EMITTER International Journal of Engineering Technology
Vol. 7, No. 1, June 2019
ISSN: 2443-1168

Nuclei Detection and Classification System Based On Speeded Up Robust Feature (SURF)

Neneng Nur Amalina[1], Kurniawan Nur Ramadhani[2], Febryanti Sthevanie[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
Jl. Telekomunikasi No. 01, Terusan Buah Batu, Bandung, (022) 7564108
[1]nenengna@students.telkomuniversity.ac.id, [2]kurniawannr@telkomuniversity.ac.id,
[3]sthevanie@telkomuniversity.ac.id

Abstract

Tumors contain a high degree of cellular heterogeneity. Various type of cells infiltrate the organs rapidly due to uncontrollable cell division and the evolution of those cells. The heterogeneous cell type and quantity in infiltrated organs determine the maglinancy level of tumor. We implemented Speeded Up Robust Feature (SURF) to analyze cells through their nuclei for better understanding of tumor by recognizing the inflammatory and non-inflammatory nuclei on histopathology image of colon cancer. We used three classifiers algorithm to classify the nuclei as performance comparison, those were k-Nearest Neighbor (k-NN), Random Forest (RF), and State Vector Machine (SVM) to ensure the consistency of SURF result. Based on the experimental result, the highest F1 score for nuclei detection was 0.722 using Determinant of Hessian (DoH) thresholding = 50 as parameter. For classification of nuclei, Random Forest algorithm produced F1 score of 0.527, the highest score as compared to the other classifier.

Keywords: nuclei, Speeded Up Robust Feature, histopathology image, feature extraction, classifier.

1. INTRODUCTION

Nucleus (plural: nuclei) is the most prominent organelle present in the middle of eukaryotic cells. It regulates all cell activity. Nucleus also controls the cell growth and reproduction in addition to store the cell's DNA which inherited through cell division[1]. Every type of cell has different nucleus characteristics. Tumors contain a high degree of cellular heterogeneity. Various type and number of cells infiltrate the organs rapidly due to uncontrollable cell division and the evolution of those cells. The

heterogeneous cell type and its quantity in infiltrated organs determine the level maglinancy of the tumor [2]. Therefore, the analysis of those various cells through their nuclei is needed for better understanding of tumor but also specify its proper treatment.

One way to explore the types of cell is to use multiple protein markers which can mark different cells in cancer tissues [3]. However, this way require deep biological understanding of tumors to identify informative markers and also it needs to be conducted in laboratory with adequate instrumentation [4]. An alternative and effective approach to identify cell nuclei is using morphological clue via image analysis based on image processing method [5].

Some image processing method had been implemented for nuclei detection, one of them is SIFT (Scale Invariant Feature Transform) [6]. However, SIFT requires a lot of space memory and high time computation, especially for big size images and complicated images because SIFT have to convolute the images with Gaussian filter in various scales [7]. In 2006, Herbert Bay presented the SURF (Speed Up Robust Feature) which is inspired from SIFT [8]. SURF is a further improvement of the SIFT, where SURF uses an integral image and Fast-Hessian filter box to speed up its time computation [9]. According to several studies comparing the performance of SIFT and SURF [10] [11] [12], SURF turns out to be more robust against different image transformations and works three times faster than SIFT.

In this paper, we proposed a system to detect the centroid location of nuclei and also classify the nuclei using SURF (Speeded Up Robust Feature) in histopathology images. We used SURF detector to detect the nuclei, and SURF descriptor to generate the feature extraction from the nuclei. Nuclei are classify into two types, inflammatory nuclei and non-inflammatory nuclei. We used three classifiers to classify the nuclei, those were k-Nearest Neighbor (k-NN), Random Forest (RF), and State Vector Machine (SVM) to ensure the consistency of SURF performances.

The organization of paper is as follows: related work cell detection and classification is given in section II, proposed method is explained in section III, section IV shows big picture and explanation about system design, section V shows the result from experiment and performance analysis of the proposed system, and conclution with recommendation for future works in section VI.

2. RELATED WORKS

There were many researches conducted that proposed detection system of certain types of cells or cellular subunits in microscopy images [13]. Most methods that had been used for cell and nucleus detection is segmentating and thresholding by morphological operation.

