Computer Science and Application 计算机科学与应用, 2021, 11(2), 277-284
Published Online February 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.112028

Hans汉斯

基于原型的乳腺癌病理图像分类算法

陈　霧，穆国旺

河北工业大学，天津
Email: muguow@hebut.edu.cn

收稿日期：2021年1月8日；录用日期：2021年2月3日；发布日期：2021年2月10日

摘　要

乳腺癌病理图像的自动分类是开发乳腺癌计算机辅助诊断系统的关键。针对乳腺肿瘤病理图像的特点，提出了一种基于LBP原型的特征提取方法，首先，随机地从训练图像中选取若干子图像作为原型并提取其LBP特征；其次，对于任意输入图像，计算图像中所有和原型同样大小的子图像和原型的LBP特征的余弦距离，然后，通过池化操作得到最终的特征；最后，利用SVM集成的方法对图像进行分类。在BreakHis数据集上对算法进行了验证，结果表明，本文提出的特征提取方法优于一些传统的方法。

关键词

计算机辅助诊断，乳腺癌病理图像分类，机器学习，局部二值模式，支持向量机

Algorithm for Breast Cancer Histopathological Image Classification with Prototypes

Wu Chen, Guowang Mu

Hebei University of Technology, Tianjin
Email: muguow@hebut.edu.cn

Received: Jan. 8th, 2021; accepted: Feb. 3rd, 2021; published: Feb. 10th, 2021

Abstract

Automatic classification of breast cancer pathological images is the key to the development of computer-aided diagnosis system for breast cancer. According to the characteristics of breast tumor pathological images, a feature extraction method based on LBP prototypes was proposed.

文章引用: 陈霧, 穆国旺. 基于原型的乳腺癌病理图像分类算法[J]. 计算机科学与应用, 2021, 11(2): 277-284.
DOI: 10.12677/csa.2021.112028

Firstly, some sub-images were randomly selected from the training images as prototypes and their LBP features were extracted. Secondly, for arbitrary input images, the cosine distances between the LBP features of all sub-images of the same size and those of the prototypes were calculated, and then, the final features were obtained by pooling operation. Finally, the images were classified by integration of SVMs. Algorithm is verified on the BreakHis dataset. The experimental results show that the proposed feature ex-traction method is superior to some traditional methods.

Keywords

Computer-Aided Diagnosis, Breast Cancer Histopathological Image Classification, Machine Learning, Local Binary Pattern, SVM

1. 引言

乳腺癌(Breast Cancer, BC)是全世界女性发病率和死亡率最高的癌症，而组织病理学分析是最可靠的癌症诊断方法[1] [2] [3]。组织病理学分析是一项高度耗时的专业工作，通常是需要由病理学专家对乳腺癌组织病理学(Breast Cancer Histopathological, BCH)图像进行观察，根据经验来判断肿瘤是良性还是恶性，或者说对 BCH 图像人工进行分类，它十分依赖于病理学家的经验，并且难以避免病理学专家受到疲劳和注意力下降等因素的影响造成误诊。在此背景下，迫切需要相应的计算机辅助诊断来减轻病理学专家的工作负担，其中 BCH 图像分类在计算机辅助诊断中具有重要意义[2]。

