Binary Classifier Evaluation Without Ground Truth

Maksym Fedorchuk*
*National Technical University of Ukraine
Faculty of Biomedical Engineering,
Kyiv, Ukraine
m.fedorchuk-2017@kpi.ua

Bart Lamiroy†
†Université de Lorraine – Loria (UMR 7503)
Campus scientifique – BP 239
54506 Vandœuvre-lès-Nancy Cedex – France
Bart.Lamiroy@loria.fr

***Abstract*—In this paper we study statistically sound ways of comparing classifiers in absence for fully reliable reference data. Based on previously published partial frameworks, we explore a more comprehensive approach to comparing and ranking classifiers that is robust to incomplete, erroneous or missing reference evaluation data.**

On the one hand, the use of a generalized McNemar's test is shown to give reliable confidence measures in the ranking of two classifiers under the assumption of an existing better-than-random reference classifier. We extend its use to cases where its traditional formulation is notoriously unstable. We also provide a computational context that allows it to be used for large amounts of data.

Our classifier evaluation model is generic and applies to any set of binary classifiers. We have more specifically tested and validated it on synthetic and real data coming from document image binarization.

I. Introduction

Many of the research issues in Pattern Recognition are expressed as classification problems. For instance, in Document Image Analysis alone, number of significant contributions to the field consist of more or less elaborate and specialized classifiers [6], [8], [20]. This is true for the larger Pattern Recognition domain. In many cases, there are multiple different classification approaches for a same class of problems, each often developed for specific tasks and data types. It is therefore important to be able to determine the ones most suitable for the problem at hand.

Measuring advances to the state-of-the-art therefore largely depends on the capacity of evaluating the differences in quality between various approaches, and classifiers in particular. The most commonly used approach consists of testing them on known annotated reference data (*Ground Truth* or *Golden Standard*) and then measuring their agreement/disagreement level on these data. Often the result is expressed in terms of *precision* and *recall* or similar measures [12]. The question of how accurate these measures, and how significant conclusions of ranking various classifiers can be, arises when the annotated reference data cannot be fully trusted and may contain debatable interpretations. We have started investigating these issues in previous work [7], [11]. In this paper we are pushing the reasoning to a more extreme point, by considering classifier evaluation in entire absence of reference data.

Everything that follows applies to generic binary classifiers, but we will be using image binarization algorithms for their convenience, low computational cost and broad availability for setting up an experimental protocol. The reader should be aware that all results extend easily to other binary classifiers.

The rest of this paper is organized as follows: Section II outlines the general context of our work and previous efforts related to evaluating classifier performance in absence of reference data or *Ground Truth*; in Section IV we outline the underlying theoretical basis to the *Pseudo-Metrics* and *Reference Method* approaches used in this paper. Extensive experimental validation is reported in Section V, and a final analysis and conclusion can be found in Section VI.

II. Classifier Evaluation with Uncertain Reference Data

A. Origins of Reference Data Uncertainty

In this paper we are assuming unreliable reference data. Although it is common practice to consider the data on which methods are evaluated as untainted and free of any error, it can be quite easily proven that this is only the case on very rare occasions [10]. In the specific context of document image binarization, the problem has also been raised by others [1], [2]. The origins of this general uncertainty on reference data are multiple, and they are not always necessarily to be considered as "errors". In many cases, they can be traced back to genuine ambiguity in the data, or legitimate differences in interpretation that are open to discussion and debate [10] or may be influenced by the application context and subsequent actual use of the classification results, too.

It is beyond the scope of this paper to fully cover the reasons and origins of reference data uncertainty. The core of our argument is to claim that not taking it into account when evaluating and ranking classifiers, one necessarily introduces biases that may lead to false or incomplete conclusions [10]. This paper explores some tools to provide a statistically sound framework framing this bias, and experimentally establishes their scope of validity.

B. Comparing Binary Classifiers

In general, comparison of classifiers, be they binary or not, is done using reference data. Classifier output is then compared with *Ground Truth* and ranked using traditional metrics such as *F-Measure, PSNR, NRM, Cross correlation, etc.*.