Veta *et al.* [14] developed an automated nuclei segmentation method which is relied on the direction of gradient to identify nuclei centroid in hematoxylin and eosin (H&E) stained breast cancer histopathology images.

There are four main step in this system: 1) pre-processing image with color unmixing and morphological operators, 2) marker-controlled watershed segmentation, 3) post-processing the false region and 4) merging of the results. The evaluation of two subset data revealed that the proposed method has good performance in both detection and segmentation accuracy. The mean sensitivity for the first subset was 0.875 and the second subset is 0.853.

Yuan et al. [15] develop a system to classi y cellular components into three categories: cancer, lymphocyte and stromal. This system used an SVM classifier to provide initial classification of cells based on morphological features, a kernel smoother to account for the neighbors of a cell, and a hierarchical model for incorporating global view of the image by multiresolution information flow.

F. Mualla, *et al.* [6] presented a novel system for cell detection in brightfield microscope image with no manual parameter tuning is needed in training or in detection. The system was designed invariance to cell size and orientation. SIFT keypoints with random forest was used to the cell and the image background. The result show that system yields a high detection accuracy and high invariance scores in reasonable computation time. Detection error from system was between approximately zero and 15.5%.

3. ORIGINALITY

This paper proposed Speeded Up Robust Feature (SURF) as algorithm to be implemented in system that can detect and classify various type of nuclei. H&E stained histology images of colorectal adenocarcinomas is used as image dataset in this system. Color deconvolution is applied on dataset before implementing SURF to separate the contribution of each specific staineds. K-Nearest Neighbor, random forest, and support vector machine is used and compared for classification process. The parameter of classifiers is tuned several times to find the most suitable classifier based on the dataset. The purpose of this system implementation is to be able to recognize nuclei found in infiltrated organs for better understanding of the cellular subtances, tumor maglinancy, and exploring its proper treatment.

4. SYSTEM DESIGN

This section will describe about system design of our proposed method. There are two main steps of the system as despicted in figure 1 below. The first stage is detection of nuclei and the second step is classification of nuclei.