BCH 图像分类已经成为医学图像分类领域上的一个研究热点，目前已经有很多学者对此进行了研究。目前关于 BCH 图像分类算法可以分为两大类：基于传统机器学习的分类算法和基于深度学习的分类算法。基于传统机器学习的分类算法步骤为：预处理[4] [5] [6] [7]、特征抽取和选择[8] [9] [10] [11]、分类[11] [12] [13]。其中特征提取和选择、分类是图像分类中的关键。Spanhol 等[11]将 LBP (Local Binary Pattern，局部二值模式)、CLBP、LPQ、GLCM、ORB 和 PFTAS 等特征，分别与 SVM，RF，QDA，Nearest Neigbor 分类器相结合用于 BCH 图像良恶性分类。Vibha Gupta 等[13]提出了一种将 Gabor、OCLBP、颜色、纹理等多种特征融合，并采用投票机制的异构集成分类器对 BCH 图像进行分类的方法。Shukla K. K.等[14]提出了利用形态学特征对 BCH 图像进行自动检测和分类的方法，使用直方图均衡化改善图像的局部对比度，采用 TWS 进行分割，并对 MLP，LMT，RF，Rotation Forest，SMO，Naïve Bayes，J-Rip 和 PART 等多种分类器进行了比较研究。由于深度学习的发展，越来越多的文献使用深度学习的方法对 BCH 图像分类。Pimkin [15]等人采用卷积神经网络(CNN)架构来进行 BCH 图像分析，提出对图像小块进行分类以增加有效样本的数量，然后应用集成技术对原始图像进行预测。Marami [16]等人提出一种集成了 4 个改进的 incepin-v3 神经网络的自动分类方法。而 Spanhol [8]等人提出了一种基于提取图像小块来训练 CNN，然后将这些小块结合起来进行最终分类的方法。

虽然人们已经对 BCH 图像分类进行了很多的研究，但是，现有的识别方法正确率还不能满足实际应用的要求。究其原因，在有关 BCH 图像分类的文献中所使用的很多特征提取方法，其中，Gabor、LBP、CLBP、LPQ 等特征主要应用于人脸识别等应用场合，为了得到好的识别结果，首先要将人脸区域分割出来，使得每幅图只包含一个人脸，并且要通过归一化处理使得不同图像中的关键部位(如眼睛、鼻子)尽量

对齐。但是，在每一幅 BCH 图像中却包含大量不同形态的细胞，其中既有肿瘤细胞，也有正常细胞，而且肿瘤细胞在图像中的位置是随机的。因此，对整幅图像提取 LBP、CLBP、LPQ 等特征，包含了大量对分类结果无用的区域的颜色或纹理信息，对分类不仅无用而且造成干扰。采用其它一些纹理或颜色特征，例如 ORB、PFTAS、OCLBP 等特征，存在同样的问题。

去除这种干扰的一种理想方法是，将图像中的肿瘤细胞都分割出来，然后对每个肿瘤细胞进行分类，最后对各个细胞分类的结果进行综合，得到最终的分类结果。但是，要将图像中每个肿瘤细胞分割出来的算法比较复杂，而且工作量巨大。

为此，本文提出了一种基于原型的高级特征提取方法。首先，随机地从所有训练样本中选取若干个小的图像块(本文取 10 × 10 的图像块)，称为原型。然后，对于任意一幅图像和一个原型，取和原型同样大小的所有子块，计算各个子块的 LBP 特征和原型的 LBP 特征的余弦距离，并取其中最小的若干个距离的平均值作为和该原型相关的高级特征。因此，最终的高级特征的维数和原型的个数相同。最后，采用 SVM 分类器集成的方法对 BCH 图像进行分类。在 BreakHis 数据集上进行实验，取得了很好的分类效果。

2. 基于 LBP 原型的特征

2.1. 局部二值模式

LBP 是由 Ojala 等人[17]在 1994 年提出的一种用来描述图像局部纹理特征的算子。原始的 LBP 算子定义在一个 3 × 3 的窗口内。以窗口中心像素为阈值，与相邻的 8 个像素的灰度值比较，若周围的像素值大于中心像素值，则该位置被标记为 1，否则标记为 0，得到一个 8 位二进制数，将这个值作为窗口中心像素点的 LBP 编码，它反映了该像素附近区域的纹理信息。像素 (x_c, y_c) 的 LBP 编码用公式可以表示为：

