Unsupervised Feature Extraction via Deep Learning for Histopathological Classification of Colon Tissue Images: Supplementary Material

Can Taylan Sari and Cigdem Gunduz-Demir*, *Member, IEEE*

***Abstract*—Histopathological examination is today's gold standard for cancer diagnosis. However, this task is time consuming and prone to errors as it requires a detailed visual inspection and interpretation of a pathologist. Digital pathology aims at alleviating these problems by providing computerized methods that quantitatively analyze digitized histopathological tissue images. The performance of these methods mainly rely on features that they use, and thus, their success strictly depends on the ability of these features successfully quantifying the histopathology domain. This technical report contains the supplementary material for the new unsupervised feature extractor that we developed for effective representation and classification of histopathological tissue images [1].**

***Index Terms*—Deep learning, feature learning, histopathological image representation, digital pathology, automated cancer diagnosis, saliency, colon cancer, hematoxylin-eosin staining.**

I. Introduction

We recently developed a new unsupervised feature extractor, which we called *DeepFeature* [1], for representation and classification of histopathological images. This extractor defines salient subregions around cytological tissue components and characterizes them in an unsupervised way. In this characterization, it learns the local features of the salient subregions by a deep belief network consisting of consecutive RBMs and quantizes them by clustering these local features by the k-means algorithm. At the end, it represents and classifies the image with the distribution of its quantized subregions.

In [1], we presented experimental results for two different classification datasets. This technical report provides supplementary experiments. It first presents the receiver operating characteristic (ROC) curves for these classifications together with their area under the curve (AUC) metrics. It then provides the parameter analysis for the second dataset; the analysis for the first dataset was already given in [1]. In the original paper, we discussed how the proposed classification system can be used in a digital pathology setup, in which typically lower magnifications are used to scan a slide. Thus, the produced images usually have a larger field of view and may be homogeneous or heterogeneous. To this end, we implemented a simple colon adenocarcinoma detection algorithm and presented its visual results. The last section of this technical report provides the visual results obtained for additional images.

C. T. Sari is with the Department of Computer Engineering, Bilkent University, Ankara TR-06800, Turkey (e-mail: can.sari@bilkent.edu.tr).

*C. Gunduz-Demir is with the Department of Computer Engineering and Neuroscience Graduate Program, Bilkent University, Ankara TR-06800, Turkey (e-mail: gunduz@cs.bilkent.edu.tr).

TABLE I
FOR THE FIRST DATASET, THE AREA UNDER THE CURVE (AUC) METRICS OF THE PROPOSED *DeepFeature* METHOD AND THE COMPARISON ALGORITHMS. THESE METRICS ARE CALCULATED ON THE TEST SAMPLES OF THIS DATASET.

	Norm.	Low	High	Arith. mean	Harm. mean
DeepFeature	0.9974	0.9895	0.9942	0.9937	0.9937
Handcrafted features					
CooccurrenceMatrix	0.9618	0.9615	0.9418	0.9550	0.9549
GaborFilter	0.9728	0.9584	0.9452	0.9588	0.9587
LocalObjectPattern [2]	0.9901	0.9756	0.9841	0.9833	0.9832
TwoTier [3]	0.9996	0.9907	0.9872	0.9925	0.9925
Deep learning for supervised classification					
AlexNet	0.9990	0.9848	0.9750	0.9863	0.9862
GoogLeNet	1.0000	0.9913	0.9859	0.9924	0.9923
Inception-v3	1.0000	0.9882	0.9827	0.9903	0.9902
Deep learning for feature extraction (salient points)					
SalientStackedAE	0.9982	0.9885	0.9888	0.9918	0.9918
SalientConvolutionalAE	0.9984	0.9651	0.9293	0.9643	0.9635
Deep learning for feature extraction (random points)					
RandomRBM	0.9950	0.9837	0.9935	0.9907	0.9907
RandomStackedAE [4]	0.9976	0.9836	0.9811	0.9874	0.9874
RandomConvolutionalAE	0.9927	0.9528	0.9224	0.9560	0.9551

II. ROC Curves and AUC Analysis

This section presents the ROC curve and AUC analysis. Although this analysis is well defined for binary classifications, there is no consensus on how to obtain the ROC curves for multi-class classification problems. In our experiments, we follow the following procedure for both our proposed method and the comparison algorithms used in [1]. In this procedure, we generate a ROC curve for each class separately, by considering only the posterior probabilities that the multi-class SVM classifier outputs for this particular class (we do not consider the posteriors of the other classes). We threshold these posteriors with the threshold values across the $[0, 1]$ interval and obtain the true positive rate (TPR) and the false positive rate (FPR) for each threshold. We then use these rates to generate the ROC curve.

