

Original Article

Support system for pathologists and researchers

Takumi Ishikawa, Junko Takahashi, Mai Kasai, Takayuki Shiina, Yuka Iijima, Hiroshi Takemura, Hiroshi Mizoguchi, Takeshi Kuwata¹

Department of Mechanical Engineering, Tokyo University of Science, Tokyo, ¹Pathology Division, Cancer Center Hospital East, Chiba, Japan

E-mail: *Hiroshi Takemura - takemura@rs.tus.ac.jp

*Corresponding author

Received: 24 December 14

Accepted: 11 May 15

Published: 23 June 2015

This article may be cited as:

Ishikawa T, Takahashi J, Kasai M, Shiina T, Iijima Y, Takemura H, et al. Support system for pathologists and researchers. *J Pathol Inform* 2015;6:34.

Available FREE in open access from: <http://www.jpathinformatics.org/text.asp?2015/6/1/34/158911>

Copyright: © 2015 Ishikawa T. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Aims: In Japan, cancer is the most prevalent cause of death; the number of patients suffering from cancer is increasing. Hence, there is an increased burden on pathologists to make diagnoses. To reduce pathologists' burden, researchers have developed methods of auto-pathological diagnosis. However, virtual slides, which are created when glass slides are digitally scanned, saved in a unique format, and it is difficult for researchers to work on the virtual slides for developing their own image processing method. This paper presents the support system for pathologists and researchers who use auto-pathological diagnosis (P-SSD). Main purpose of P-SSD was to support both of pathologists and researchers. P-SSD consists of several sub-functions that make it easy not only for pathologists to screen pathological images, double-check their diagnoses, and reduce unimportant image data but also for researchers to develop and apply their original image-processing techniques to pathological images.

Methods: We originally developed P-SSD to support both pathologists and researchers developing auto-pathological diagnoses systems. Current version of P-SSD consists of five main functions as follows: (i) Loading virtual slides, (ii) making a supervised database, (iii) learning image features, (iv) detecting cancerous areas, (v) displaying results of detection. **Results:** P-SSD reduces computer memory size random access memory utilization and the processing time required to divide the virtual slides into the smaller-size images compared with other similar software. The maximum observed reduction in computer memory size and reduction in processing time is 97% and 99.94%, respectively. **Conclusions:** Unlike other vendor-developed software, P-SSD has interoperability and is capable of handling virtual slides in several formats. Therefore, P-SSD can support both of pathologists and researchers, and has many potential applications in both pathological diagnosis and research area.

Key words: Auto-pathological diagnosis, machine learning, OpenSlide library, virtual slide

Access this article online

Website:

www.jpathinformatics.org

DOI: 10.4103/2153-3539.158911

Quick Response Code:



INTRODUCTION

In Japan, cancer is the most prevalent cause of death with a mortality of approximately 400,000 people in 2012,

accounting for about 30% of all deaths.^[1] The number of patients suffering from cancer is increasing every year. On the other hand, the number of pathologists remains almost constant. The increasing shortage of pathologists

means that the burden of work that each pathologist shoulders becomes heavier, with shortened time of review of individual cases, often resulting in false diagnosis. To prevent false diagnoses and to screen for unimportant information, support systems for auto-pathological diagnosis (P-SSD) are needed.

Pathological virtual slides [Figure 1], which are created when glass slides are digitally scanned, have been increasingly used for both diagnosis and research. Many methods for an auto-pathological diagnosis system have been proposed.^[2] Most of them calculate features, learn the features, and then detect cancer using a computer. Otsu and Kurita^[3] proposed the high-order local auto-correlation (HLAC) image feature and Nosato *et al.*^[4] applied this feature to cancer detection. Takahashi *et al.*^[5] proposed a method that detects cancerous areas with HLAC and support vector machines (SVM).^[6] Ishibashi *et al.*^[7] applied a wavelet transformation to calculate the frequency features from pathological images. Doyle *et al.*^[8] proposed a multi-resolution Bayesian classifier with AdaBoost and applied that method to pathological images.