$$\mathrm{LBP}(x_c, y_c) = \sum_{p=1}^{8} s(I(p) - I(c)) * 2^P \tag{1}$$

其中，p 表示 $N \times N$ 窗口除中心像素点外的第 p 个像素点；$I(c)$ 表示中心像素点的灰度值；$I(p)$ 表示第 p 个像素点的灰度值，$s(x)$ 公式如下：

$$s(x) = \begin{cases} 1, x \geq 0 \\ 0, \text{otherwise} \end{cases} \tag{2}$$

在将局部二值模式(LBP)应用于图像分类时，一般要将图像分成若干子块，统计每个子块中像素的 LBP 编码的直方图，并将它们串接起来作为图像的特征向量。原始的局部二值模式有 256 种，因此，每个子块的 LBP 编码的直方图是一个 256 维的向量。

Ojala 等人[17]通过实验证明，在实际图像中，绝大多数 LBP 模式最多只包含两次从 1 到 0 或从 0 到 1 的跳变。当某个 LBP 所对应的循环二进制数从 0 到 1 或从 1 到 0 最多有两次跳变时，该 LBP 称为一个均匀(uniform)二值模式。如 00000000 (0 次跳变)，10001111 (先由 1 跳到 0，再由 0 跳到 1，共两次跳变)都是均匀二值模式。于是，二进制模式由原始的 256 种减少为 59 种，将其用 0~58 进行编码(例如，1~58 表示 58 种均匀二值模式，0 表示非均匀二值模式)。这样直方图从原来的 256 维变成 59 维，这使得特征向量的维数更少，并且可以减少高频噪声带来的影响。本文我们采用均匀二值编码(每个像素的均匀二值模式取值 0~58)。

2.2. 基于 LBP 原型的特征提取算法

在 BCH 图像分类中，每幅图像由很多不同形状的细胞组成，细胞可以分为正常细胞和癌细胞两种。在不把图像中所有细胞分割出来的情况下，目前的特征提取方法将所有细胞放在一起考虑，所提取的特

征缺乏针对性，因此识别率普遍不高。本文设计了一种基于 LBP 原型的特征，试图通过原型来捕捉正常细胞和癌细胞的信息。本文特征提取的过程如图 1 所示，具体步骤如下：

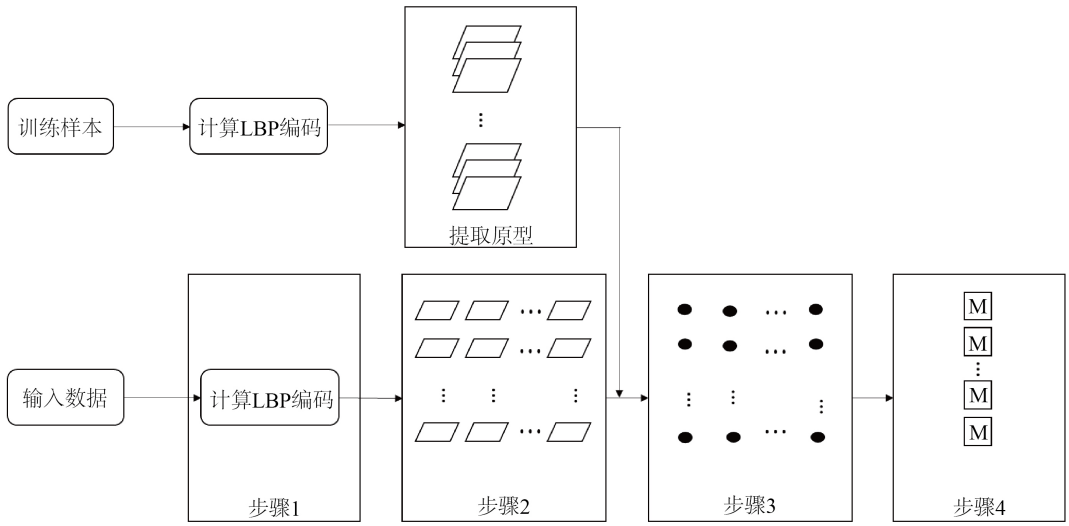

Figure 1. Schematic diagram of advanced feature based on prototype
图 1. 基于原型的高级特征提取示意图