After obtaining the ROC curve for each class separately, we calculate the area under this curve. Tables I and II report the class-specific AUC metrics obtained on the test samples of the first and second datasets, respectively. Note that the last two

TABLE II
FOR THE SECOND DATASET, THE AREA UNDER THE CURVE (AUC) METRICS OF THE PROPOSED *DeepFeature* METHOD AND THE COMPARISON ALGORITHMS. THESE METRICS ARE CALCULATED ON THE TEST SAMPLES OF THIS DATASET.

	Norm.	Low (grade1)	Low (grade1-2)	Low (grade2)	High	Arith. mean	Harm. mean
DeepFeature	0.9991	0.9752	0.9284	0.9206	0.9727	0.9592	0.9582
Handcrafted features							
CooccurrenceMatrix	0.9808	0.9083	0.8228	0.7971	0.9541	0.8926	0.8867
GaborFilter	0.9692	0.9100	0.8056	0.8234	0.9483	0.8913	0.8864
LocalObjectPattern [2]	0.9899	0.9622	0.9084	0.8946	0.9612	0.9433	0.9419
TwoTier [3]	0.9997	0.9651	0.8865	0.9001	0.9725	0.9448	0.9427
Deep learning for supervised classification							
AlexNet	0.9974	0.9802	0.8939	0.9132	0.9766	0.9523	0.9505
GoogLeNet	1.0000	0.9893	0.9326	0.8764	0.9764	0.9549	0.9527
Inception-v3	0.9999	0.9773	0.9015	0.9234	0.9677	0.9540	0.9526
Deep learning for feature extraction (salient points)							
SalientStackedAE	0.9998	0.9736	0.9259	0.9130	0.9590	0.9543	0.9532
SalientConvolutionalAE	0.9991	0.9337	0.8539	0.8397	0.9530	0.9159	0.9119
Deep learning for feature extraction (random points)							
RandomRBM	0.9951	0.9588	0.8923	0.9167	0.9693	0.9465	0.9450
RandomStackedAE [4]	0.9993	0.9544	0.8750	0.8894	0.9560	0.9348	0.9325
RandomConvolutionalAE	0.9906	0.9185	0.8549	0.8244	0.9157	0.9008	0.8972

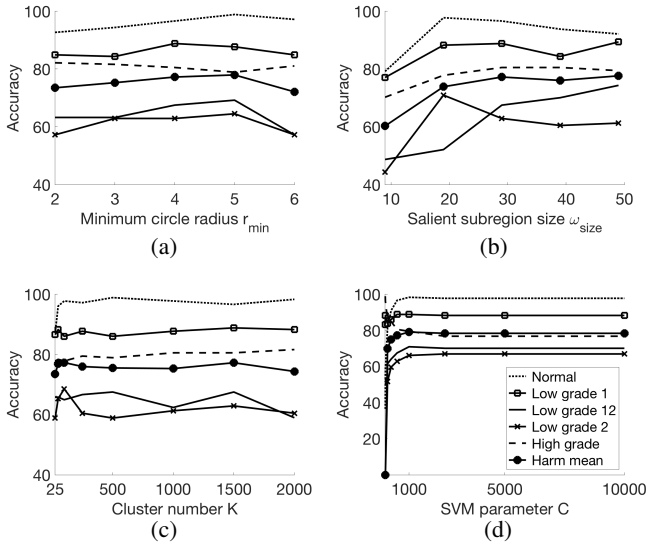


Fig. 1. For the second dataset, test set accuracies as a function of the model parameters: (a) minimum circle radius r_{min}, (b) size of a salient subregion ω_{size}, (c) cluster number K, and (d) SVM parameter C. The parameter analysis for the first dataset were given in [1].

columns of these tables present the averages of these class-specific AUC metrics. Here we provide the arithmetic mean of the class-specific AUC metrics as well as their harmonic mean since the arithmetic mean can sometimes be misleading when values to be averaged differ greatly. These tables indicate the effectiveness of our proposed *DeepFeature* method for the representation and classification of histopathological images. It yields better results than the other algorithms, which is also consistent with our findings reported in [1]. The ROC curves used in the calculation of these AUC values are presented in Figs. 2 and 3 for the first dataset, and in Figs. 4 and 5 for the second one.

