基于机器学习技术的非染色细胞凋亡检测分类新方法

冯婧文 撒昱*

(天津大学精仪学院生物医学工程系, 天津 300072)

摘要 细胞凋亡检测和分类在生物医学研究中具有重要意义。该研究目的是在我们多年研究的基础上建立了一种基于偏振衍射成像流式细胞检测系统和机器学习技术的凋亡检测新方法, 具有更高的时间效率和更好的应用前景。化学诱导K562和HL60细胞凋亡。通过荧光激活细胞分选仪结合荧光双染技术将细胞分为三个亚群: 健康细胞、早期凋亡细胞、晚期凋亡/坏死细胞。应用偏振衍射成像流式细胞检测技术采集每个亚群细胞的衍射图像。基于局部二值模式算法提取图像纹理特征生成训练和测试数据集, 研究了不同的分类算法, 建立凋亡分类模型。对模型性能和时间效率做出对比分析, 筛选出具有更高时间效率的模型。新方法可实现90%的分类准确度, 并在时间效率上具有优势。该研究成功开发了一种快速的无染色细胞凋亡检测新方法, 使检测后的细胞可直接用于后续实验。

关键词 细胞凋亡检测; 非染色标记; 衍射图像; 机器学习

A Stain-Free Apoptosis Detection and Classification Method Based on Machine Learning Technique

Feng Jingwen, Sa Yu*

(*School of Precision Instrument and Opto-Electronics Engineering, Tianjin University, Tianjin 300072, China*)

Abstract Cell apoptosis detection and classification are very important in biological and medical studies. In this study, we established an apoptosis detection and classification method based on the polarization diffraction imaging flow cytometry system and machine learning techniques, which has higher time efficiency and applicability comparing with the previous result. K562 and HL60 cells were induced to undergo apoptosis. The cells were sorted into three subpopulations (viable, early apoptotic and late apoptotic/necrotic cells) using fluorescence-activated cell sorter in combination with double fluorescent stain after the apoptosis induction, and then measured by polarization diffraction imaging flow cytometry for diffraction image acquisition. Different classification algorithms were explored. The model performance and efficiency were analyzed to obtain a high-efficiency model. The new method can achieve a high accuracy of 90% and has higher time efficiency. A fast stain-free apoptosis detection method was developed. Cells after measurement and classification can be directly used in further experimental studies.

Keywords apoptosis detection; stain-free; diffraction image; machine learning

细胞凋亡是一种由生理或病理性因素引起的, 在基因调控下完成的程序性细胞死亡过程, 近年来的研究显示细胞凋亡与多种疾病的发生相关[1-3]。识别细胞凋亡对于临床和基础医学研究具有重要意义。

基于凋亡的形态和生化特点, 产生了多种凋亡

收稿日期: 2019-02-27 接受日期: 2019-04-24
国家自然科学基金(批准号: 81171342)资助的课题
*通讯作者。Tel: 13642165356, E-mail: sayu@tju.edu.cn
Received: February 27, 2019 Accepted: April 24, 2019
This work was supported by the National Natural Science Foundation of China (Grant No.81171342)
*Corresponding author. Tel: +86-13642165356, E-mail: sayu@tju.edu.cn
网络出版时间: 2019-08-12 15:08:22 URL: http://kns.cnki.net/kcms/detail/31.2035.Q.20190812.1508.028.html

检测技术: (1)透射电子显微镜技术, (2)琼脂糖凝胶电泳技术, (3)脱氧核苷酸末端转移酶介导的dUTP缺口末端标记技术, (4)酶联免疫吸附技术, (5)检测凋亡相关细胞内源成分, 如细胞内Ca^{2+}浓度增加、凋亡相关酶类如caspase 3活化, 及凋亡相关基因的表达等等, (6)基于流式细胞技术和Annexin V-FITC/PI荧光双染的凋亡检测[4]。以上技术均需要对细胞做特殊处理, 使检测后的细胞不能用于后续的实验研究, 或影响结果的准确性。同时, 实验操作步骤较多、抗体造价高, 增加了上述检测方法的成本。因此, 一些学者开发了非染色标记的凋亡检测技术。基于延迟相衬显微镜和图像处理技术的非染色凋亡检测方法, 该方法准确度可以达到90%[5-6]。基于光晶体生物传感图像处理的非染色细胞凋亡检测方法, 用于分析药物的细胞毒性[7]。但是, 此类方法不能分离出凋亡诱导后不同状态的细胞。