2.2.1. 准备阶段

对于训练集中的每幅图像，计算各个像素的 LBP 编码，得到 LBP 图像；然后从所有 LBP 图像中，随机选取 K 个大小为 $m\times m$ 的子图像(patch)，作为原型。记第 k 个原型为 $\mathbf{P}_k$，统计每个原型的 LBP 编码直方图，记为 $\mathbf{H}\left(\mathbf{P}_k\right)$，$k=1,2,\cdots,K$。选取 m 数值的大小和图像中细胞的大小相近。

2.2.2. 特征提取阶段

对于任一幅输入图像 $I\left(x,y\right)$，本文的特征提取过程分为以下几个步骤：

步骤 1：计算每个像素的 LBP 编码，得到 LBP 图像 $I_{\mathrm{LBP}}\left(x,y\right)$，$x=1,2,\cdots,M$，$y=1,2,\cdots,N$。

步骤 2：在 LBP 图像 $I_{\mathrm{LBP}}\left(x,y\right)$ 中取所有 $m\times m$ 的子图像 $\mathbf{B}_{i,j}$，$i=1,2,\cdots,M-m+1$，$j=1,2,\cdots,N-m+1$，统计每个子图像的 LBP 编码直方图，记为 $\mathbf{H}\left(\mathbf{B}_{i,j}\right)$。

步骤 3：对于每个原型 $\mathbf{P}_k\left(k=1,2,\cdots,K\right)$，计算 $\mathbf{H}\left(\mathbf{B}_{i,j}\right)$ 和 $\mathbf{H}\left(\mathbf{P}_k\right)$ 的余弦距离，记为 $d_k\left(i,j\right)$，$i=1,2,\cdots,M-m+1$，$j=1,2,\cdots,N-m+1$。

步骤 4：对 $d_k\left(i,j\right)$，采用类似于卷积网络中的池化操作，得到最终的特征向量 $f=\left(f_1,f_2,\cdots,f_K\right)$。本文考虑了三种池化操作：

(1) 取最小距离：

$$f_k=\min_{i=1,2,\cdots,M-m+1;\,j=1,2,\cdots,N-m+1} d_k\left(i,j\right),k=1,2,\cdots,K \tag{3}$$

(2) 对最小的 n 个距离取均值：

设 $d_k\left(1\right),d_k\left(2\right),\cdots,d_k\left(n\right)$ 是 $\left\{d_k\left(i,j\right)\right\}$，$i=1,2,\cdots,M-m+1$，$j=1,2,\cdots,N-m+1$ 中最小的 n 个，则令 f_k 是 $d_k\left(1\right),d_k\left(2\right),\cdots,d_k\left(n\right)$ 的平均值，即：

$$f_k=\frac{1}{n}\sum_{i=1}^{n} d_k\left(i\right) \tag{4}$$

(3) 基于阈值的操作：

令 f_k 是小于某个阈值 δ 的所有距离之和，即：

$$f_k = \sum_{d_k(i,j)<\delta} d_k(i,j) \tag{5}$$

3. SVM 分类器集成

在提取基于 LBP 原型的特征之后，采用多个 SVM 集成的方法对 BCH 图像进行分类。每次从 K 个特征中随机地取 p 个特征，训练一个支持向量机。最后，从中选出在验证集上性能最优的 s 个支持向量机，作为最终的分类器。对于任意一幅输入图像，提取基于 LBP 原型的特征，再分别用 s 个支持向量机分类，将分类的结果按照投票的方法进行融合，得到最终的分类结果，如图 2 所示。