III. PARAMETER ANALYSIS

The proposed *DeepFeature* method has four external parameters: minimum circle radius r_{min}, size of a salient subregion ω_{size}, cluster number K, and SVM parameter C. The selection of these parameters and their analyses for the first dataset were given in [1]. This section gives the parameter analysis for the second dataset. In this analysis, for each parameter, the values of the other three parameters are fixed and the test set accuracies are measured as a function of the parameter of interest.

The minimum circle radius r_{min} determines the size of the smallest circular object (tissue component) to be located. Its larger values may cause an inadequate representation since they cause not to define smaller objects, which may correspond to important small tissue components such as nuclei, and salient subregions around them. This lowers the accuracy. Its smaller values define noisy objects and using their salient subregions slightly decreases the accuracy. This analysis is depicted in Fig. 1(a).

The size of a salient subregion ω_{size} determines the locality of the deep features extracted from salient subregions. When ω_{size} is too small, it is not sufficient to accurately characterize the subregion, which significantly decreases the accuracy. After a certain point, it does not affect the accuracy too much, but of course, increases the complexity of the required deep neural network. This analysis is depicted in Fig. 1(b).

The cluster number K determines the number of labels used to quantize the salient subregions (components). Its smaller values may result in defining the same label for components of different types. This may lead to an ineffective representation, decreasing the accuracy. Its larger values only slightly affect the performance. This analysis is depicted in Fig. 1(c).

The SVM parameter C controls the trade-off between the training error and the margin width of the model. Unfortunately, there is no foolproof method for its selection and its value must be determined empirically. As shown in Fig. 1(d),

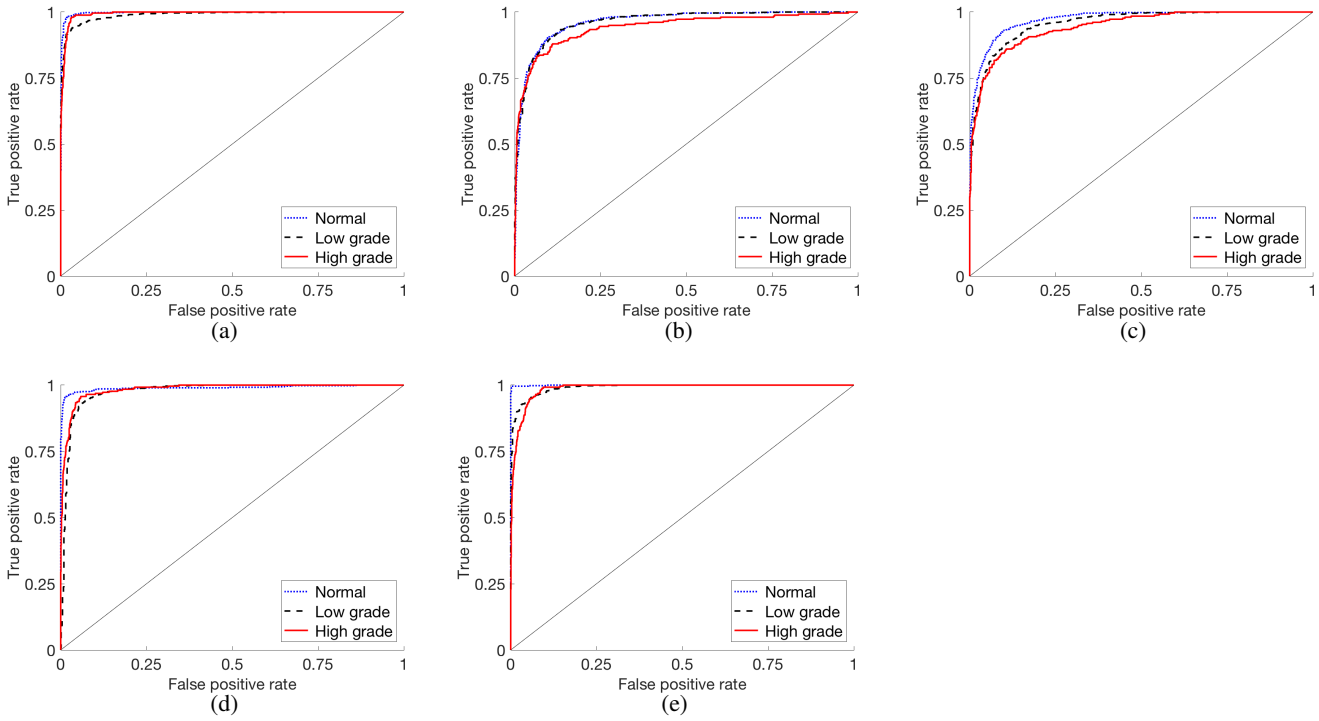


Fig. 2. ROC curves for the test samples of the first dataset. These curves are obtained for the proposed *DeepFeature* method and the comparison algorithms that use handcrafted features: (a) *DeepFeature*, (b) CooccurrenceMatrix, (c) GaborFilter, (d) LocalObjectPattern [2], and (e) TwoTier [3] methods.

our application necessitates the use of C in the range between 250 and 1000.