In previous work [7] we have started to develop a probabilistic approach that can assess performance without reference data. The general idea behind this work is to replace the

traditional *Ground Truth* by the overall consensus between classifiers, and by ranking them with respect to their relative individual disagreement with this consensus. These probabilistic metrics have been proven to correlate well with traditional ones that rely on *Ground Truth*.

An other approach consists in pairwise comparison of classifiers [18] rather than establishing an absolute metric-based ranking. This paired testing is based on using a (third) reference classifier as a benchmark. The authors show how the comparison of output of the classifiers relates to the underlying performance. They use a generalized McNemar's test for measuring the statistical significance for expressing the agreement between the tested classifiers and the reference one and measure the probability that one outperforms the other. We use the same test in Section IV-B.

III. Tested Classification Algorithms

The performance evaluation metrics described above make the assumption one tries to evaluate a collection of classifiers addressing the same classification problem. As a consequence, in order to set up a valid experimental protocol, it is important to have a set of those for which the operational conditions that can be easily controlled and reproduced, and that are convenient for large runs of repeated execution within a reasonable amount of time and computational resources. They should also reflect settings that are representative of real problems.

We have chosen two specific binary classification contexts. The first consists in using image binarization algorithms (*cf.* Section III-A). Image binarization is a well studied problem and many algorithms exist. They can be seen as pixel-wise classifiers, classifying each point into foreground or background classes. Their advantage is also to easily provide large quantities of data.

The second is a more peculiar setup. We train a simple perceptron neural network on the MNIST handwritten digit dataset [13] to recognize one single digit, thus obtaining a binary classifier. Using the same network topology, we train perceptron multiple times (each time obtaining a new classifier), changing the data sampling. The result gives a range of classifiers supposedly solving the same problem (identifying the same digit) but having slightly different behavior, given the changed learning sampling.

It is important to note that the scope of this paper is not to evaluate binarization or perceptron back-propagation as such, but to use these contexts to generate binary classification data in a controlled environment, in order to validate our various models for comparing classifiers without ground truth. We could have taken any other collection of binary classifiers.

A. Binarization Algorithms

The general approach for every binarization system is the same: if a pixel (i, j) in the input image has a higher gray level value than a given threshold, then this pixel is labeled as background, otherwise, it is labeled as foreground. Individual binarization approaches differ in how the threshold is computed: global algorithms calculate one for the entire image, while local algorithms may have different threshold values for each pixel of the image, depending on their surrounding region. For our experiments, we have used Otsu's global binarization method [16]; a local version of Otsu's method[1]; Bernsen's method [3]; Niblack's method [15]; Breadly's method [5]; local medium value; a modified version of Gatos' method [8]; Wolf's method [22]; Kittler's method [9]; Sauvola's method [19].

The chosen algorithms have significant reported performance differences. Most of them rank higher than others in at least some specific context or for particular types of images according to the DIBCO evaluation campaigns [17].

All these algorithms depend on operational parameters and their performance is sometimes sensitive to subtle changes. We applied consistent and near-optimal parameters for all experiments, either by applying the recommended published parameters, either experimentally determining parameters. It should be clear to the reader that the scope of this paper concerns *Ground Truth*-less performance metrics, and not binarization. Therefore, whether the choice of parameters is actually optimal or not is of no incidence to the conclusions we will eventually draw form our study on the various metrics we will be developing in Section VI.

B. Tested Machine-Learning Algorithms

Artificial neural network classifiers rely on large amounts of annotated data to "learn" the weights of their synaptic links. We assume most readers are familiar with the basics of neural networks: information flows through a neural network in two ways. The usual *feed-forward* (and actual classification) phase consists of giving it input by activating the neurons of the first layer. This data is propagated through the subsequent layers by modulating their influence with the weight of each synaptic link between neurons. If the weights of the synaptic links are set correctly, the output of the network corresponds to the classification result that is expected for the given input. When the network is learning (being trained) the weights of the links are progressively and iteratively adapted by feeding it patterns of information with known expected outcomes using a *back-propagation* gradient descent optimization.