Although widely utilized as a replacement for glass slides in pathology, virtual slides pose major challenges for data storage, processing, and interoperability. As there is no common and no standard data format such as DICOM for virtual slides, each vendor defines their own data formats, analysis tools, viewers, and software libraries. Thus, researchers and pathologists need specific knowledge to work directly with these images. To handle the virtual slides, researchers usually convert the unique format of the virtual slides to a standard image format like JPEG and BMP using a specific viewer, and then save the converted images (CVT image). The CVT images still have a high resolution and often occupy several gigabytes of computer memory, which makes it difficult to open the saved CVT images. For example, a single

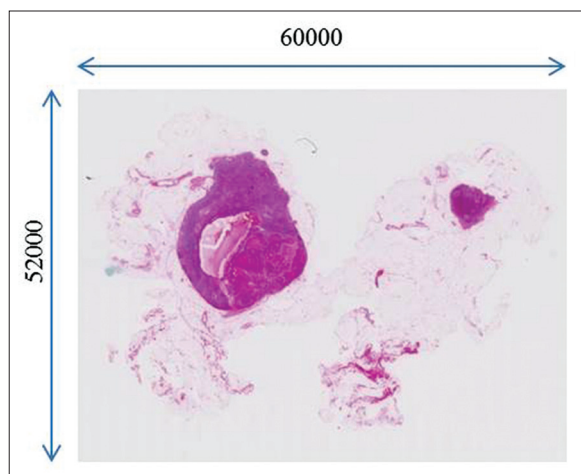


Figure 1: Virtual slide. The size of this virtual slide is 60,000 × 52,000 pixels

42 mm × 24 mm slide glass is about 60,000 × 52,000 pixels in size. To access the CVT image, the researchers need to save it after dividing it into smaller-sized images.

There are several open software packages for bio-image processing. NDP.view^[9] is the specific viewer software for “ndpi” format images, which is one of the most commonly used slide image format. With NDP.view, pathologists and researchers can convert and save a slide to a standard display format like JPEG or BMP using the screen capture function. The GNU image manipulation program (GIMP)^[10] implements several image-processing methods and can apply them to pathological images after NDP.view converts and saves the slide into several images. ImageJ^[11] also applies image-processing methods to pathological images. One of the most interesting characteristics of ImageJ is its plugin functionality. Deroulers *et al.*^[14] introduced “NDPITools” to ImageJ as a plugin. This plugin enables ImageJ to handle the “ndpi” format image. Using this plugin, ImageJ first converts a virtual slide image to a standard image format and then saves those images. Next, ImageJ applies image-processing methods to the pathological images. However, neither GIMP nor ImageJ deal directly with the virtual slide images. They first need to convert the virtual slide image into a software-compatible image format and save that image (including dividing the images into several parts) before applying the image-processing techniques. This procedure takes more time and requires more computer memory than directly applying image-processing techniques to the virtual slide images. There are other several commercially available software, which can handle different format, however, the researcher cannot apply their original image-processing technique and cannot access image data directly in their software. In addition, to make a supervised database function is very important for the researcher who works on the auto-pathological diagnosis research area. If there are huge pathological image database with the supervised cancerous area, the research of auto-pathological diagnosis probably improved tremendously.

To circumvent these issues, we proposed and developed the P-SSD.^[12] P-SSD consists of several sub-functions that make it easy not only for pathologists to screen pathological images and double-check their diagnoses but also for researchers to apply image-processing techniques to pathological images. Moreover, the standard pathological database collected by P-SSD is available to both pathologists and researchers. The original P-SSD application^[15] converts virtual slides into the standard format image such as JPEG and BMP, while the present version of P-SSD loads virtual slides using the OpenSlide library.

The rest of this paper is organized as follows. In section 2, we describe the concept of P-SSD and recent works. In

section 3, we discuss P-SSD and compare it with other software packages. Some brief conclusions are drawn in the last section.

METHODS

We originally developed P-SSD to support both pathologists and researchers developing auto-pathological diagnoses systems.^[13] Figure 2 illustrates the flow of P-SSD. P-SSD consists of five main functions as follows: (i) Loading virtual slides, (ii) making a supervised database, (iii) learning image features, (iv) detecting cancerous areas, (v) displaying results of detection. P-SSD can load various types of virtual slides and make a supervised database of the virtual slides. Compared to the original version presented in^[13] the present version of P-SSD loads virtual slides using the OpenSlide library.^[14] P-SSD is built with Visual Studio C++, Win32 API, OpenSlide library, and OpenCV library.