为了弥补现有凋亡检测技术的不足, 在之前的研究中, 我们基于偏振衍射成像流式细胞检测和机器学习技术, 开发了一种非染色标记的凋亡检测方法, 可以区分存活细胞、早期凋亡细胞、晚期凋亡/坏死细胞, 在独立测试数据集上可以达到90%以上的准确度[8]。上述研究中通过灰度共生矩阵(grey-level co-occurrence matrix, GLCM)算法提取衍射图像的纹理特征。研究过程中, 我们通过对时间效率的分析发现, 纹理特征的提取是该非染色凋亡检测技术的关键限速环节。因此, 在本研究中我们主要考察了局部二值模式(local binary patterns, LBP)算法建立的纹理特征的分类效果。LBP是一种纹理非参数模型, 最早由Ojala等[9]提出, 其计算简单高效、分类效果好。近年来LBP算法不断得到改进和扩展, 被广泛地应用于计算机视觉和纹理分类的研究中。通过对衍射图像的分析和研究, 采用一致性LBP(uniform LBP)算法建立细胞凋亡衍射图像的特征数据集, 用于训练模型, 并对其结果做出评价, 同时与经过降维处理后的GLCM算法模型进行时间效率的比较。

1 材料与方法

1.1 细胞培养

人髓性白血病细胞系K562和人早幼粒细胞白血病细胞系HL60悬浮于含10%胎牛血清的RPMI-1640培养基中, 置于37 ℃含5%二氧化碳的恒温培养箱培养。

1.2 凋亡诱导和细胞分选

分别用20 microg/mL顺铂孵育K562细胞24 h和1 mmol/L的双氧水孵育HL60细胞2 h来建立细胞凋亡模型。通过流式细胞仪结合Annexin V-FITC/PI染色将凋亡诱导后的细胞标记为三个亚群: 存活细胞、早期凋亡细胞、晚期凋亡/坏死细胞并检测凋亡率, 确定诱导条件。凋亡诱导后通过Annexin V-PE和SYTOX® Green双染, 利用荧光激活细胞分选设备将三个亚群的细胞分离出来, 用于非染色细胞检测实验研究中的图像采集。

1.3 细胞衍射图像采集

基于本实验室自主开发的偏振衍射成像流式细胞检测系统, 分别采集两个细胞系中上述三个亚群细胞的衍射图像。每一个细胞样本分别获得一对衍射图像, 即s-偏振和p-偏振衍射图像, 如图1所示。

1.4 组织图像数据集

本研究中共使用了5 580对K562细胞衍射图像, 其中5 000对作为训练数据集, 包括2 000对存活细

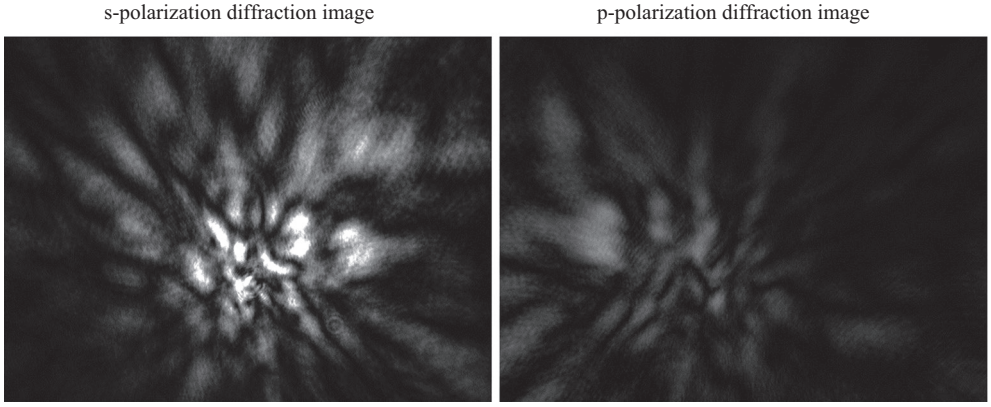


图1 K562细胞的一对衍射图像

Fig.1 A pair of diffraction image of K562 cell

胞、2 000对早期凋亡细胞、1 000对晚期凋亡细胞。580对作为测试数据集, 其中三个亚群的衍射图像对分别为250、250、80对。共使用了8 284对HL60细胞衍射图像, 7 000对作为训练数据集, 其中包括存活细胞3 500对、早期凋亡细胞1 000对、晚期凋亡细胞2 500对。1 284对作为测试数据集, 其中三个亚群衍射图像对分别为498、288、498对。