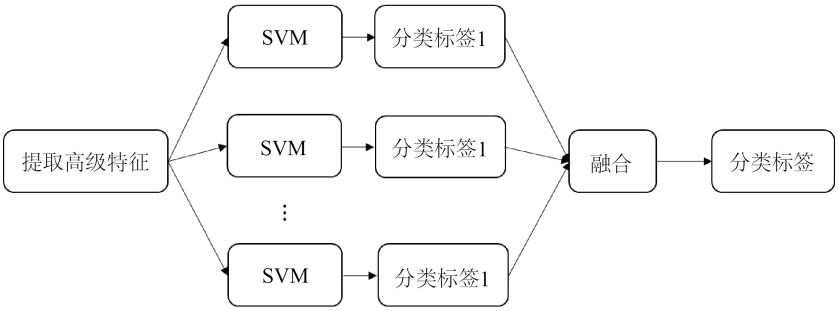

Figure 2. Flow chart of algorithm based on SVM Classifier ensemble
图 2. 基于 SVM 分类器集成的算法流程图

4. 实验结果

本文采用 BreaKHis 数据集对算法进行验证，该数据集在 2014 年由巴西 P&D 实验室采集，包括来自 82 位患者的 7909 幅已标注的乳腺肿瘤病理组织切片的电子显微图像，其中良性肿瘤图像 2480 幅，恶性肿瘤图像 5429 幅[11]。样本使用来自苏木精－伊红(HE)染色的乳房组织活检切片，并由 P&D 实验室的病理学家标记。每个病例的诊断均由经验丰富的病理学家完成，并通过免疫组织化学分析等辅助检查确认。图像采用 RGB 色彩空间，图像分辨率为 700 × 460，采用四种不同的放大倍数(40 倍，100 倍，200 倍和 400 倍)。

本文选择放大倍数为 40 的数据进行实验，测试遵循 BreaKHis 数据库协议，选择五倍交叉测试并取五折分类精度的均值作为最终分类结果。按照文献[11]，采用基于患者水平的识别率。设 N_P 为患者 p 的癌症图像，N_{rec} 是该患者被正确分类的图像数，则该患者评分定义为

$$\text{Patient Score} = \frac{N_{rec}}{N_P}$$

总的识别率定义为

$$\text{ACC} = \frac{\sum \text{Patient Score}}{\text{Total number of patients}}$$

在以下实验中，首先通过预处理将图像转化 400 × 256 的灰度图像，然后提取基于 LBP 原型的特征，最后，采用 SVM 分类器集成的方法对图像进行分类。

本文实验中，每次随机取 200 个特征，训练一个线性的支持向量机，重复 100 次，得到 100 个 SVM 分类器，从中选出最优的 k 个(本文 $k=5$)，对它们的分类进行采用投票的方法进行融合，得到最终的分类结果。

4.1. 原型图像大小对识别率的影响

首先，我们通过实验研究原型图像大小 m 对识别率的影响。在提取基于原型的特征时，取原型个数 $K=2000$，原型图像大小 m 分别取 6、8、10 和 12，并且采用池化方法(1)，识别结果见表 1。可见，当原型大小为 8×8 时，分类精度最高达到 77.83%，且方差也较小，效果较好。

Table 1. Algorithm classification results under different patch
表 1. 不同 patch 大小下算法分类结果

patch 大小	分类精度	方差
6	77.02%	2.46%
8	77.83%	2.08%
10	77.47%	2.27%
12	76.52%	2.04%

4.2. 原型个数对识别率的影响

为了研究原型个数对识别率的影响，我们固定原型的大小 $m=8$，并且采用池化方法(1)，原型的个数 K 分别取 1000、2000、3000 和 4000，识别的结果如表 2。

Table 2. Algorithm classification results under different patch numbers
表 2. 不同 patch 数量下算法分类结果

patch 数量	分类精度	方差
1000	77.02%	2.46%
2000	77.83%	2.08%
3000	77.70%	2.29%
4000	77.20%	1.39%