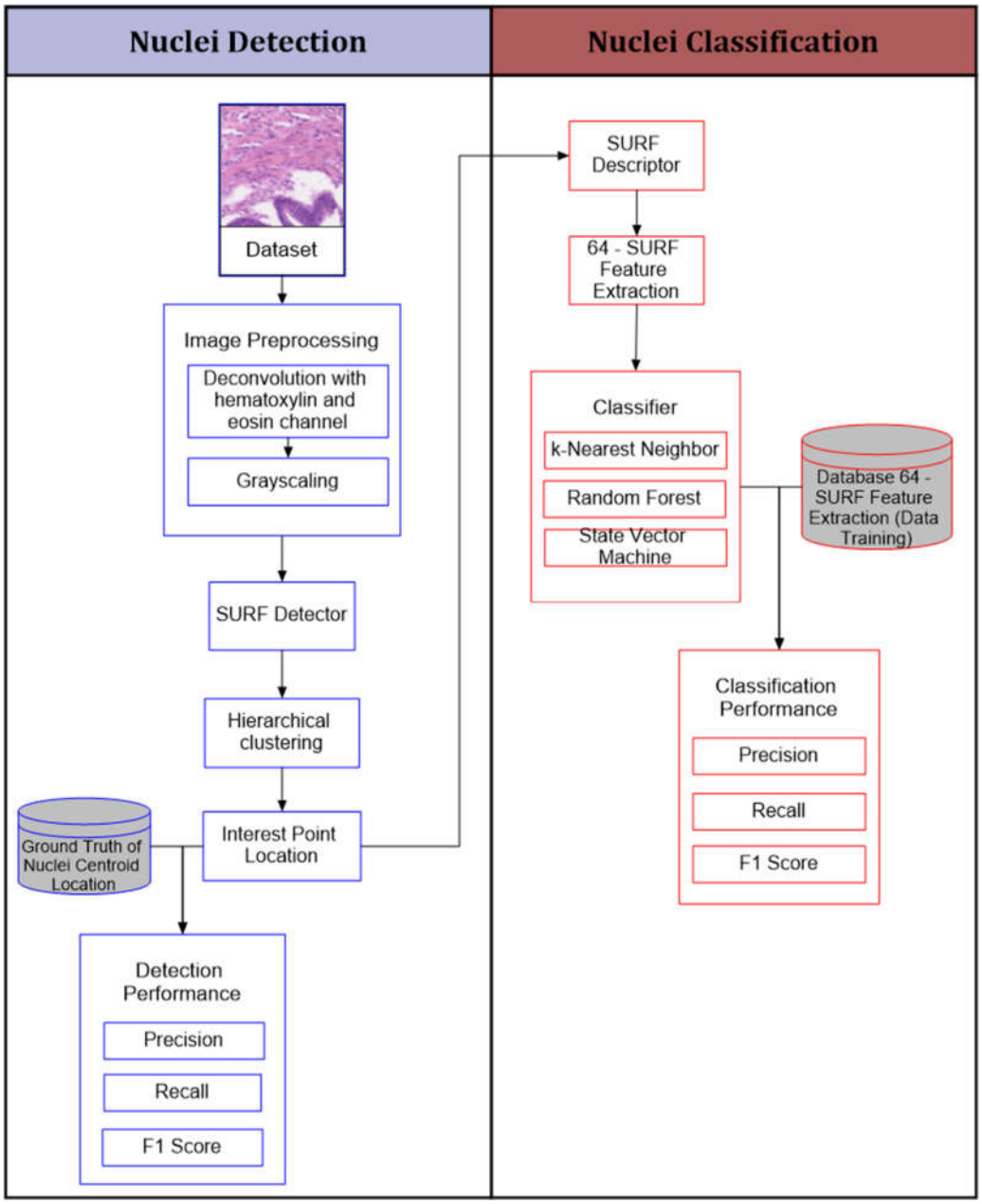

Figure 1. System Design of Nuclei Detection and Nuclei Classification

The purpose of nuclei detection is to detect all nuclei in an image by locating the position of their centroid regardless of the class label of the nuclei. This step will produce x- and y-coordinate of interest point and its Determinant of Hessian value. First, all of the image data have to be preprocessed by deconvoluting them with their original channel and converting them into grayscale images. To locate the position of nuclei centroid in an image, SURF detector is used to generated all of the interest points there. The interest points that had been generated are clustered with average hiearchical clustering by their dissimilarity. The clustered interest

point are obtained by cutting the dendrogram at a cutoff equal to 0.5. Clustered interest point signified the detection of nuclei around that region. Each centroid of clustered interest point is the estimation location for actual nuclei centroid.

In nuclei classification step, SURF descriptor will produce 64 dimension of SURF feature for every nuclei that had been detected from nuclei detection step. Those features will be classified based on sets of SURF feature from image data training. There are three classifiers that are used to classify the detected nuclei, those are k-Nearest Neighbor (k-NN), Random Forest (RF), and State Vector Machine (SVM). Due to the limited number of dataset and the imbalance numbers of nuclei present in an image, this paper classifies two type of nuclei, inflammatory nuclei and non-inflammatory nuclei (fibroblast, epithelial, and miscellaneous cells).

4.1 Dataset

Image data that used for dataset in this system consists of 70 H&E stained histology images of colorectal adenocarcinomas with 500x500 pixels in size. The image dataset is obtained from H&E stained histology images of colorectal adenocarcinomas Warwick University [3].