细胞实验中使用的试剂和设备、实验技术细节和结果、衍射成像流式细胞仪的结构和参数以及两个细胞系分选出的三种不同状态细胞的衍射图像示例详见我们之前的工作[8]。

1.5 特征提取算法

采用uniform LBP算法提取细胞衍射图像的纹理特征。LBP算子通过计算指定邻域内的像素点与观察像素点的差值来建立纹理特征模型, 差值大于0标记为1, 反之记为0。由此可以得到一个二值序列, 将该二值序列转换为10进制, 即为观察像素点的LBP值。设指定邻域内有N个像素, 中心像素点的LBP值可以表示为:

$$\mathrm{LBP}(x_c, y_c) = \sum_{n=0}^{N-1} \mathrm{S}(i_n - i_c) 2^n \tag{1}$$

其中(x_c, y_c)表示中心观察像素点, i_n为邻域内任一像素的灰度值, i_c为中心像素灰度值, S为计算的差值分配权重。

Uniform LBP, 采用圆形算子, 位于两个像素点之间的点通过插值计算, 通过统计二进制模式中从0到1和从1到0的跳变的次数有效的减少了LBP特征的维度, 同时实现了灰度与旋转的不变性[10]。

$$\mathrm{LBP}^{\mathrm{ri}}_{\mathrm{N,R}} = \begin{cases} \sum_{n=0}^{N-1} \mathrm{S}(i_n - i_c), & \mathrm{U}(\mathrm{LBP}_{\mathrm{N,R}}) \le 2 \\ \mathrm{N+1}, & \text{else} \end{cases} \tag{2}$$

$$\mathrm{U}(\mathrm{LBP}_{\mathrm{N,R}}) = \left|\mathrm{S}(i_{N-1} - i_c) - \mathrm{S}(i_0 - i_c)\right| + \sum_{n=1}^{N-1} \left|\mathrm{S}(i_n - i_c) - \mathrm{S}(i_{n-1} - i_c)\right| \tag{3}$$

N是指定邻域内像素点的个数，R是邻域的半径。

在本研究中我们将一幅衍射图像切割成64副均等的子图像, 分别计算每一幅子图像的LBP值生成新的LBP图像, 计算LBP子图像的直方图, 将其顺序连接作为特征向量用于衍射图像的分类。

1.6 分类算法

本研究中选择了线性核支持向量机(linear SVM)、径向基核支持向量机(RBF SVM)、逻辑斯蒂回归(logistic regression, LB)和随机森林(random forest, RF)分类算法分别构建模型, 以便筛选得到性能更好的模型。

1.7 模型评价方法

通过十折交叉验证的平均分类准确度(accuracy)和平均绝对误差(mean absolute error, MAE), 独立测试数据集上的查准率(precision)、查全率(recall)、F1因子(F1)、受试者工作曲线下面积(area under receiver operating characteristic curve, AUC)等参数评价分类模型的泛化能力和性能。

1.8 时间效率比较

在研究过程中, 我们发现该技术的主要限速环节是图像纹理特征的提取。因此, 在维持分类准确度不变的前提下, 通过限制特征提取的参数来实现时间效率的提升是一种有效的方法。

经过对前述研究中提取的GLCM二阶统计量的相关性分析, 保留了三个完全独立的二阶统计量分别是对别度、相关性和角二阶矩。GLCM模型的特征向量由288为降为144维。

对于基于LBP算法的纹理特征提取算法, 我们计算了观察像素点毗邻的8个像素点, 覆盖了其二阶邻域, 减少考察像素点会损失信息, 因此我们考虑通过减少子图像的划分来减少特征提取的时间和特征向量的维度, 但是分类准确度会随之下降, 因此仍然划分为64副子图像。

2 结果

2.1 基于LBP特征的分类模型在两个细胞系上的表现

基于我们构建的LBP特征提取算法, 一对细胞衍射图像可以生成1 280个特征向量, 特征向量的维度要高于GLCM算法生成的特征向量。利用该特征构建的分类模型的交叉验证结果和在独立测试数据集上的表现如表1所示。观察发现, 增加训练样本可以有效改善欠拟合问题, 在两个细胞系上相对表现最佳的是基于RBF SVM 分类模型。对独立样本的预测查准率可以达到90%, 召回率接近于查准率, 同时具有较高的F1因子和AUC值, 且平均绝对误差相对最低。