IV. Visual Results of the Detection Algorithm

In [1], we discussed how the proposed image representation and classification system can be used in a digital pathology setup. To that end, we outlined a simple algorithm for an example application, in which the aim is to detect low-grade and high-grade colon adenocarcinomatous regions on large images as well as those containing normal colon glands. The visual results of this colon adenocarcinoma detection algorithm were given in [1]. This section provides the visual results for additional images. The results given in Fig. 6 indicate that this algorithm is good at detecting the regions of interest on many large heterogeneous images.

This section also discusses probable misclassifications of the detection algorithm, on illustrative examples. First, it may incorrectly output the cancer grade because of an error in the SVM classifier. The examples for such type of a misclassification are illustrated in Figs. 7(a) and 7(b). In these examples, low-grade cancerous regions are incorrectly classified as high-grade. Here it is worth to noting that the detection in the latter example is much more difficult since it contains heterogeneous tumor (with multiple grades). The second type of error may occur due to the existence of unannotated non-epithelial regions. These regions are left as unannotated in our datasets, on which the classification system was trained, since colon adenocarcinoma mainly affects epithelial cells and non-epithelial regions are not that informative for the diagnosis of this cancer type. When these regions are small, incorrect classifications can be compensated by correct classifications of nearby regions and the reject action. On the other hand, when they are large, such compensation may or may not be possible and the system may give incorrect results since there is no separate class for these non-epithelial regions. Such an example is given in Fig. 7(c). Defining an extra class(es) will definitely improve the accuracy on these regions. This is left as future research work of our study.

References

[1] C. T. Sari and C. Gunduz-Demir, "Unsupervised feature extraction via deep learning for histopathological classification of colon tissue images," *IEEE Trans. Med. Imaging*, submitted 2018.

[2] G. Olgun, C. Sokmensuer, and C. Gunduz-Demir, "Local object patterns for tissue image representation and cancer classification," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 4, pp. 1390-1396, Jul. 2014.

[3] T. Gultekin, C. F. Koyuncu, C. Sokmensuer, and C. Gunduz-Demir, "Two-tier tissue decomposition for histopathological image representation and classification," *IEEE Trans. Med. Imaging*, vol. 34, no. 1, pp. 275-283, Jan. 2015.

[4] J. Arevalo, A. Cruz-Roa, V. Arias, E. Romero, and F. A. Gonzalez, "An unsupervised feature learning framework for basal cell carcinoma image analysis," *Art. Intel. Medicine*, vol. 64, no. 2, pp. 131-145, Jun. 2015.