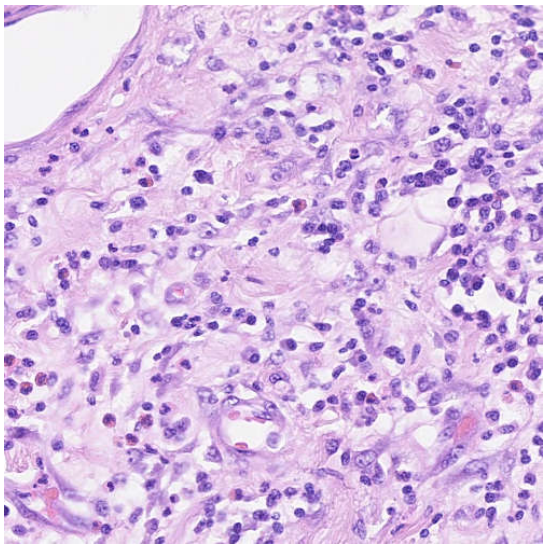

Figure 2. Sample of dataset

There are four types of nuclei in the dataset, such as inflammatory, fibroblast, epithelial, and miscellaneous nuclei. Four types of nuclei is grouped into two group of nuclei, those are inflammatory nuclei and non inflammatory nuclei (fibroblast nuclei, epihtelial nuclei, and miscellaneous nuclei). The dataset are divided into 60 training images and 10 testing images. We used 60 training images because the dimension of SURF descriptor is 64 for one nuclei. In 60 training images, there are 8000 nuclei consist of 4000 inflammatory nuclei and 4000 non-inflammatory nuclei.

While in 10 testing images, there are 5726 nuclei consist of 2839 inflammatory nuclei and 2887 non-inflammatory nuclei.

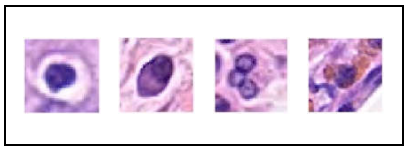

Figure 3. Set of Inflammatory Nuclei

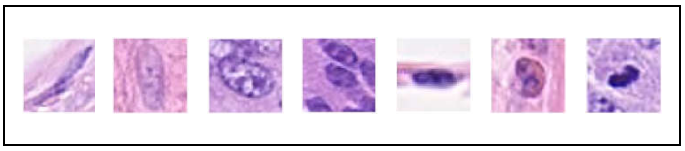

Figure 4. Set of Non-Inflammatory Nuclei

4.2 Image Preprocessing

There are two image preprocessing steps before implementing SURF detector to the system, deconvoluting and grayscaling. Preprocessing steps would enhance the quality of image data and also clarify the boundary between nuclei and background. To distinguish nuclei with other subtances in an image data, the image dataset should be deconvoluted with standart color matrix of hematoxylin channel since it is hematoxylin and eosin stained histology images. This way would be reduce grayscale image noises. The value of hematoxylin channel matrix is as following:

$$He = [0.6500286 \quad 0.704031 \quad 0.2860126] \tag{1}$$

4.3 SURF Detector

SURF Detector is used to locate interest points of an image. Interest points is obtained based on the determinant value of the Hessian matrix. Hessian matrix from image I in point X = (x, y) with scale σ is defined as:

$$H\ (X,\sigma) = \begin{bmatrix} L_{xx}(X,\sigma) & L_{xy}(X,\sigma) \\ L_{xy}(X,\sigma) & L_{yy}(X,\sigma) \end{bmatrix}, \tag{2}$$

where $L_{xx}(X,\sigma), L_{xy}(X,\sigma)$, and $L_{yy}(X,\sigma)$ are the convolution of Gaussian second order derivative with image I at the point X.

The second order Gaussian derivatives used for the Hessian matrix need to be discretized and cropped, so it can be used to calculate the approximated convolution effectively. The approximated and discretized filters are referred to D_{xx} as $L_{xx}(X,\sigma)$, D_{xy} as $L_{xy}(X,\sigma)$, and D_{yy} as $L_{yy}(X,\sigma)$.

$$det(H_{approx}) = D_{xx}D_{yy} \quad (0.9D_{xy})^2 \tag{3}$$

The Determinan of Hessian approximation value represents the blob response of an interest point in the image at position X = (x, y). This approximation value has to be performed on every octave and predefined scale then localized them using non-maximum suppression [16].