2.2 LBP-SVM(RBF)分类模型在测试集上的混淆矩阵

表2的混淆矩阵展示了LBP-SVM(RBF)分类模

表1 LBP特征模型在两个细胞系上的表现

Table 1 The performance of the model with LBP features

细胞系 Cell line	算法 Algorithm	准确率 Accuary	绝对平均误差 MAE	查准率 Precision	查全率 Recall	F因子 F1	受试者工作曲线下面积 AUC
K562	SVM (linear)	0.85	0.10	0.92±0.06	0.91±0.07	0.91±0.07	0.92±0.06
	SVM (RBF)	0.90	0.07	0.94±0.07	0.93±0.08	0.93±0.08	0.95±0.08
	RF	0.82	0.25	0.87±0.12	0.87±0.13	0.86±0.13	0.94±0.07
	LR	0.73	0.18	0.73±0.07	0.73±0.07	0.73±0.08	0.87±0.05
HL60	SVM (linear)	0.85	0.10	0.74±0.04	0.70±0.06	0.70±0.06	0.74±0.04
	SVM (RBF)	0.91	0.06	0.90±0.03	0.89±0.03	0.89±0.03	0.90±0.02
	RF	0.77	0.27	0.75±0.01	0.75±0.00	0.75±0.00	0.89±0.02
	LR	0.86	0.09	0.85±0.05	0.83±0.06	0.82±0.07	0.94±0.02

型在两个细胞系上分类的细节。观察发现, 对于早期凋亡细胞的识别和分类效果相对较差, 这与我们之前的研究结果是一致的。

2.3 模型学习曲线

在机器学习任务中, 我们希望获得更为准确的、推广性更好的模型。但是存在一个无可避免的问题, 即偏差–方差窘境。偏差是模型预测结果与真实值之间的差距, 方差是同一个模型在来自同一个分布整体的不同数据集上习得的差异。方差与偏差在一定意义上是矛盾的存在。训练数据集较小时, 模型不能很好的习得细节信息, 此时偏差较大, 但不同数据集的方差较小。为了减小偏差, 需要增加训练样本, 随着样本数的增加, 方差随之增大。因此, 为了获得更好的分类模型, 需要通过绘制学习曲线, 来对方差和偏差做出权衡, 考察要达到模型偏差和方差稳定所需要的训练样本数量。图2是基于LBP-SVM(RBF)分类模型的学习曲线, 可以看出在K562细胞系数据集上达到偏差和方差稳定需要4 000例样本, 而HL60细胞系的数据集上则需要5 600例样本。

2.4 模型时间效率比较

对降维处理后的GLCM模型和LBP模型的时间效率做了分析比较。如表3所示。在分类准确度保持90%不变的前提下, LBP算法在纹理特征提取的时间效率上要优于GLCM模型, 但是LBP模型由于特征向量的维度比较高, 分类模型构建的时间相对较长。各个模型在单个样本的预测时间上没有明显的差别。

3 讨论

本研究对我们之前的工作是一次有效的验证, 即细胞散射光生成的衍射图像中包含三维结构和折射率信息, 可以用于细胞凋亡的识别和分类。基于特征向量的不同, 可能获得不同的具有相对最佳效果的分类模型。结果显示, 基于LBP特征和SVM(RBF)分类算法的模型具有更好的分类效果。混淆矩阵显示, 模型对早期凋亡的细胞分类效果相对较差。分析可能的原因是, Annexin V检测早期细胞凋亡时, 是通过结合质膜上发生转位的磷脂酰丝氨酸, 该现象发生在细胞器皱缩和DNA被内源性核酸内切酶切割之前[11]。相对于凋亡后期和坏死细胞, 细胞在进入凋亡早期时形态学改变并不明显, 增加了以形态学和折射率变化为检测依据的衍射成像分