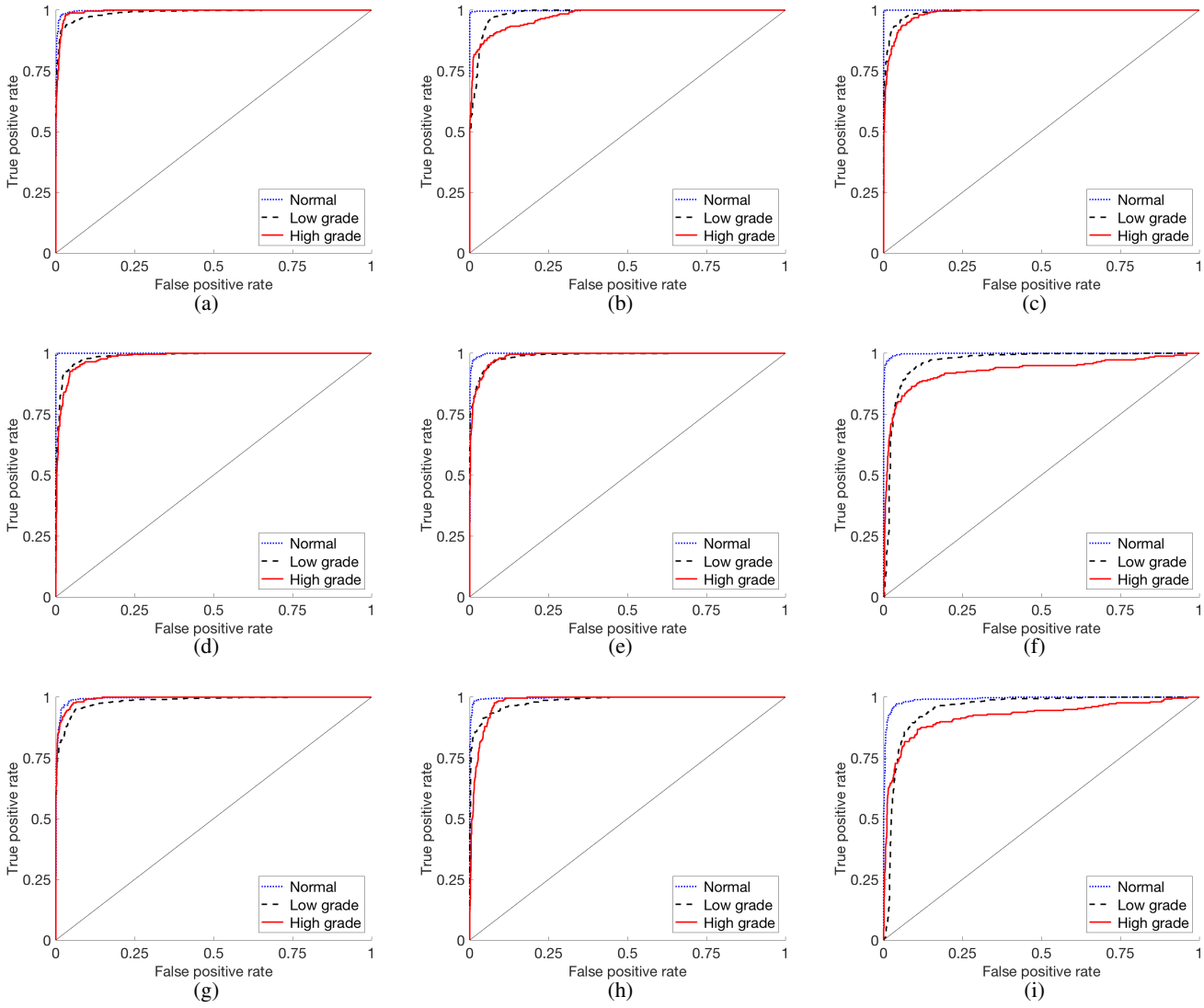


Fig. 3. ROC curves for the test samples of the first dataset. These curves are obtained for the proposed *DeepFeature* method and the deep learning based comparison algorithms: (a) *DeepFeature*, (b) AlexNet, (c) GoogLeNet, (d) Inception-v3, (e) SalientStackedAE, (f) SalientConvolutionalAE, (g) RandomRBM, (h) RandomStackedAE [4], and (i) RandomConvolutionalAE methods.