What may be less known is that convergence to a usable classifier largely depends on the learning data and the order in which it is presented to the network, and that, consequently, training the same network on the same data, but by shuffling their order, may result in significantly different weight distributions (and thus, different networks) while maintaining very similar classification results. Furthermore, in order to avoid over-fitting, the resulting networks never achieve a 100% classification result. The outcome is a series of similar classifiers, with globally equivalent performance, but differing slightly on individual inputs.

Our setup thus provides a large quantity of similar classifiers that occasionally disagree on individual training samples. This

[1]Local Otsu's method and Gatos' method were slightly modified with respect to their published versions [7].

offers a reproducible and controllable framework for simulating the more broader context of multiple classifier evaluation on "uncertain" data.

IV. Statistics and Tests for Comparing Classifiers

Now that we have set up the general context and experimental environment of our work, the following sections will describe how to compare multiple classifiers without knowledge of ground truth. Section IV-A addresses traditional ranking metrics, revisited in the light of unknown reference evaluation data, Section IV-B addresses a statistic confidence measure that two given classifiers can be ranked in a specific order.

A. Comparing Classifiers Using Probabilistic Metrics

Given the legitimate objections to Ground Truth-based evaluation expressed in [1], [2], [10] the idea of using performance metrics that can be used in absence of Ground Truth has been experimented in [7]. The main idea behind the approach is to replace the standard Ground Truth with a consensus metric resulting from the collection of compared methods. In this section we recall the reformulation of traditional metrics in probabilistic terms with respect to this consensus metric. We are using the notations introduced in [12]: the statistical equivalent of GT is an array of expressing the probability $P(\delta_i)$ for each data item δ_i to belong to the class Δ^+. These probabilities can be written:

$$P(\delta_i) = \sum_{k=1..s} \frac{S_k(\delta_i)}{s}$$

Where $S_n(\delta_i) \in \{0, 1\}$ represents the classification result of data item δ_i by classifier S_n, and s the number of classifiers.

1) F-Measure: under the hypothesis of equivalent distribution of all data items δ_i, probabilistic equivalents of Precision and Recall can be defined as

$$Pr(S_k) = \frac{\sum_{1..d} P(\delta_i) S_k(\delta_i)}{\sum_{1..d} S_k(\delta_i)} \quad Rc(S_k) = \frac{\sum_{1..d} P(\delta_i) S_k(\delta_i)}{\sum_{1..d} P(\delta_i)}$$

where d is the total number of elements classified by classifier S_k as belonging to one of classes. $Pr(S_k)$ and $Rc(S_k)$ can be combined into a corresponding probabilistic equivalent of the F-Measure by computing their harmonic mean.

2) Negative Rate Metric: can be expressed in function of the probabilistic equivalents of False Negative, False Positive, True Positive and True Negative values. Resulting in

$$\text{NR}_{FN}(S_k) = 1 - \frac{\sum_{1..d} P(\delta_i) S_k(\delta_i)}{\sum_{1..d} P(\delta_i)}$$

$$\text{NR}_{FP}(S_k) = \frac{\sum_{1..d} (1 - P(\delta_i)) S_k(\delta_i)}{\sum_{1..d} P(\delta_i)}$$

NRM is the average of the negative rate of false positive and false negative values:

$$\text{NRM} = \frac{\text{NR}_{FN} + \text{NR}_{FP}}{2}$$

In contrast to F-Measure and PSNR, the lower the value for this metric, the better the classifier.

3) Normalized Cross Correlation: is the normalized correlation between the probability that the elements δ_i belong to class Δ^+ given the majority voting $P(\delta)$ and the result given by classifier S_k.

$$\text{NCC} = \frac{\sum_{1..d} \left(S_k(\delta_i) - \bar{S_k}\right)(P(\delta_i) - \bar{P_\delta})}{\sqrt{\sum_{1..d} (S_k(\delta_i) - \bar{S_k})^2 \sum_{1..d} (P(\delta_i) - \bar{P_\delta})^2}}$$

The higher this value, the better both arrays correlate with each other.