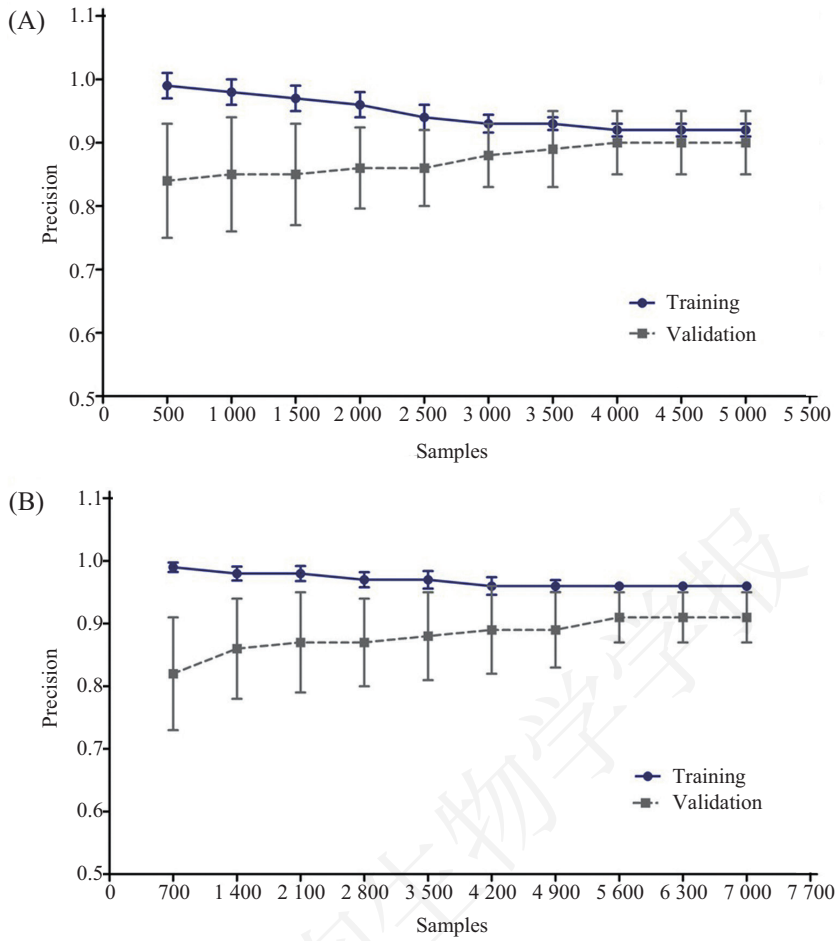
表2 LBP特征和SVM(RBF)组成的模型的混淆矩阵

Table 2 Confusion matrix of the model consisting of LBP features and SVM (RBF) on test dataset

Sample classes 样本类别	K562细胞个数 K562 cell number			HL60细胞个数 HL60 cell number		
	V	E	L/N	V	E	L/N
V	240	7	3	493	0	5
E	26	224	0	10	205	73
L/N	1	2	77	41	1	456

V: 活细胞; E: 早期凋亡细胞; L/N: 晚期凋亡或坏死细胞。

V: viable cell; E: early apoptotic cell; L/N: late apoptotic/necrotic cell.



A: 模型对K562细胞凋亡分类的学习曲线; B: 模型对HL60细胞凋亡分类的学习曲线。

A: the learning curve of the model on K562 cell apoptosis classification; B: the learning curve of the model on HL60 cell apoptosis classification.

图2 LBP-SVM(RBF)凋亡分类模型的学习曲线

Fig.2 Learning curve of the LBP-SVM(RBF) model

表3 模型时间效率比较

Table 3 Time efficiency of the models

特征 Feature	特征提取时间(s) Time for feature extraction (s)	特征维度 Teature dimension	算法 Algorithm	模型构建时间(s) Time for model building (s)	单个细胞识别时间(ms) Time for single cell identification (ms)
GLCM	4.54±0.60	144	SVM (linear)	29.72±1.33	0.015 0±0.001 0
			SVM (RBF)	12.91±1.64	0.014 0±0.000 6
			RF	5.86±0.26	0.013 0±0.001 0
			LR	17.12±0.11	0.013 0±0.001 0
LBP	1.12±0.03	1 280	SVM (linear)	47.21±0.54	0.013 0±0.001 0
			SVM (RBF)	39.42±0.09	0.013 0±0.001 0
			RF	9.34±0.45	0.013 0±0.000 6
			LR	171.37±0.40	0.014 0±0.001 0

表中数据为均值加减标准差。

The data in the table is the mean plus or minus standard deviation.