Loading virtual slides

There are many types of the virtual slide formats. Researchers need specific knowledge about the virtual slide type to handle these unique formats and researchers need to obtain new knowledge for each unique format. To handle the many unique formats of virtual slides, P-SSD uses the OpenSlide library. OpenSlide library is an open-access library and was specifically developed for loading and manipulating virtual slides in diverse vendor formats. It currently handles the following vendor formats:

- Aperio SVS

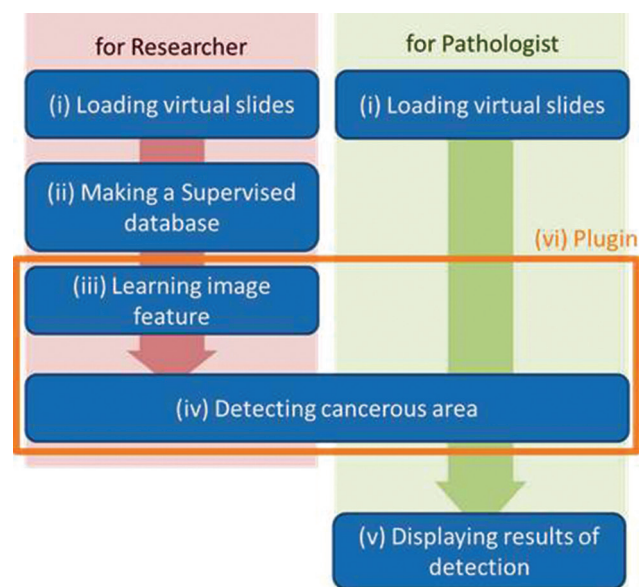


Figure 2: Concept of support system for pathological diagnosis (P-SSD). P-SSD consists of six main functions as follows: (i) Loading virtual slides, (ii) making a supervised database, (iii) learning image features, (iv) detecting cancerous areas, (v) displaying results of detection, (vi) plugin

- Aperio TIFF
- Hamamatsu VMS
- Hamamatsu VMU
- Hamamatsu NDPI
- Leica SCN
- MIRAX MRXS (“MIRAX”)
- Sakura SVSLIDE
- Trestle TIFF
- Generic tiled TIFF.

OpenSlide library enables not only pathologists to load the virtual slides but also researchers to apply image processing to a virtual slide without first saving the slide in a standard image format. Using P-SSD, it is easy for pathologists and researchers to work on the pathological images. Moreover, with P-SSD, both pathologists' time and computer resource requirements are less than other software.

Making a supervised database

Support system for pathological diagnosis can make a supervised database. Figure 3 shows an example of a virtual slide paint canvas in P-SSD. There is a map image and menu. The map image is on the lower right and the menu is on the upper left. In order to obtain a supervised database for researchers, the virtual slide images are supervised by pathologists who mark outlines of cancerous areas on the canvas, as shown in Figure 4. The size and number of supervised areas is displayed to the right of the supervised area. Pathologists are able to view the pathological image as a whole because the image is converted into a low-resolution image with P-SSD. Pathologists mark the outlines using a mouse or a pen tablet, and can change the colors of the outlines. Moreover, the pathologists can zoom in and out on a region of interest with a mouse wheel, changing the resolution. The maximum magnification depends on a scanning machine. After marking the cancerous area, the size of the area surrounded by the outlines is calculated and displayed on the screen. Next, the supervised data is output as a series of points in the XML file, and a set of

