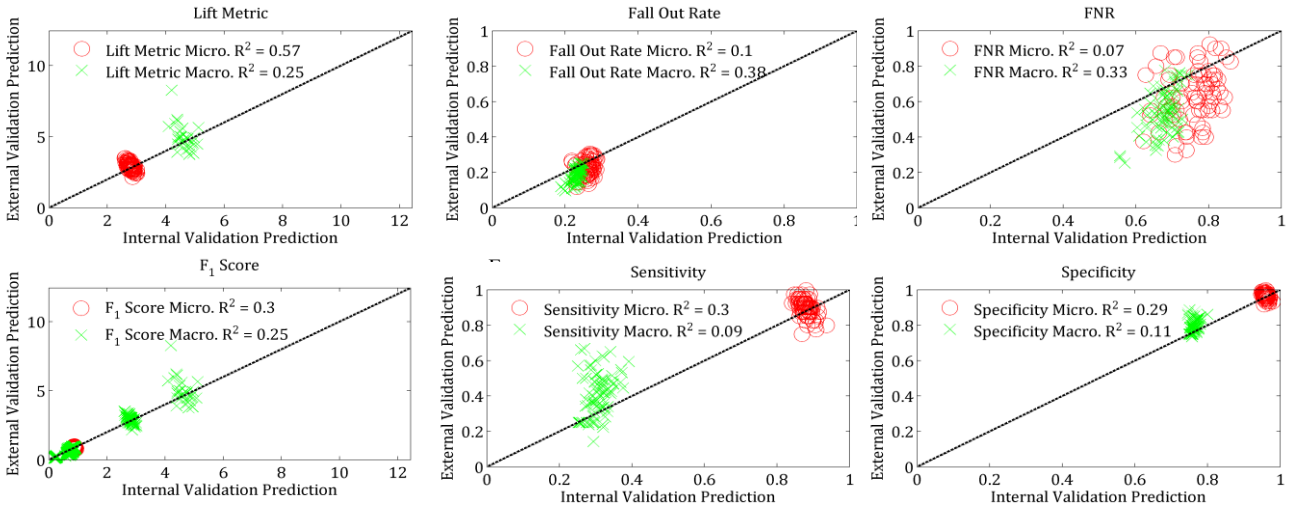

Figure 3: Correlation between internal and external validation for the top six metrics. The lift metric (A) has the highest internal and external validation correlation. Fall out rate (B), False Negative Rate (C), F_1 Score (D), Sensitivity (E) and Specificity (F) metrics display moderate ability to track external validation as a function of internal validation.

Positive Predictive Value (PPV): Measures the positive class identification proclivity of a classifier.

$$PPV = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Negative Predictive Value (NPV): Measures the negative class identification proclivity of a classifier.

$$NPV = \frac{True\ Negatives}{True\ Negatives\ +\ False\ Negatives}$$

Fall Out rate: Measures a classifiers rate of false positives

$$Fall\ Out\ Rate = \frac{False\ Positives}{Total\ Number\ of\ Negatives}$$

False Discovery Rate: Measures a classifiers tendency to positively identify a class incorrectly.

$$False\ Discovery\ Rate = 1 - PPV$$

False Negative Rate: Measures a classifiers tendency to reject a class incorrectly.

$$False\ Negative\ Rate = \frac{False\ Negatives}{True\ Positives\ +\ False\ Negatives}$$

F Score: The F_1 score is a weighting of a metrics precision and recall accuracy. F_1 as is defined is a harmonic weighting of two entities (precision and recall). The $F_{0.5}$ score on the other hand places a larger emphasis on precision while F_2 score weights recall higher [18].

$$F_1\ Score = \frac{2PPV \times Sensitivity}{PPV + Sensitivity}$$

$$F_{0.5}\ Score = \frac{1.25PPV \times Sensitivity}{0.25PPV + Sensitivity}$$

$$F_2\ Score = \frac{5PPV \times Sensitivity}{4PPV + Sensitivity}$$

Mathews Correlation Coefficient (MCC): Mathew's correlation coefficient is a metric that has been traditionally defined to be prevalence independent unlike accuracy and ranges from -1 to 1 where 1 is perfect prediction while -1 is worse than random prediction [19].