观察表 2 可知，原型数量 $K=2000$ 时，得到的识别率最高，为 77.83%。

4.3. 不同池化操作对识别率的影响

为了研究不同池化操作对结果的影响，我们固定原型的大小 $m=8$，原型个数 $K=2000$，分别采用三种不同的池化操作。在池化方法(2)的操作中，取 $n=50$。在池化方法(3)的操作中，取阈值 $\delta=0.3$，即将所有小于 0.3 的距离值求和，得到最终的特征。实验结果如表 3。

Table 3. Algorithm classification results under different pooling operations
表 3. 不同池化操作下算法分类结果

池化类型	分类精度	方差
方法(1)	77.83%	2.08%
方法(2)	79.66%	1.99%
方法(3)	74.39%	2.96%

由上表可知，采用池化方法(2)的识别率最高，达到了 79.66%，且方差最小，其次是池化方法(1)。

4.4. SVM 和 SVM 集成的比较

我们还通过实验对本文提出的 SVM 集成分类方法和简单的 SVM 进行比较，结果见表 4，可见本文提出的分类器集成方法显著提高了识别的准确率。

Table 4. Classification results
表 4. 分类结果

算法	分类精度	方差
SVM	74.70%	2.15%
SVM 集成	79.66%	1.99%

4.5. 和现有方法的比较

表 5 是本文提出的方法和现有文献中一些方法在 BreakHis 数据集上分类精度的对比。通过比较可以发现，针对乳腺癌病理图像分类的问题，本文提出的特征优于传统的一些特征，甚至优于一些深度学习特征。

Table 5. Existing BCH image classification results
表 5. 现存 BCH 分类结果

算法	分类精度
LBP + SVM [11]	74.20%
LPQ + SVM [11]	73.70%
ORB + SVM [11]	71.90%
ResNet34 [15]	76.00%

5. 实验结果

本文针对乳腺癌病理图像识别，首先提出了一种基于 LBP 原型的特征，其基本思想是，先随机地从训练图像中选取若干子图像作为原型，并提取其 LBP 特征，以此来捕获正常细胞和癌细胞的形态或模式；然后，对于每幅输入图像，提取图像中和原型同样大小的所有子图像的 LBP 特征，并计算和原型的余弦距离，最后对到同一原型的距离进行池化操作，得到该原型的一个距离值，作为最终的一个特征；其次，提出了一种基于 SVM 分类器集成的乳腺癌病理图像分类算法。在 BreaKHis 数据库上对算法进行了验证，并讨论了算法中不同参数对分类结果的影响，实验结果表明本文算法的有效性。

参考文献

[1] Boyle, P. and Levin, B. (2008) World Cancer Report 2008. Tech. Rep., International Agency for Research on Cancer, Lyon.

[2] Gurcan, M.N., Boucheron, L.E., Can, A., *et al*. (2009) Histopathological Image Analysis: A Review. *IEEE Reviews in Biomedical Engineering*, **2**, 147-171. https://doi.org/10.1109/RBME.2009.2034865

[3] Desir, C., Petitjean, C., Heutte, L., *et al*. (2012) Classification of Endomicroscopic Images of the Lung Based on Random Subwindows and Extra-Trees. *IEEE Transactions on Biomedical Engineering*, **59**, 2677-2683. https://doi.org/10.1109/TBME.2012.2204747

[4] Arevalo, J., Gonzalez, F.A., Ramospollan, R., *et al*. (2015) Convolutional Neural Networks for Mammography Mass Lesion Classification. *International Conference of the IEEE Engineering in Medicine and Biology Society*, Milano, 25-29 August 2015, 797-800. https://doi.org/10.1109/EMBC.2015.7318482