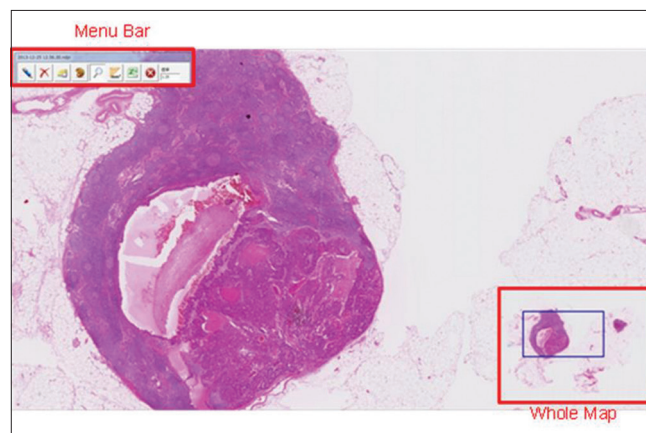


Figure 3: Paint canvas in the support system for pathological diagnosis. Upper-left corner is a menu bar and lower-right corner is a whole map

the supervised data and images are stored in the database. The space of the series of points in XML file depends on the supervised image resolution. The more high resolution, the more narrow space of points. Each supervised data file includes initial of the pathologist's name manually. P-SSD can handle multiple supervised data on the same image. Figure 5 shows the output file; lines 1–11 show the first cancerous area. The third and seventh lines indicate the X-coordinates of the outline, and the fourth and eighth lines indicate the Y-coordinate of the outline. Lines 12–14 show the second cancerous area. When that data are redisplayed, the series of points are linked to form outlines. By saving the supervised data in text format, as opposed to image format, the computer memory requirements are drastically reduced.

Learning image features

Support system for pathological diagnosis can learn the features of cancerous and noncancerous tissues. Numerous methods of cancer detection have been proposed; the most common cancer detection methods are implemented using machine learning. These methods calculate features from pathological images supervised by pathologists, and then build a learning model using machine learning methods. User can select supervised data files. If user selects multiple supervised data files with the same image, P-SSD builds a learning model using multiple supervised data files. Typically, the virtual slide is divided into image patches and the features are calculated from the smaller image patches. P-SSD can divide the pathological images into image patches. The image patches are rectangular images; the size of the image patches varies and is determined by the researcher. Features are calculated individually from every image patch.

The learning model is used in two ways. First, it is used for cancer detection in the non-supervised pathological virtual slides, as explained in the subsequent section.

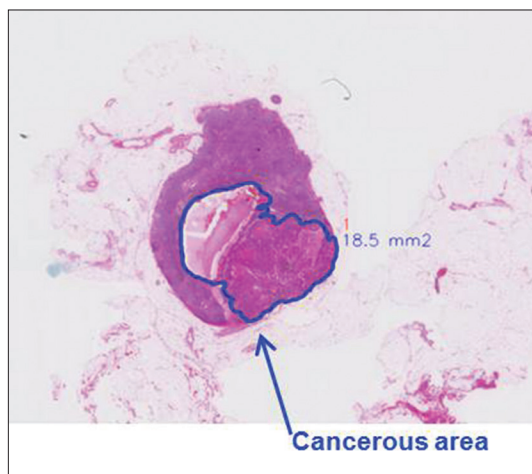


Figure 4: Supervised virtual slide. Blue mark is cancerous area and there is the size of cancerous area next to the blue mark

Second, the learning model is used to classify the database into cancerous and non-cancerous tissues. The labeled databases can then be used by students studying pathology and other people in the field.

The default features^[15] and machine learning methods are as follows:

- Features: Grayscale HLAC and Wavelet Feature
- Machine learning method: SVM (LibSVM).

The HLAC feature is a morphological feature and invariant to parallel translation. The wavelet feature is a frequency feature and quantifies the density of cell nuclei. Pathologists usually diagnose cancers considering the shape and the density of cell nuclei. By using HLAC and Wavelet features, P-SSD can diagnosis cancers like a pathologist. P-SSD collects all features into a single feature vector after calculating each feature as follows:

$$x_{HW} = \frac{x_H}{x_W} \tag{1}$$

where x_{HW} is the grayscale HLAC and Wavelet combination vector, x_H is the grayscale HLAC feature vector, and x_W is the Wavelet feature vector. SVM is one of the most common machines learning methods. P-SSD utilizes LibSVM,^[16] which is an open-source SVM program. Essentially, SVM is a binary linear classifier. However, it can classify features into multiple classes with kernel methods. There are many types of SVM and kernel methods. Table 1 shows the available SVMs and kernels in P-SSD. Finally, P-SSD can do cross-validation, which is a technique for assessing the accuracy of a statistical analysis independent of the particular data set.

```

1 <line number="1"> //First cancer area
2 <point>
3 <x>X1_VALUE</x> //x-coordinate of points
4 <y>Y1_VALUE</y> //y-coordinate of points
5 </point>
6 <point>
7 <x>X2_VALUE</x>
8 <y>Y2_VALUE</y>
9 </point>
10 ...
11 </line>
12 <line number="2"> //Second cancer area
13 ...
14 </line>
    
```