4) Peak Signal-Noise Rate:

$$\text{PSNR} = -ln\left(\sum_{1..d} \frac{\left(S_k\left(\delta_i\right) - P_\delta(\delta_i)\right)^2}{d}\right)$$

The higher the value of PSNR, the higher the similarity of the two arrays.

We refer the interested reader to [7] for further details.

B. Comparing two Classifiers Using a Reference Method

In the previous section, the probabilistic metrics are used to rank classifiers from best to worst. We show in Section V-B that this ranking comes very close to the one obtained with *Ground Truth*. However, they don't provide a measure of confidence in the ranking that is obtained, nor do they provide a level of reliability.

This can be solved by using a statistical significance test like McNemar's. Use of the McNemar test in classifier performance analysis is not new [4]. However, the general state-of-the-art applies it to traditional configurations, where *Ground Truth* is available. In [18], the authors explore the use of a general McNemar test for expressing a confidence measure in the ranking of two classifiers when *Ground Truth* is either unreliable or unavailable, provided a "decent" reference classifier is available. In this section we apply their work to our context.

Let A and B be two classifiers to be compared, and R a "decent" reference classifier; "decent" meaning that the correct classification rate of $R > 50\%$. Furthermore let $N_{A\bar{B}R}$ (resp. $N_{\bar{A}BR}$) be the number of classification samples where A and R agree with one another, and disagree with B (resp. B and R agree with one another, and disagree with A). Let $N_{AB} = N_{A\bar{B}R} + N_{\bar{A}BR}$.

[18] establishes that the p-value related to the test's significance can be computed with a binomial probability distribution $\mathcal{B}$ of parameters $\sigma = N_{AB}; \mu = 0.5$

$$p = \begin{cases} 2\,\mathcal{B}\left(N_{A\bar{B}R} \leq n \leq N_{AB}\right) & \text{if } N_{A\bar{B}R} > N_{\bar{A}BR} \\ 2\,\mathcal{B}\left(0 \leq n \leq N_{A\bar{B}R}\right) & \text{if } N_{A\bar{B}R} < N_{\bar{A}BR} \\ 1 & \text{if } N_{A\bar{B}R} = N_{\bar{A}BR} \end{cases} \quad (1)$$

If the p-value is smaller than a given rejection threshold, it can be safely assumed that the sign of $N_{A\bar{B}R} - N_{\bar{A}BR}$ reflects the classification quality of A *vs.* B. On the other hand, if p-value is above the rejection threshold, then the comparison is considered non conclusive.

V. Empirical Comparison of Classifiers

In this section we establish that the previous approaches can be used when *Ground Truth* is unavailable or unreliable. In order to achieve this we compare our methods to traditional *Ground Truth*-based evaluations. It is obvious that this is uniquely for benchmarking and experimental validation. Field use of our metrics does not require *Ground Truth*.

A. Using Pseudo-Metrics

The first question we are trying to answer is how good the probabilistic metrics described in Section IV-A are in ranking classifiers in comparison with *Ground Truth* based approaches. We used three different test configurations: one using a set of artificially generated binary images, one using a collection of real images from the Digital Image Binarization Contests (DIBCO 2009–2013 [17]) with binarization algorithms described in Section III-A and one using MNIST data [13] and a set of classifiers built by machine learning algorithms as described in Section III-B.

For each configuration we used the same testing protocol:

1) we rank all tested classifiers using known *Ground Truth* and using standard metrics, giving for each metric μ_i a vector $\mathcal{R}_{GT}(\mu_i) = (r_1, r_2, \ldots r_{10})$ with the rank of each classifier;
2) we rank all tested classifiers without relying on *Ground Truth* by using the probabilistic metrics mentioned in Section IV-A, giving for each metric μ_i a vector $\mathcal{R}_P(\mu_i) = (r_1, r_2, \ldots r_{10})$ with the rank of each classifier;
3) we compute the correlation between the pairs of $(\mathcal{R}_{GT}, \mathcal{R}_P)$ for each metric μ_i.