Before convoluting the image with Gaussian second order derivative, integral image of image I should be computed first. The entry of an integral image $I_{\Sigma}(X)$ represent the sum of all pixel in image I within a rectangular region from point $X = (0, 0)$ to point $X = (x, y)^T$ The using of sum from image I within a rectangle area may reduce its time computation [17].

$$I_{\Sigma}(X) = \sum_{i=0}^{i \le x} \sum_{j=0}^{j \le y} I(i, j) \tag{4}$$

4.4 SURF Descriptor

The purpose of a descriptor is to generate a unique and robust description of a feature. The SURF descriptor extract the description of a feature by determining the orientation and calculating the value of Haar Wavelet responses around the interest point. The interest point orientation is obtained by calculating the Haar Wavelet response from the horizontal and vertical directions around the interest point in the radius of 6s, where s is the scale from detected interest point. The wavelet response in the horizontal and vertical direction is refered to as dx and dy respectively. The SURF descriptor interest area is divided into 4 × 4 subareas that is described by the values of a wavelet response in the dx and dy, defined by the vector:

$$V = \left(\Sigma d_x, \Sigma d_y, \Sigma |d_x|, \Sigma \left|d_y\right|\right) \tag{5}$$

For each subarea a vector V is calculated, based on 5×5 samples. The descriptor for a interest point is the 16 vectors for the subareas concatenated. Finally the descriptor is normalized, to achieve invariance to contrast variations that will represent themselves as a linear scaling of the descriptor.

5. EXPERIMENT AND ANALYSIS

Experiment and analysis consist of two stages: nuclei detection performance evaluation and nuclei classification performance evaluation. Each stage is describe in section 5.1 and 5.2.

5.1 Nuclei Detection Performance Evaluation

A nucleus is successfully detected if there is at least one interest point found in the region within the radius of 6 pixels from the actual centroid of the nucleus. If there are multiple interest points found in those region, only the closest one to the annotated centroid is considered as true positive. Several experiments were performed with different Determinant of Hessian (DoH) threshold as parameter to find the best accuracy. The higher value of DoH threshold, the lower amount of interest point is generated by system. The following table is experimental result of nuclei detection in testing image data with DoH threshold = [0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100] as parameter.

Table 1. Experimetal result of nuclei detection on various DoH threshold

DoH Threshold	Run time (s)	Median Distance	Precision	Recall	F1 Score
0	145.568	**2.211**	0.261	**0.968**	0.405
10	89.221	2.329	0.505	0.924	0.65
20	64.524	2.487	0.579	0.885	0.698
30	52.925	2.661	0.625	0.841	0.715
40	47.628	2.783	0.653	0.803	0.718
50	43.06	2.945	0.68	0.776	**0.722**
60	39.567	3.163	0.7	0.745	0.719
70	36.807	3.359	0.717	0.72	0.715
80	33.874	3.726	0.727	0.682	0.701
90	31.488	4.171	0.735	0.648	0.685
100	**29.344**	4.655	**0.741**	0.619	0.67

Based on the table 1 above, the shortest running time is 29.344s with DoH threshold = 100. The shortest median distance of of interest point to its annotated centroid is 2.211 with DoH threshold = 0. While DoH threshold = 100 recorded the highest precision with 0.741, DoH threshold = 0 recorded the lowest precision with 0.261. Otherwise, the highest recall is 0.968 generated by DoH threshold = 0 and the lowest recall is 0.619 generated by DoH treshold = 100. According to precision and recall, the highest F1 score from the experiment is 0.722 reached by DoH threshold = 50.