Figure 5: Example of the supervised data code. Lines 1–11 show the first cancerous area. The third and seventh lines indicate the X-coordinates of the outline, and the fourth and eighth lines indicate the Y-coordinate of the outline. Lines 12–14 show the second cancerous area

Detecting the cancerous area

Support system for pathological diagnosis can detect cancerous areas in tissue. After the virtual slide is divided into image patches, the features are calculated from the image patches as they were during the learning process. Then, using the supervised learning data, the cancerous area is detected in the unsupervised pathological images. Figure 6 shows an example P-SSD output, which is in XML form. In the XML file, the central coordinates of the image patches are shown in the third and fourth lines. The sixth line shows the detection results. If there is a cancerous cell in the region of interest, P-SSD outputs TRUE, otherwise, P-SSD outputs FALSE.

Displaying results of the detection stage

The system displays the detected cancerous area of a pathological image by painting in red those image patches that include cancer-based on the XML file. Each area is shown in a rectangular shape. Pathologists can double-check the cancerous area with the results of the image detection. Finally, P-SSD can screen and sort out the resulting images with detected cancer from ones without it beforehand so that pathologists can work with a set of images that likely contain cancerous tissue.

Software assessment

NDP.view divides the virtual slide into smaller-size images manually and NDPTool (ImageJ) divides them automatically. P-SSD does not need to divide the virtual slide into the normal images before P-SSD works on the virtual slide. In the assessment, the image memory size needed by P-SSD is compared with the size needed by NDP.view and NDPTool. In addition, the processing time of P-SSD, until the pathologists supervise, is compared with one of NDP.view and NDPTool. The size of a sample slide is 74,000 × 77,000 pixels, and its memory size is 450 MB. The PC display size is 1920 × 1080 pixels.

RESULTS AND DISCUSSION

Table 2 shows the number and the computer memory size of the images from whole virtual slide images in

```

1 <brock>
2 <point>
3 <x>X_1_VALUE</x> //x-coordinate
4 <y>Y_1_VALUE</y> //y-coordinate
5 </point>
6 <result>TRUE</result> //the result of detection
7 </brock>
8 ...
    
```

Figure 6: Example of the resulting XML code.The central coordinates of the image patches are shown in the third and fourth lines. The sixth line shows the detection results. If there is a cancerous cell in the region of interest, support system for pathological diagnosis (P-SSD) outputs TRUE, otherwise, P-SSD outputs FALSE

“ndpi” format. To save BMP format images, 2698 images of 1920 × 1080 pixel images are saved using NDP.view. To save TIFF format images, 16 images of 16,000 × 13,000 pixel images are saved using NDPTool. After added original memory size, the total computer memory size needed by NDP.view and NDPTool are about 16 GB and 740 MB. The computer memory size needed by P-SSD is only 450 MB, the original slide memory size. Compared with the computer memory size required by NDP.view and NDPTool to work on virtual slides, P-SSD saves 97% and 39% of the computer memory size respectively.

Using P-SSD, the time spent before the pathologists supervise can be reduced. Table 3 shows the processing time spent dividing the slide into the smaller-size image using NDP.view and NDPTool when one virtual slide is saved. The screen size was 1920 × 1080 pixels, and one virtual slide image was about 2336 screen-size images. NDP.view needs many operations to save one virtual slide into software-compatible images because the user saves 2336 screen-size images using the screen-capture function. Average time to save one screen-size image is 17 s, and then to save 2336 screen-size images, it takes approximately 13 h. NDPTool can divide the virtual slide automatically. However, it takes approximately 3 min to obtain TIFF images. With P-SSD, a user does not need to divide the virtual slide before processing. After NDP.view converts the virtual slide into software-computable images, each CVT image is marked using GIMP or other painting software to make white and black binary mask images. Table 4 shows the processing time of supervising