Metrics with a higher correlation value more reliably reproduce *Ground Truth* based ranking than metrics with lower correlation.

1) Artificial Data: in order to establish baseline observations, we have generated an artificial 1000×1000 black and white image and use this as *Ground Truth*. This image is then randomly modified in order to obtain 10 new artificial images with a controlled level of "errors" compared to the reference image. These are supposed to imitate binary classifiers with different performance levels. In our experiments we used the following artificial error levels: $(0.5\%, 1\%, \ldots 5\%)$ and $(5\%, 10\%, \ldots 50\%)$. The results are reported in Tables I and II. When comparing the raw metric values between the *Ground Truth*-based metrics and their probabilistic counter-part, we obtain the best results with the NRM metric. However, their impact is marginal in the sense that it doesn't influence the ranking (*cf.* Table II).

Metric	Correlation up to 1% error: 0.1%, 0.2%... 1%	Correlation up to 5% error: 0.5%, 1%... 5%	Correlation up to 50% error: 5%, 10%... 50%
PSNR	0.998 ($4\,10^{-4}$)	0.997 ($6\,10^{-4}$)	0. 967 ($2\,10^{-3}$)
F-Measure	0.999 ($5\,10^{-5}$)	**0.999** ($7\,10^{-5}$)	0. 997 ($1\,10^{-3}$)
NCC	**0.999** ($7\,10^{-5}$)	**0.999** ($7\,10^{-5}$)	0. 997 ($5\,10^{-3}$)
NRM	**0.999** ($7\,10^{-5}$)	**0.999** ($7\,10^{-5}$)	**0. 997** ($5\,10^{-4}$)

TABLE I
AVERAGE CORRELATION AND STANDARD DEVIATION OF METRIC VALUES ON ARTIFICIAL IMAGES.

Metric	Correlation up to 1% error: 0.1%, 0.2%... 1%	Correlation up to 5% error: 0.5%, 1%... 5%	Correlation up to 50% error: 5%, 10%... 50%
PSNR	1	1	1
F-Measure	1	1	1
NCC	1	1	1
NRM	1	1	1

TABLE II
AVERAGE CORRELATION OF RANKING ON ARTIFICIAL IMAGES.

2) Real Data: we used the 56 manually ground truthed DIBCO 2009–2013 contest images and applied the binarization algorithms described in Section III-A. The results are reported in Table III. We grouped the results by year and type of image (printed, hand written) as done for the contests. This allows us to compute average correlation and standard deviation. Best results are obtained for PSNR (highest correlation) and F-Measure (lowest standard deviation). NRM is definitely not suited for our purposes.

Metric	Avg. Correlation	Std. Dev
PSNR	**0.856**	0.06
F-Measure	0.845	**0.051**
NCC	0.783	0.234
NRM	0.373	0.163

TABLE III
AVERAGE CORRELATION AND STANDARD DEVIATION OF RANKING ON REAL IMAGES.

3) Neural Network Classifiers: showed highly unstable correlation results. The main reason is that the overall performance of the classifiers is very similar, and differences between them are so small (0.3% – 0.5%) that the pseudo-metrics run into stability issues (very small variations immediately induce large effects on the ranking).

We can make following conclusions after these tests:

1) Pseudo-metrics have high correlation with *Ground-Truth* based metrics for classifier ranking and can therefore be applied in absence of reference data.
2) Classifiers showing only marginal performance differences create instabilities and failure to correctly distinguish them by using pseudo-metrics.
3) We did not report experiments where we used increasing numbers of classifiers from 3 to 10. They confirmed the findings reported in [7]: the more classifiers one uses the more robust and reliable the probabilistic metrics be-

come, and the higher the correlation with *Ground Truth* based metrics (resp. the lower the standard deviation).