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Classification Error: Is a direct measure of classification accuracy.

$$Classification\ Error = \frac{|f(x_i) \neq y_i|}{n}$$

Where $f(x_i)$ is the application of a classifier to the feature vector x_i and where y_i is the true label of the feature vector x_i. 'n' is the total number of samples

Lift [20]: The lift metric examines how much better a classifier performs relative to random chance.

$$Lift = \frac{TP \times n}{(TP + FP)(TP + FN)}$$

Where TP = True Positive, FN = false negative, etc, 'n' is the total number of samples. Thus for each of these metrics, a micro and macro version were implemented. All the metrics are bound between 0 and 1 except for MCC as previously stated, F_1, $F_{0.5}$, F_2 and lift where the metrics are only bounded on the lower range by 0. PPV or positive predictive value places and false negative rate force classifiers to have a large number of positive (in a binary case) class while the negative predictive value emphases classifiers that have a bias towards negative (in a binary case) class.

III. Results and Discussion

We evaluate the metrics using the correlation between each metrics internal and external validation results. We present both the micro and macro version of each metric when applicable. It must be noted that the internal validation result was obtained based on the average of the metric for the optimized parameter in the inner loop of the nested cross validation. Table 2 summarizes the correlation results while Figure 3 explicitly plots the 6 top metrics for better visualization. The lift metric (micro) has the highest correlation between internal and external validation. In short, it serves as the best measure of which classifier generalizes to the unseen data (here modeled by external validation). Given the definition of the lift metric, it would stand to reason that it elevates the correlation of internal

Table 2: Correlation between internal and external validation performance of models selected using different metrics. The statistical significance of each correlation value is shown to validate the existence /nonexistence of correlation. The classification error in external validation for the models optimized relative to each metric is also listed. The metric with the highest correlation is highlighted in red.

Metric		R^2 Value	P Value	Error
Classification Error		0.29	2.35e-7	0.03
PPV	Micro	0.11	2.60e-3	0.93
Sensitivity	Micro	0.30	1.57e-7	0.25
	Macro	0.09	6.80e-3	0.90
Specificity	Micro	0.29	2.91e-7	0.25
	Macro	0.11	3.00e-3	0.90
F_1 Score	Micro	0.30	1.49e-7	0.25
	Macro	0.06	0.22	0.63
$F_{0.5}$ Score	Micro	0.29	2.16e-7	0.25
	Macro	0.04	0.27	0.25
F_2 Score	Micro	0.29	2.30e-7	0.25
	Macro	0.03	0.37	0.25
MCC	Micro	0.26	1.59e-6	0.25
	Macro	0.02	0.45	0.25
NPV	Micro	0.26	1.19e-6	0.25
	Macro	0.25	2.81e-6	0.25
Fall Out Rate	Micro	0.10	4.0e-3	0.35
	Macro	0.38	1.05e-9	0.33
False Discovery Rate	Micro	0.11	2.6e-3	0.35
False Negative Rate	Micro	0.07	1.9e-2	0.30
	Macro	0.33	2.39e-8	0.33
Lift	Micro	**0.58**	**3.83e-16**	**0.25**
	Macro	0.26	3.8e-3	0.20

and external validation results for most metrics. The lift (micro) metric, which as seen in Figure 3 is the most robust to model generalization, was original defined to represent how much better a classifier performs relative to the random chance. This form of model quantification generalizes best among all the metric considered.

Since a robust classifier is commonly optimized and validated using misclassification rate, we also list misclassification rates in Table 2 for models optimized using different metrics and the pipeline depicted in Figure 2. When optimized for lift metric, we were able to develop a robust classifier with a low external misclassification rate of 0.25 (in a four-class classification). Classification error is very low when classifier is optimized using classification error but the model and metric is often biased towards higher prevalent class when there is difference in prevalence (such as in our study). Also, as indicated by correlation values models optimized using classification error are less generalizable than models optimized using prevalence aware metrics such as Lift.