Table 1: Available SVM and kernel in P-SSD

SVM	C-SVM (basic), nu-SVC, one-class SVM, epsilon-SVR, nu-SVR
Kernel	Linear kernel, polynomial kernel, RBF kernel, sigmoid kernel

SVM: Support vector machines, P-SSD: Support system for pathological diagnosis, SVR: Support vector regression, RBF: Radial basic function, nu-SVC: nu-support vector classification, nu-SVR: nu-support vector regression

Table 2: The number and computer memory size of images after dividing the virtual slide into smaller-size images

	NDP.view	ImageJ	P-SSD
The number of images after dividing	2698	16	
The size of images after dividing	15.6 GB	290 MB	
The total size	16 GB	740 MB	450 MB

P-SSD: Support system for pathological diagnosis

Table 3: The processing time to divide the virtual slide into smaller-size images

	NDP.view	NDPTool
Time required for outputting and saving images per one virtual slide	12 h 44 min	3 min

one virtual slide in P-SSD and NDP.view. In the case of supervising whole virtual slides, the average time to supervise one virtual slide is about 2 1/2 min. Over 100 h are required for making one supervised virtual slide. In contrast, with P-SSD, pathologists mark virtual slides directly, and it is approximately 4 min to supervise one virtual slide. P-SSD saves 99.94% of the processing time to supervise one virtual slide compared with NDP.view.

By using the OpenSlide library, P-SSD can load virtual slides stored in diverse vendor formats. However, researchers and pathologists need P-SSD or other software to handle virtual slides of diverse vendor formats. It narrows the possibility of virtual slides that can be used because loading and manipulating virtual slides depend on these software packages. With the standard virtual slide format, virtual slides will be a more flexible tool. In future work, we will define the standard virtual slide formats. Furthermore, we will develop the plugin function to make it easy to apply feature calculation and learning techniques.

Example of using support system for pathological diagnosis

In this example, we learned HLAC and Wavelet feature vectors from nine virtual slides supervised by the pathologist, and the detected cancerous area from one virtual slide by calculating the feature vectors. P-SSD learned and detected cancer using SVMs after calculating HLAC features from 256 × 256 pathological image patches. The learning model is used only for cancer detection. The type of SVM and kernel are C-SVM and RBF, respectively. The size of each test images is

737,728 × 50,816 pixels. The resulting image, shown in Figure 7, depicts the detected cancerous area in blue. Each blue-painted area is 256 × 256 pixels (because the results of the cancer detection stage are output for each 256 × 256 pixel chunk). Learning time and detection time were 6349.868 s (705.54 s per slide) and 569.382 s, respectively.

CONCLUSION

This paper has presented the support system for pathologists and researchers using auto-pathological diagnosis. P-SSD consists of five functions: (i) Loading virtual slides, (ii) making a supervised database, (iii) learning image features, (iv) detecting cancerous areas, (v) showing results of detection. P-SSD enables pathologists and researchers to handle virtual slides in diverse vendor formats directly without converting them to other image formats. In order to obtain a supervised database for researchers, the virtual slides are supervised by pathologists who mark outlines of cancerous areas. P-SSD builds a leaning model by learning the feature vectors of cancer in the pathological images. Then, with the learning model, P-SSD detects cancerous areas by distinguishing cancerous and noncancerous tissue and shows the result of cancer detection to pathologists.

In future work, we are introducing a plugin function to P-SSD, which will make it easy to apply feature calculation and learning techniques. In addition to the plugin function, we will define standard virtual slide formats. With P-SSD, pathologists can double-check the pathological diagnoses and researchers can apply their techniques to pathological images more easily. P-SSD can support not only pathologists, but also researchers of P-SSD. Moreover, P-SSD reduces the time and computer resources required to save the virtual slide images in standard image formats.