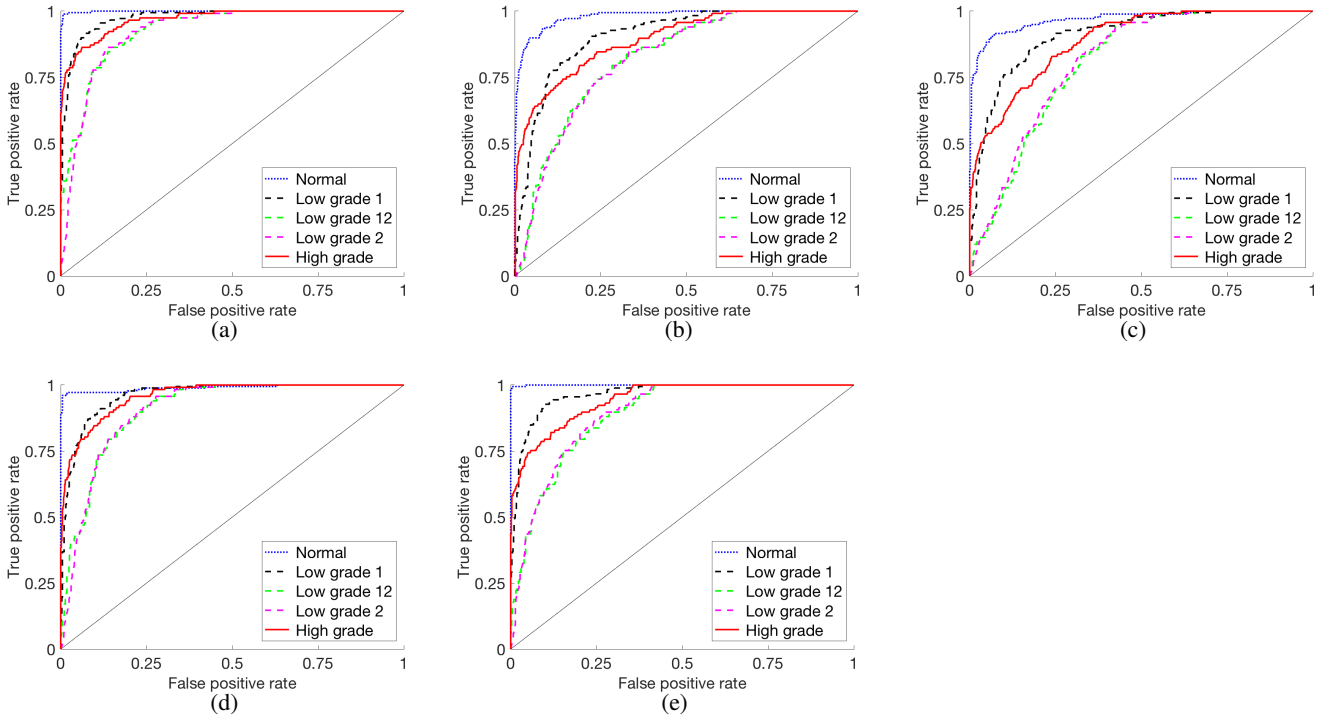


Fig. 4. ROC curves for the test samples of the second dataset. These curves are obtained for the proposed *DeepFeature* method and the comparison algorithms that use handcrafted features: (a) *DeepFeature*, (b) CooccurrenceMatrix, (c) GaborFilter, (d) LocalObjectPattern [2], and (e) TwoTier [3] methods.