析技术的识别难度。时间效率分析的结果可以看出, LBP算法计算高效, 在独立测试集表现不变, 查准率接近90%的前提下, 其获得分类特征向量所需时间相对更短。在实际应用中, 每一组待检测细胞都需经历衍射图像采集和特征提取的过程, 这样一来, LBP模型在一定程度上有效地减少了实验研究

过程所需的时间, 提高了检测和分类的效率。同时可以看出, 由于生成的特征向量维度高, 分类模型构建的时间相对也较长。但是, 在基于衍射成像流式细胞检测系统的凋亡分类研究中, 针对一种细胞样本, 分类模型只需构建一次或有限几次, 因此, 实际应用中的效率较少受到模型构建时间的影响。与基于传统流式细胞检测技术的凋亡检测方法相比, 该技术不需对细胞样本进行染色或其他特殊处理, 简化了实验步骤, 最大程度上保持了细胞功能和结构的完整性。经该方法检测和分类的细胞回收后可直接用于后续的实验研究, 无需考虑荧光抗体标记和荧光染料带来的影响。在严格无菌的实验条件下, 回收的细胞可以继续培养, 用于药物筛选、功能实验等。由于该技术是一种流式细胞检测技术, 不能给出凋亡发生的组织学定位。 在凋亡检测中, 需要根据实验目的对检测技术进行选择。

我们基于细胞衍射图像的非染色凋亡检测分类技术仍在不断发展中, 随着后期研究的进一步深入, 硬件系统的进一步开发, 有望实现实时分选, 可以反向对已获得的实验结果做验证。值得指出, 本文报告的方法是对Annexin V/PI双染流式细胞凋亡分选技术的一项补充, 由于研究的依据是前者分选出的纯化样品, 受到实验条件的限制无法针对晚期凋亡细胞与坏死细胞进行非染色分类研究。随着荧光激活流式细胞分选技术的发展, 当可以分离得到相应的纯化样品后, 有望展开进一步的研究推动本文报道的非染色流式凋亡分类技术的发展。

4 结论

本研究成功建立了一种快速的基于偏振衍射成像流式细胞检测系统的非染色细胞凋亡识别分类的方法。该方法在时间效率上优于之前的基于GLCM纹理特征的方法, 在以快速检测为目标的实验研究中具有应用优势。

参考文献 (References)

1 Carson DA, Ribeiro JM. Apoptosis and disease. Lancet 1993; 341(8855): 1251-4.

2 Favaloro B, Allocati N, Graziano V, Di Ilio C, De Laurenzi V. Role of apoptosis in disease. Aging (Albany NY) 2012; 4(5): 330-49.

3 Feuerstein GZ, Young PR. Apoptosis in cardiac diseases: stress- and mitogen-activated signaling pathways. Cardiovasc Res 2000; 45(3): 560-9.

4 高超, 华子春. 细胞凋亡检测方法新进展. 中国细胞生物学学报(Gao Chao, Hua Zichun. Progress on detection of apoptosis. Chinese Journal of Cell Biology) 2011(5): 564-9.

5 Huh S, Ker DF Su H, Kanade T. Apoptosis detection for adherent cell populations in time-lapse phase-contrast microscopy images. Med Image Comput Assist Interv 2012; 15(Pt 1): 331-9.

6 Huh S, Kanade T. Apoptosis detection for non-adherent cells in time-lapse phase contrast microscopy. Med Image Comput Assist Interv 2013; 16(Pt 2): 59-66.

7 Chan LL, Gosangari SL, Watkin KL, Cunningham BT. A label-free photonic crystal biosensor imaging method for detection of cancer cell cytotoxicity and proliferation. Apoptosis 2007; 12(6): 1061-8.

8 Feng J, Feng T, Yang C, Wang W, Sa Y, Feng Y. Feasibility study of stain-free classification of cell apoptosis based on diffraction imaging flow cytometry and supervised machine learning techniques. Apoptosis 2018; 23(5/6): 290-8.

9 Ojala T, Pietikäinen M, Harwood D. A comparative study of texture measures with classification based on featured distributions. Pattern recogn 1996; 29(1): 51-9.

10 Ojala T, Pietikäinen M, Mäenpää T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE T Pattern Anal 2002; 24(7): 971-87.

11 Vermes I, Haanen C, Steffens-Nakken H, Reutellingsperger C. A novel assay for apoptosis flow cytometric detection of phosphatidylserine expression on early apoptotic cells using fluorescein labelled annexin V. J Immunol methods 1995; 184(1): 39-51.