Table 4: The processing time of supervising one virtual slide

	NDP.view	P-SSD
Time required for supervising one virtual slide	107 h 10 min	4 min

P-SSD: Support system for pathological diagnosis

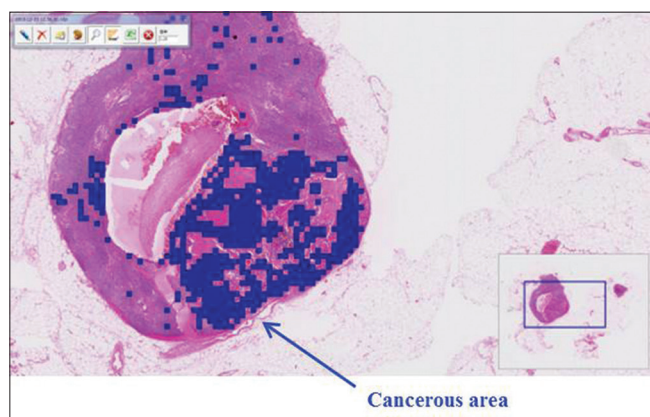


Figure 7: Result of detecting cancer area. Blue are detected cancerous areas

REFERENCES

1. National Cancer Center: Center for Cancer Control and Information Services Statistics; 2013. Available from: http://www.ganjocho.jp/pro/statistics/en/table_download.html. [Last cited on 2014 Oct 03].
2. Gurcan MN, Boucheron LE, Can A, Madabhushi A, Rajpoot NM, Yener B. Histopathological image analysis: A review. *IEEE Rev Biomed Eng* 2009;2:147-71.
3. Otsu N, Kurita T. A New Scheme for Practical, Flexible and Intelligent Vision Systems. In Proc. IAPR Workshop on Computer Vision; 1988.
4. Nosato H, Sakanashi H, Murakawa M, Higuchi T, Otsu N. Histopathological Diagnostic Support Technology using Higher-order Local Autocorrelation Features. In Bio-inspired Learning and Intelligent Systems for Security; 2009.
5. Takahashi J, Takemura H, Mizoguchi H, Kuwata T. System to Detect Abnormal Cells in Pathological Images using Higher-order Local Autocorrelation Features and Color Spaces. In: Proceedings of the 2012 International Conference of Information Science and Computer Applications; 2012.
6. Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;20:273-97.
7. Ishibashi Y, Hara A, Okayasu I, Kurihara K. Development of histopathological information database system which enables image retrieval using image features and text information. *Jpn Soc Comput Stat* 2011;24:3-21.

8. Doyle S, Feldman M, Tomaszewski J, Madabhushi A. A boosted Bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies. *IEEE Trans Biomed Eng* 2012;59:1205-18.
9. Hamamatsu.com: NDP.view2 U12388-01. Available from: <http://www.hamamatsu.com/jp/en/U12388-01.html>. [Last cited on 2014 Oct 03].
10. gimp.org: Gimp 2.8. Available from: <http://www.gimp.org>. [Last cited on 2014 Oct 03].
11. National Institutes of Health: ImageJ. Available from: <http://www.imagej.nih.gov/ij/>. [Last cited on 2014 Oct 03].
12. Deroulers C, Ameisen D, Badoual M, Gerin C, Granier A, Lartaud M. Analyzing huge pathology images with open source software. *Diagn Pathol* 2013;8:92.
13. Ishikawa T, Takahashi J, Takemura H, Mizoguchi H, Kuwata T. Beyond Supporting Pathological Diagnosis: Concept of Support System for Pathologist and Researcher. In: *Proceedings of the 9th Asian-Pacific Conference on Medical and Biological Engineering*; 2014.
14. Goode A, Gilbert B, Harkes J, Jukic D, Satyanarayanan M. OpenSlide: A vendor-neutral software foundation for digital pathology. *J Pathol Inform* 2013;4:27.
15. Ishikawa T, Takahashi J, Takemura H, Mizoguchi H, Kuwata T. Gastric Lymph Node Cancer Detection of Multiple Features Classifier for Pathology Diagnosis Support System. *Systems, Man, and Cybernetics (SMC)*, In: *Proceedings of the 2013 IEEE International Conference on*; 2013.
16. Chang CC, Lin CJ. LIBSVM: A library for support vector machines. *ACM Trans Intell Syst Technol* 2011;2:1-27.