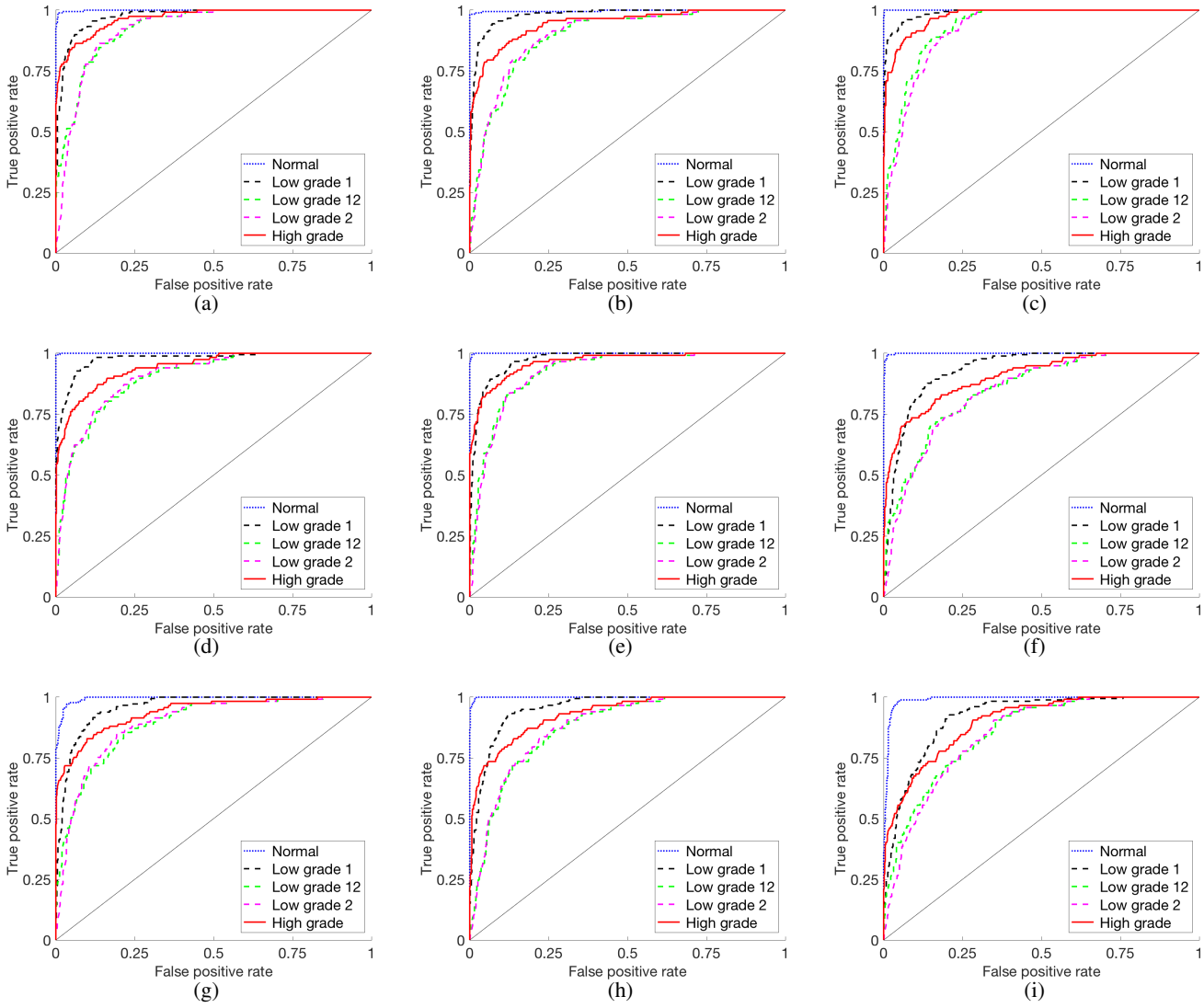


Fig. 5. ROC curves for the test samples of the secomd dataset. These curves are obtained for the proposed *DeepFeature* method and the deep learning based comparison algorithms: (a) *DeepFeature*, (b) AlexNet, (c) GoogLeNet, (d) Inception-v3, (e) SalientStackedAE, (f) SalientConvolutionalAE, (g) RandomRBM, (h) RandomStackedAE [4], and (i) RandomConvolutionalAE methods.