B. Using the McNemar Test and a Reference Method

In order to evaluate the second approach, we need to slightly adapt our experimental protocol. Indeed, as described in Section IV-B and in [18] the method based on the generalized McNemar test only allows for comparing 2 classifiers at the time, provided there is a third reference classifier available with a better than 50% classification rate.

1) Experimental Protocol: since the approach only allows comparison between two classifiers we have used the following protocol for all experiments in this section and in Section V-C. For a given reference classifier $\mathcal{R}$ and a set of classifiers $\{S_k\}$, we generate all possible pairs (S_i, S_j) and compare them with using $\mathcal{R}$. When the p-value < 0.05, the global rank value of S_i (resp. S_j) is incremented if $S_j < S_i|_{\mathcal{R}}$ (resp. $S_j < S_i|_{\mathcal{R}}$). In the end, classifiers are ranked from highest to lowest rank values.

2) Testing Reference Classifier Influence: the method requires the reference classifier $\mathcal{R}$ to have better than 50% classification performance.We have evaluated experimentally in how far the quality of this classifiers influences on the overall robustness. We therefore used the same technique as described in Section V-A1 and generated artificial images with controlled error rates for $\mathcal{R}$, and progressively increased the error rate by 1% steps. We ran 100 tests for each value of classification performance.

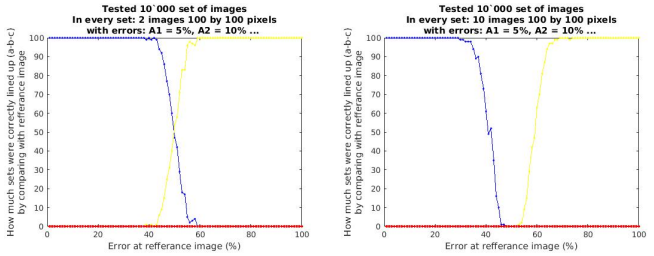

Fig. 1. Influence of reference classifier quality on classifier comparison

When trying to correctly rank 2, 3 and 5 classifiers while varying the quality of $\mathcal{R}$, ranking remained correct up to a classification error rate of 49% for $\mathcal{R}$. The classifiers used for this test have different controlled error rates according to a benchmark initial classifier and were set with steps of 5% (5%, 10%, 15%...). When increasing the number of classifiers, sensitivity to $\mathcal{R}$ increased. Ranking 10 classifiers yielded correct ranking orders more than 50% of the time for classification error rates of $\mathcal{R}$ up to 47%. This is further confirmed in the next experiment.

Fig. 1, shows the results for ranking 2 and 10 classifiers respectively with varying classification errors of $\mathcal{R}$.

3) Quality in Function of Number of Classifiers: contrary to our findings in Section V-A, the more classifiers are used for evaluation, the lower the ranking quality. This is reported in Table IV showing the results on artificially generated images.

# Classfiers	Error betwn Classifiers	Max Error – 90% Correct Ranking	Max Error – 50% Correct Ranking
2	5%	49%	49%
3	5%	49%	49%
10	5%	47%	48%
2	3.8%	47%	49%
3	3.8%	45%	47%
10	3.8%	37%	42%

TABLE IV
RANKING QUALITY IN FUNCTION OF $\mathcal{R}$ ERROR RATE AND NUMBER OF CLASSIFIERS

Experiments were performed 100 times for every percent of error in $\mathcal{R}$.

4) Testing the Sensitivity to Classifier Similarity: our tests also showed that the minimum required quality difference between classifiers that will allow them to be ranked correctly more than 50% of times is 3.8% when comparing 10 classifiers, and 3.6% when ranking 5 classifiers.

Results on a machine learned classifiers were inconclusive. Average ranking correlation with *Ground Truth* ranking was around 0.6, but in more than half of cases the p-value was too large, and ranking had to be rejected. This is due to the fact that the differences between the classification systems were around 0.3% – 0.5%, as already mentioned in Section V-A.