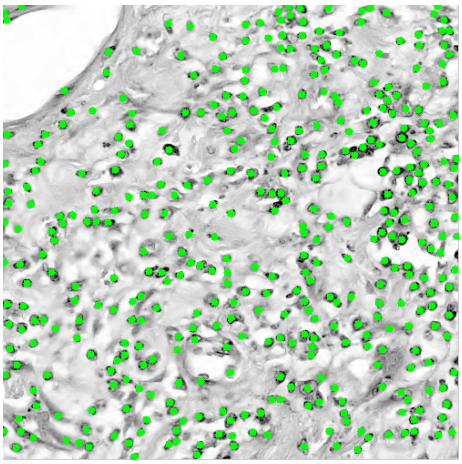

Figure 5. Annotated nuclei. Green dots is the location of nucleus centroid

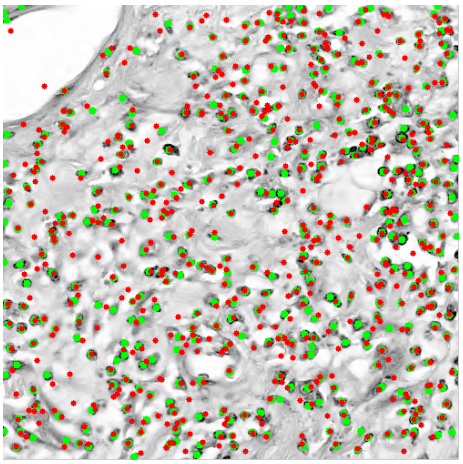

Figure 6. Result of nuclei detection. Red dots is the location of detected nuclei generated by system

From the figure 5 and figure 6 above, some nuclei on the side of image is fail to be detected. And it seemed that SURF overdetected the nuclei and detected the background as a nucleus. For the detection result, the implementation of SURF with DoH threshold as parameter is influencing the accuracy of the system because DoH threshold affect the number of generated interest point in an image. The higher value of DoH threshold, the higher precision is generated but it would be lowering recall. Otherwise, the lower value of DoH threshold, the higher recall is generated but precision would be getting down. Too much or too little interest point might lower the accuracy of the system. Figure 7 and figure 8 below is graphic representing about the influence of threshold DoH to the result of precision, recall, and F1 score.

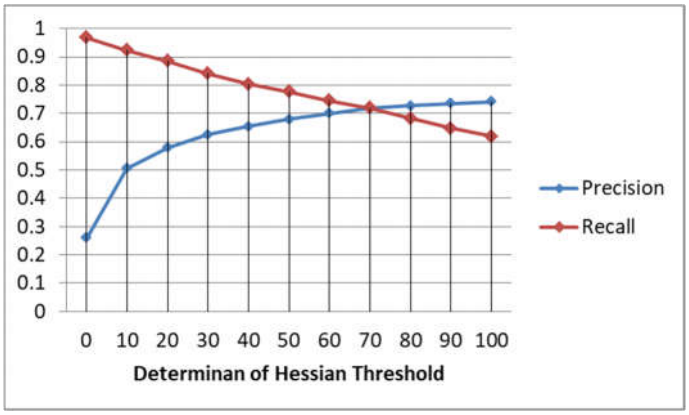

Figure 7. Precision and recall generated from different value of DoH threshold

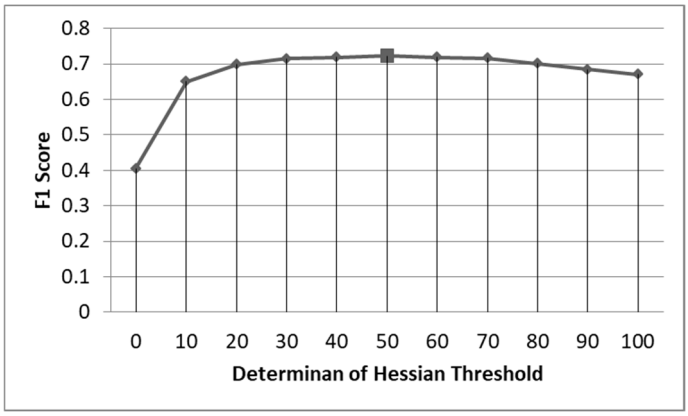

Figure 8. F1 score generated from different value of DoH Threshold

5.2 Nuclei Classification Performance Evaluation

We used three classifer to classify the 64 dimension SURF feature of the nuclei that had been detected from nuclei detection. The classifier were k-NN, random forest, and SVM. Every classifier has different characteristics. For k-NN, we observed the best value of k parameter started from k=1 to k=13 with interval 2. Meanwhile for random forest, we set the number of tree by 300, 400, and 500 and chose the best result from several attemption for every number of tree due to randomness of random forest. Table 2 shows the experimental result of nuclei classification using three classifier:

Table 2. Experimental result of Nuclei Classification by k-NN, random forest, and SVM