[5] Nejad, E.M., Affendey, L.S., Latip, R.B., *et al*. (2017) Classification of Histopathology Images of Breast into Benign and Malignant Using a Single-Layer Convolutional Neural Network. In: *Proceedings of the International Conference on Imaging*, *Signal Processing and Communication*, ACM, New York, 50-53. https://doi.org/10.1145/3132300.3132331

[6] Araujo, T., Aresta, G., Castro, E., *et al*. (2017) Classification of Breast Cancer Histology Images Using Convolutional Neural Networks. *PLoS ONE*, **12**, e0177544. https://doi.org/10.1371/journal.pone.0177544

[7] Zheng, Y., Jiang, Z., Xie, F., *et al*. (2017) Feature Extraction from Histopathological Images Based on Nucleus-Guided Convolutional Neural Network for Breast Lesion Classification. *Pattern Recognition*, **71**, 14-25. https://doi.org/10.1016/j.patcog.2017.05.010

[8] Spanhol, F.A., Oliveira, L.S., Cavalin, P.R., *et al*. (2017) Deep Features for Breast Cancer Histopathological Image Classification. *Systems*, *Man and Cybernetics*, Banff, 5-8 October 2017, 1868-1873. https://doi.org/10.1109/SMC.2017.8122889

[9] Doyle, S., Agner, S., Madabhushi, A., *et al*. (2008) Automated Grading of Breast Cancer Histopathology Using Spectral Clustering with Textural and Architectural Image Features. 2008 5*th IEEE International Symposium on Biomedical Imaging*: *From Nano to Macro*, Paris, 14-17 May 2008, 496-499. https://doi.org/10.1109/ISBI.2008.4541041

[10] Wang, P., Hu, X., Li, Y., *et al*. (2016) Automatic Cell Nuclei Segmentation and Classification of Breast Cancer Histopathology Images. *Signal Processing*, **122**, 1-13. https://doi.org/10.1016/j.sigpro.2015.11.011

[11] Spanhol, F.A., Oliveira, L.S., Petitjean, C., *et al*. (2015) A Dataset for Breast Cancer Histopathological Image Classification. *IEEE Transactions on Bio-Medical Engineering*, **63**, 1455-1462. https://doi.org/10.1109/TBME.2015.2496264

[12] Cruzroa, A., Ovalle, J.E., Madabhushi, A., *et al*. (2013) A Deep Learning Architecture for Image Representation, Visual Interpretability and Automated Basal-Cell Carcinoma Cancer Detection. *Medical Image Computing and Computer Assisted Intervention*, Nagoya, 22-26 September 2013, 403-410. https://doi.org/10.1007/978-3-642-40763-5_50

[13] Gupta, V. and Bhavsar, A. (2017) Breast Cancer Histopathological Image Classification: Is Magnification Important. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Honolulu, 21-26 July 2017, 17-24. https://doi.org/10.1109/CVPRW.2017.107

[14] Shukla, K.K., Tiwari, A. and Sharma, S. (2017) Classification of Histopathological Images of Breast Cancerous and Non Cancerous Cells Based on Morphological Features. *Biomedical and Pharmacology Journal*, **10**, 353-366. https://doi.org/10.13005/bpj/1116

[15] Pimkin, A., Makarchuk, G., Kondratenko, V., *et al*. (2018) Ensembling Neural Networks for Digital Pathology Images Classification and Segmentation. *International Conference Image Analysis and Recognition*, Póvoa de Varzim, 27-29 June 2018, 877-886. https://doi.org/10.1007/978-3-319-93000-8_100

[16] Marami, B., Prastawa, M., Chan, M., *et al*. (2018) Ensemble Network for Region Identification in Breast Histopathology Slides. *International Conference Image Analysis and Recognition*, Póvoa de Varzim, 27-29 June 2018, 861-868. https://doi.org/10.1007/978-3-319-93000-8_98

[17] Ojala, T., Pietikäinen, M. and Mäenpää, T. (2002) Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **24**, 971-987. https://doi.org/10.1109/TPAMI.2002.1017623