We can make following conclusions after these tests:

1) The less classifiers we compare, the better the results and a higher tolerance in the quality of $\mathcal{R}$ can be accepted.
2) Again, when classifiers have very similar performance levels, ranking becomes unstable and results are unreliable.
3) However, McNemar's rejection test efficiently identifies statistical inconsistent comparisons (and notably reduces the influence of the previous case, when performance levels are too similar).

C. Comparing both Statistical Approachess

Finally, we compared both approaches of classifier ranking. We did not test machine learnt classifiers since the previous evaluations for both the reference method and the probabilistic metrics were inconclusive or insufficient.

In order to reduce possible bias from the way results are measured, we have used two other metrics besides correlation for comparing ranking results: word-edit distance [14] and sequence alignment cost [21]. These two additional metrics were chosen considering that the order of the evaluated classifiers is eventually a more reliable way to make conclusions about the approach than correlation of this order with the one obtained by using *Ground Truth*.

We compared the probabilistic F-Measure with the reference method on real and artificial binary images. The results are reported in Tables V and VI. For the real images, we detailed the results in function of the DIBCO datasets, per year. The same comparison was done on 9000 artificial images in which we used, for every test, a set of 10 artificially-generated images with differences of 5%, 0.5% and 0.1%.

	DIBCO 2013	DIBCO 2012	DIBCO 2011	DIBCO 2009	Total
Avg. Correlation					
F-Measure	0.597	**0.719**	0.70	0.636	**0.664**
Reference	**0.605**	0.461	**0.72**	**0.904**	0.654
Avg. Word-Edit Distance					
F-Measure	6.3	5.6	5.1	5.5	5.6
Reference	**4.6**	**5.4**	**4.7**	**2.9**	**4.5**
Avg. Sequence Alignment Cost					
F-Measure	12.8	11.8	10.9	11.9	11.8
Reference	**8.81**	**11.4**	**9.75**	**6.00**	**9.23**

TABLE V
RELIABILITY IN CLASSIFIER RANKING ON REAL DATA

	5%	0.5%	0.1%	Total
Avg.Correlation				
F-Measure	1	1	1	1
Reference	0.9988	0.9903	0.9006	0.9632
Avg. Word-Edit Distance				
F-Measure	0	0	0	0
Reference	0.2	0.6	3.9	1.56
Avg. Sequence Alignment Cost				
F-Measure	0	0	0	0
Reference	0.2	1.2	7.6	3

TABLE VI
RELIABILITY IN CLASSIFIER RANKING ON ARTIFICIAL DATA

After these tests, we can make the following conclusions:

1) On real images overall average ranking correlation is slightly better for F-Measure, but results are on par with the reference classifier method, and actually significantly better for the latter in most DIBCO series. Furthermore, the latter also outperforms F-Measure on the word-edit distance and sequence alignment. Overall, the reference classifier method does a better job ranking classifiers.
2) On the artificial images, tests show better correlation, sequence alignment cost and word edit distance with the F-measure. Also, the reference approach significantly decreases in reliability as classifiers converge to lower error rates.

VI. Conclusion

We can conclude that both approaches are reliable substitutes for performance evaluation when *Ground Truth* is unavailable or unreliable. However, this mainly holds provided the classifiers present sufficient levels of difference. Although F-measure seems to perform better than the reference method on artificial data, this doesn't seem to be consistently the case on real data, and will require further investigation.

It would be useful to extend the validation of each of the statistically based approaches on different types of real data and associated binary classifiers (*e.g.* speech recognition, image recognition ...). Further work on this topic will extend both proposed methods to multiclass classifier evaluation.

VII. Acknowledgement

Perceptron neural networks are the contribution of V. Razinina. She and M. Fedorchuk acknowledge Erasmus+ funding between NTUU–KPI and the Université de Lorraine.