Classifier		Inflammatory Nuclei		Non-imflammatory Nuclei		Average F1-Score
		Precision	Recall	Precision	Recall	
k-NN	k = 1	0.412	0.532	0.579	0.475	0.499
	k = 3	0.472	0.538	0.546	0.479	0.509
	k = 5	0.466	0.537	0.548	0.48	0.508
	k = 7	0.47	0.54	0.555	0.482	0.512
	k = 9	0.467	0.539	0.559	0.482	0.512
	k = 11	0.47	0.54	0.561	0.483	**0.513**
	k = 13	0.468	0.54	0.557	0.482	0.512
Random Forest	T = 300	0.322	0.557	0.717	0.483	0.52
	T = 400	0.316	0.555	0.71	0.481	0.515
	T = 500	0.328	0.567	0.725	0.487	**0.527**
SVM		0.399	0.546	0.628	0.484	**0.514**

According to the table 2, the range of F1 score is from 0.499 to 0.527. For k-NN classifier, the highest F1 score is 0.513 using k=11, the highest F1 score for random forest is 0.527 with T=500, and SVM produces F1 score of

0.514. It shows that the performance of SURF is consistent using different type of classifier. Overall, random forest is found out to be the highest F1 score classifier among the other with a slightly difference. Random forest is more suitable in our system due to the existing dataset and value distribution of generated feature extraction. Even though k-NN and SVM produced better precision value in classifying inflammatory nuclei, random forest outperformed them in classifying non-inflammatory nuclei. Figure 9 below is a comparison of the highest F1 score from the three classifiers we used.

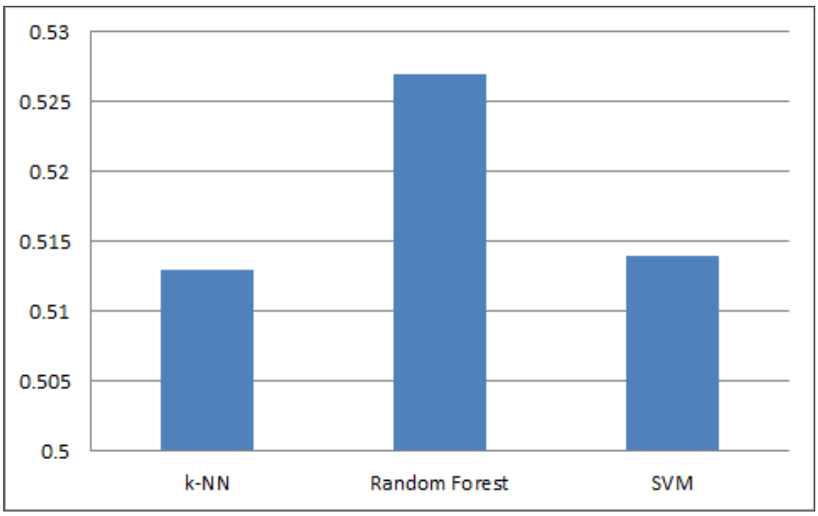

Figure 9. The higest F1 score of k-NN, random forest, and SVM

6. CONCLUSION

In this paper, we implemented Speeded Up Robust Feature (SURF) algorithm to build a detection system of nuclei on histopathology image of colon cancer. According to the analysis of experimental result, SURF is able to detect the various type of nuclei and produces F1 score of 0.722 with Determinant of Hessian (DoH) thresholding = 50 as parameter. However, based on the dataset that is used in our system, SURF descriptor needs more improvement to be able to classify the various kind of nuclei. The highest F1 score to classify the nuclei is only 0.527 using Random Forest as classifier with 500 as the number of tree. Almost half of the nuclei is fail to be classified whether it is inflammatory nuclei or non-inflammatory nuclei. For the future research, it may be better to use some combination of feature extraction or deep learning approaches like Convolutional Neural Network (CNN) to improve the accuracy of the system.

REFERENCES

[1] L. Lehninger, D. L. Nelson, and M. M. Cox, **Lehninger Principles of Biochemistry**, *Worth Pub*, 2002.

[2] H. Ochoa, K. Rao, and C. Ju´arez, **Cell Type Heterogeneity of Cytokeratin Expression in Complex Epithelia and Carcinomas as Demonstrated by Monoclonal Antibodies Specic for Cytokeratins**, Systemics *Cybernetics and Informatics*, Vol. 1, pp. 2–64, 2003.