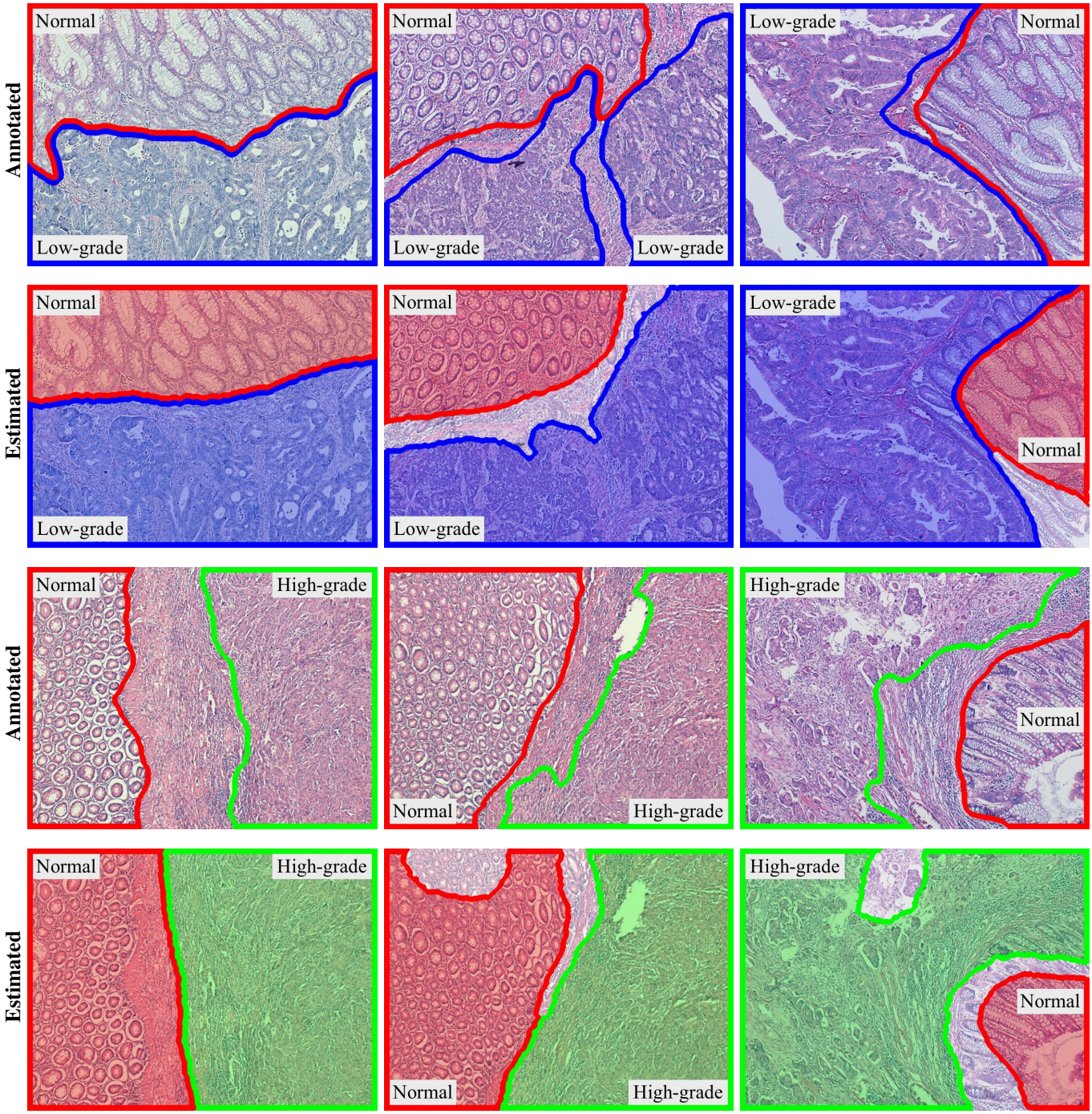
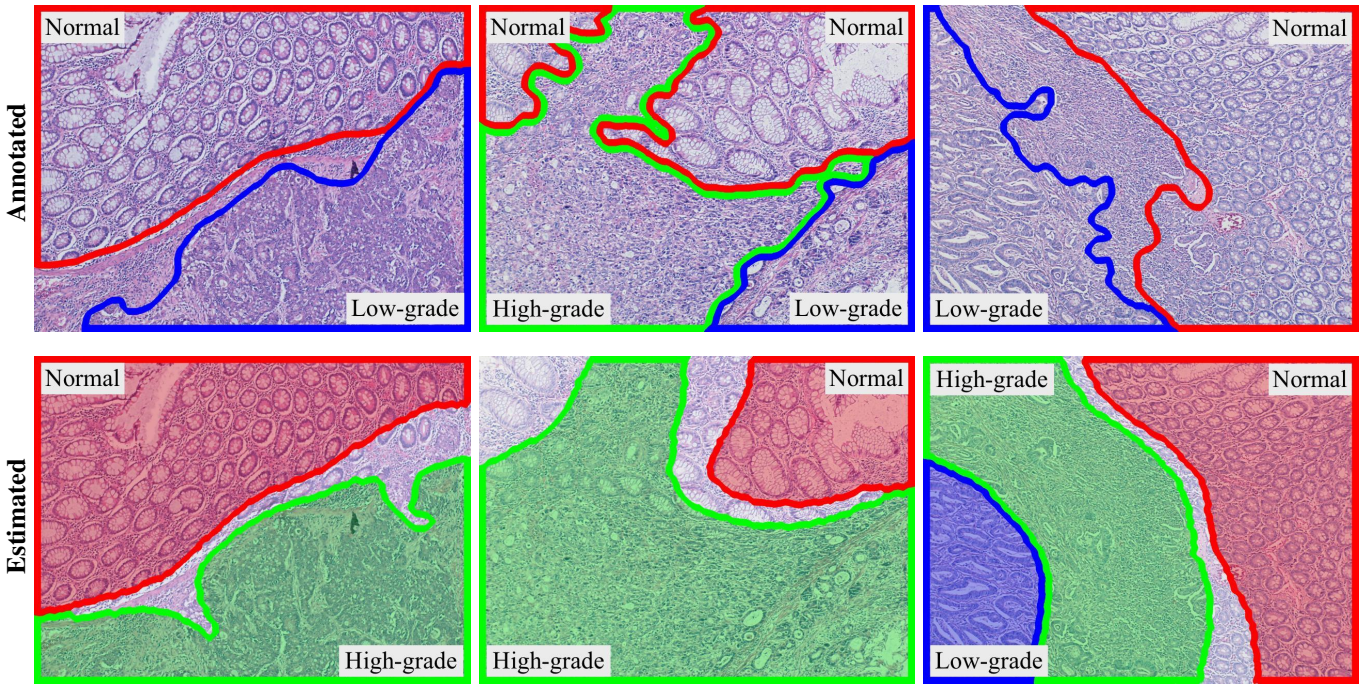


Fig. 6. Examples of large heterogeneous images together with their visual results obtained by the colon adenocarcinoma detection algorithm. The boundaries of the annotated/estimated normal, low-grade cancerous, and high-grade cancerous regions are shown with red, blue, and green, respectively.



(a) (b) (c)

Fig. 7. Examples of large heterogeneous images that contain regions whose types are incorrectly estimated by our detection algorithm. The boundaries of the annotated/estimated normal, low-grade cancerous, and high-grade cancerous regions are shown with red, blue, and green, respectively.