References

[1] E. H. Barney Smith. An analysis of binarization ground truthing. In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, pages 27–34, New York, NY, USA, 2010. ACM.
[2] E. H. Barney Smith and C. An. Effect of "ground truth" on image binarization. In *2012 10th IAPR International Workshop on Document Analysis Systems*, pages 250–254, Mar 2012.
[3] J. Bernsen. Dynamic thresholding of grey-level images. In *Proceedings of the 8th International Conference on Pattern Recognition*, pages 1251–1255, Oct 1986.
[4] B. Bostanci and E. Bostanci. An Evaluation of Classification Algorithms Using Mc Nemar's Test. In J. C. Bansal, P. K. Singh, K. Deep, M. Pant, and A. K. Nagar, editors, *Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012): Volume 1*, pages 15–26, India, 2013. Springer India.
[5] D. Bradley and G. Roth. Adaptive thresholding using the integral image. *J. Graphics Tools*, 12(2):13–21, 2007.
[6] N. Chen and D. Blostein. A survey of document image classification: problem statement, classifier architecture and performance evaluation. *International Journal of Document Analysis and Recognition (IJDAR)*, 10(1):1–16, Jun 2007.
[7] M. Fedorchuk and B. Lamiroy. Statistic Metrics for Evaluation of Binary Classifiers without Ground-Truth. In *IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)* , Kiev, Ukraine, May 2017. IEEE.
[8] B. Gatos, I. Pratikakis, and S. J. Perantonis. Adaptive degraded document image binarization. *Pattern Recognition*, 39(3):317–327, Mar 2006.
[9] J. Kittler and J. Illingworth. Minimum error thresholding. *Pattern Recognition*, 19(1):41 – 47, 1986.
[10] B. Lamiroy. Interpretation, Evaluation and the Semantic Gap ... What if we Were on a Side-Track? In B. Lamiroy and J-M. Ogier, editors, *10th IAPR International Workshop on Graphics Recognition, GREC 2013*, volume 8746, pages 213–226, Bethlehem, PA, USA, 2014. Springer.
[11] B. Lamiroy and P. Pierrot. Statistical Performance Metrics for Use with Imprecise Ground-Truth. In B. Lamiroy and R. Dueire Lins, editors, *Graphics Recognition. Current Trends and Challenges: 11th International Workshop on Graphics Recognition, GREC 2015*, volume 9657 of *LNCS*, Nancy, France, 2017. Springer.
[12] B. Lamiroy and T. Sun. Computing Precision and Recall with Missing or Uncertain Ground Truth. In Y-B. Kwon and J-M. Ogier, editors, *Graphics Recognition. New Trends and Challenges. 9th International Workshop, GREC 2011, Seoul, Korea, Sep. 15-16, Revised Selected Papers*, volume 7423 of *LNCS*, pages 149–162. Springer, Feb 2013.
[13] Y. LeCun and C. Cortes. MNIST handwritten digit database. 2010.
[14] V. I. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707, Feb 1966.
[15] W. Niblack. *An Introduction to Digital Image Processing*. Strandberg Publishing Company, Birkeroed, Denmark, Denmark, 1985.
[16] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, Jan 1979.
[17] I. Pratikakis, B. Gatos, and K. Ntirogiannis. ICDAR 2013 document image binarization contest (DIBCO 2013). In *12th International Conference on Document Analysis and Recognition, Washington, DC, USA, August 25-28, 2013*, pages 1471–1476. IEEE Computer Society, 2013.
[18] B. Raj, R. Singh, and J. Baker. A paired test for recognizer selection with untranscribed data. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5676–5679, May 2011.
[19] J. Sauvola, T. Seppanen, S. Haapakoski, and M. Pietikainen. Adaptive document binarization. In *Proceedings of the Fourth International Conference on Document Analysis and Recognition*, volume 1, pages 147–152 vol.1, Aug 1997.
[20] P. Stathis, E. Kavallieratou, and N. Papamarkos. An evaluation technique for binarization algorithms. *Journal of Universal Computer Science*, pages 3011–3030, 2008.
[21] R.A. Wagner and M.J. Fischer. The string-to-string correction problem. *J. ACM*, 21(1):168–173, Jan 1974.
[22] C. Wolf and J.-M. Jolion. Extraction and recognition ofartificial text in multimedia documents. *Formal Pattern Analysis &Applications*, 6(4):309–326, Feb 2004.