[3] K. Sirinukunwattana, S. E. A. Raza, Y.-W. Tsang, D. R. Snead, I. A. Cree, and N. M. Rajpoot, **Locality Sensitive Deep Learning for Detection and Classication of Nuclei in Routine Colon Cancer Histology Images**. *IEEE Transaction on Medical Imaging*, 2016.

[4] G. N. van Muijen, D. J. Ruiter, W. W. Franke, T. Achtsttter, W. H. Haasnoot, M. Ponec, and S. O. Warnaar, **Cell Type Heterogeneity of Cytokeratin Expression in Complex Epithelia and Carcinomas as Demonstrated by Monoclonal Antibodies Speci c for Cytokeratins Nos. 4 and 13**, *Experimental Cell Research*, Vol. 162, No. 1, pp. 97–113, 1986.

[5] H. Irshad, A. Veillard, L. Roux, and D. Racoceanu, **Methods for nuclei detection, segmentation, and classi cation in digital histopathology: A review. current status and future potential**, *Biomedical Engineering, IEEE*, Vol. 7, pp. 97–114, 2014.

[6] F. Mualla, S. Scholl, B. Sommerfeldt, A. Maier, and J. Hornegger., **Automatic Cell Detection in Brightfield Microscope Images Using SIFT, Random Forests, and Hierarchical Clustering**, *International Journal of Image Processing (IJIP)*, Vol. 32, pp. 2274–2286, 2013.

[7] L.-C. Chiu, T.-S. Chang, J.-Y. Chen, and N. Y.-C. Chang, **Fast SIFT Design for Real-time Visual Feature Extraction**, *IEEE TRANSACTIONS ON IMAGE PROCESSING*, Vol. 22, pp. 3158–3167, 2013.

[8] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, **Speeded Up Robust Features (SURF)**, *Computer Vision and Image Understanding*, Vol. 110, pp. 346–359, 2007.

[9] P. Simard, L. Bottou, P. Haffner, and Y. LeCun, **Boxlets: A Fast Convolution Algorithm for Signal Processing and Neural Networks**, *NIPS*, 1998.

[10] E. Oyallon and J. Rabin, **An Analysis of the SURF Method**, *Image Processing On Line*, pp. 176-218, 2015.

[11] D. Mistry and A. Banerjee, **Comparison of Feature Detection and Matching Approaches: SIFT and SURF**.,*GRD Journals- Global Research and Development Journal for Engineering*, Vol. 2, 2017.

[12] N. Hamid, A. Yahya, R. B. Ahmad, and O. M. Al-Qershi, **A Comparison Between Using SIFT and SURF for Characteristic Region Based Image Steganography**, *IJCSI International Journal of Computer Science*, Vol. 9, pp. 110–116, 2012.

[13] Y. Xue and N. Ray, **Cell Detection with Deep Convolutional Neural Network and Compressed Sensing**, *CoRR*, 2017.

[14] M. Veta, P. J. van Diest, R. Kornegoor, A. Huisman, M. A. Viergever, and J. P. W. Pluim, **Automatic Nuclei Segmentation in H&E Stained Breast Cancer Histopathology Images**, *PLoS ONE*, Vol. 8, No. 7, p. e70221, 07 2013.

[15] Y. Yuan, H. Failmezger , O. M. Rueda, H. R. Ali, S. Gr¨af, S. -F. Chin, R. F. Schwarz, C. Curtis, M. J. Dunning, H. Bardwell et al., **Quantitative Image Analysis of Cellular Heterogeneity in Breast Tumors Complements Henomic Pro ling**, *Science translational medicine*, vol. 4, no. 157, pp. 157ra143–157ra143, 2012.

[16] M. Brown and D. Lowe, **Invariant Features from Interest Point Groups**, *BMVC*, 2002.

[17] L. Juan and O. Gwun, **A Comparison of SIFT, PCA_SIFT and SURF,** *International Journal of Image Processing (IJIP)*, Vol. 3, pp. 143–152, 2009.