For most metrics, optimized models used kPCA with ~35 dimensions and a Gaussian kernel variance ~4. Parameter optimization and model selection took 52.3hrs on a computer running 12 Intel Xeon cores clocked at 2.66GHz. The use of greater data dimensionality (beyond 40 dimensions) and alternate kPCA kernels such as the inhomogeneous polynomial kernel remain viable avenues for future exploration.

IV. Conclusion

In this paper, we have compared 23 different classifier performance metrics and highlighted their importance in final model selection and performance. Our results for a case-study on four-class Fuhrman grading of renal carcinoma images suggests that the lift metric is a robust metric with high correlation (R^2 = 0.58, p-value=3.83e-16) between internal and external validation performance. Using Lift metric, we were able to optimize four-class classifier with misclassification rate of 0.25 in external validation. Different metrics capture different properties of models. Thus, in future work, we will design an empirical weighted estimate of different metrics to create a new data driven metric and test if it would generalize better than any single metric.

Acknowledgment

The authors would like to thank Dr. Todd H. Stokes for his assistance in image data acquisition.

References

[1] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," *Artificial Intelligence in medicine,* vol. 23, pp. 89-109, 2001.

[2] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer informatics,* vol. 2, p. 59, 2006.

[3] P. Bartels, J. Weber, and L. Duckstein, "Machine learning in quantitative histopathology," *Analytical and quantitative cytology and histology/the International Academy of Cytology [and] American Society of Cytology,* vol. 10, pp. 299-306, 1988.

[4] N. C. Institute. (March 17,2014). *The Cancer Genome Atlas*. Available: http://cancergenome.nih.gov/newsevents/multimedialibrary/images

[5] S. Kothari, J. H. Phan, R. A. Moffitt, T. H. Stokes, S. E. Hassberger, Q. Chaudry, *et al.*, "Automatic batch-invariant color segmentation of histological cancer images," in *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, 2011, pp. 657-660.

[6] S. Kothari, J. H. Phan, and M. D. Wang, "Scale normalization of histopathological images for batch invariant cancer diagnostic models," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, 2012, pp. 4406-4409.

[7] S. Kothari, J. H. Phan, A. N. Young, and M. D. Wang, "Histological image feature mining reveals emergent diagnostic properties for renal cancer," in *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on*, 2011, pp. 422-425.

[8] X. Llorà, A. Priya, and R. Bhargava, "Observer-invariant histopathology using genetics-based machine learning," *Natural Computing,* vol. 8, pp. 101-120, 2009.

[9] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological Image Analysis: A Review," *Biomedical Engineering, IEEE Reviews in,* vol. 2, pp. 147-171, 2009.

[10] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 209-216.

[11] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, "Distance metric learning with application to clustering with side-information," *Advances in neural information processing systems,* pp. 521-528, 2003.

[12] M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. Aguiar, *et al.*, "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," *Nature medicine,* vol. 8, pp. 68-74, 2002.

[13] H. F. Jelinek, A. Rocha, T. Carvalho, S. Goldenstein, and J. Wainer, "Machine learning and pattern classification in identification of indigenous retinal pathology," in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, 2011, pp. 5951-5954.

[14] J. Eble, G. Sauter, J. Epstein, and I. Sesterhenn, *Pathology and genetics of tumours of the urinary system and male genital organs*: IARC press Lyon, 2004.

[15] L. E. Boucheron and B. Adviser-Manjunath, *Object-and spatial-level quantitative analysis of multispectral histopathology images for detection and characterization of cancer*: University of California at Santa Barbara, 2008.

[16] D. Cai. (January 25,2014). *Kernel PCA*. Available: http://www.cad.zju.edu.cn/home/dengcai/Data/code/KPCA.m

[17] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management,* vol. 45, pp. 427-437, 2009.

[18] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation," *Journal of Machine Learning Technologies,* vol. 2, pp. 37-63, 2011.

[19] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Structure,* vol. 405, pp. 442-451, 1975.

[20] Carunana, Rich. "Emperical Methods in Machine Learning and Data Mining". Lecture 6: Performance Metrics. Department of Computer Science, Cornell University. 2003