Automated Quantification of Ki-67 on Gastric Epithelial Tissue based on Cell Nuclei Area Ratio

Cuantificación automatizada de Ki-67 en tejido epitelial gástrico basada en la razón de área de los núcleos celulares

Quantificação automatizada de Ki-67 em tecido epitelial gástrico com base na razão de área dos núcleos celulares

Austin Blanco-Solano[1], **Francisco Siles Canales**[1,2], **Warner Alpízar-Alpízar**[3]

Abstract

The objective was to develop an automated algorithm for the estimation of a protein (Ki-67) index based on cell nuclei area ratio of gastric epithelial tissue cells; for this purpose, digital histopathology images were used. An expert manually annotated each region of interest of the images. A proportion of Ki-67 positive and negative cells within that region was used to obtain the color distribution of the corresponding pixels. The histogram of each color distribution was modeled as a Gaussian and, later, thresholded for segmentation and classification. Finally, the Ki-67 index was estimated as the ratio between the segmented positive area of the nuclei divided by the total area of the positive and negative nuclei. The automated method has a strong correlation of 0.725 and a root mean square error of 0.293 when compared to the manual method, which gives certainty that the automated method can be used to analyze the proliferation rate. Furthermore, compared to manual classification, the presented method automatically classifies every image in the same Ki-67 category: low, intermediate, and high. Despite the small sample size, the utility of the presented method was demonstrated. However, the low number of scored images did not allow for thoroughly sampling the ranges of pixel values and intensities observed by pathologists, which will be addressed in future work.

Keywords: digital pathology; digital image processing; immunohistochemical analysis; pattern recognition; nuclei segmentation; Ki-67 quantification; gastric cancer; gastric cells

Austin Blanco-Solano, austin.blanco@ucr.ac.cr, https://orcid.org/0000-0002-5046-7886

Francisco Siles Canales, francisco.siles@ucr.ac.cr, https://orcid.org/0000-0002-6704-0600

Warner Alpízar-Alpízar, warner.alpizar@ucr.ac.cr, https://orcid.org/0000-0003-2842-4203

1 Electrical Engineering Department, and Postgraduate Studies in Electrical Engineering, Pattern Recognition and Intelligent Systems Laboratory (PRIS-Lab), Universidad de Costa Rica, San Pedro de Montes de Oca, Costa Rica.

2 Vice-presidency for Research, Surgery and Cancer Research Center (CICICA), Universidad de Costa Rica, San Pedro de Montes de Oca, Costa Rica.

3 Biochemistry Department, and Center for Research on Microscopic Structures (CIEMic), Universidad de Costa Rica, San Pedro de Montes de Oca, Costa Rica.

Resumen

El objetivo era desarrollar un algoritmo automatizado para la estimación de un índice de proteína (Ki-67), basado en la razón del área del núcleo celular de las células del tejido epitelial gástrico, utilizando imágenes digitales de histopatología. Cada región de interés de las imágenes fue anotada manualmente por un experto. Se utilizó una proporción de células Ki-67 positivas y negativas dentro de esa región para obtener la distribución de color de los píxeles correspondientes. El histograma de cada distribución de color se modeló como una Gaussiana y luego se estableció un umbral para la segmentación y clasificación. Finalmente, el índice Ki-67 se estimó como la relación entre el área positiva segmentada de los núcleos dividida por el área total de los núcleos positivos y negativos. El método automatizado tiene una fuerte correlación de 0,725 y un error cuadrático medio de 0,293, en comparación con el método manual, lo que da certeza de que el método automatizado se puede utilizar para analizar la tasa de proliferación. Además, en comparación con la clasificación manual, el método presentado clasifica automáticamente cada imagen en la misma categoría Ki-67: baja, intermedia y alta. A pesar del pequeño tamaño de la muestra, se demostró la utilidad del método presentado. Sin embargo, el bajo número de imágenes puntuadas no permitió muestrear completamente los rangos de valores de píxeles y las intensidades observadas por los patólogos, lo cual será abordado en un trabajo futuro.

Keywords: patología digital; procesamiento digital de imágenes; análisis inmunohistoquímico; reconocimiento de patrones; segmentación de núcleos; cuantificación Ki-67; cáncer gástrico; células gástricas

Resumo

[Objetivo] O objetivo foi desenvolver um algoritmo automatizado para a estimativa de um índice de proteína (Ki-67) baseado na razão da área do núcleo celular das células do tecido epitelial gástrico, utilizando imagens digitais de histopatologia. [Metodologia] Cada região de interesse nas imagens foi anotada manualmente por um especialista. Foi usada uma proporção de células Ki-67 positivas e negativas dentro dessa região para obter a distribuição de cores dos pixels correspondentes. O histograma de cada distribuição de cores foi modelado como uma Gaussiana e, em seguida, foi definido um limite para a segmentação e a classificação. Por fim, o índice Ki-67 foi estimado como a relação entre a área positiva segmentada dos núcleos dividida pela área total dos núcleos positivos e negativos. [Resultados] O método automatizado apresenta forte correlação de 0,725 e erro quadrático médio de 0,293, em relação ao método manual, o que dá certeza de que o método automatizado pode ser utilizado para analisar a taxa de proliferação. Além disso, em comparação com a classificação manual, o método apresentado classifica automaticamente cada imagem na mesma categoria Ki-67: baixa, média e alta. [Conclusões] Apesar do pequeno tamanho da amostra, foi demonstrada a utilidade do método apresentado. No entanto, o baixo número de imagens pontuadas não permitiu a amostragem completa das faixas de valores de pixels e intensidades observadas pelos patologistas, o que será abordado em trabalhos futuros.

Palavras-chave: Patologia digital; processamento digital de imagens; análise imuno-histoquímica; reconhecimento de padrões; segmentação de núcleos; quantificação de Ki-67; câncer gástrico; células gástricas.

Austin Blanco-Solano • Francisco Siles Canales • Warner Alpizar-Alpízar

Introduction and State-of-the-art

Although the incidence of gastric cancer has been decreasing, it is still the 5th most common and 7th most prevalent cancer worldwide (Bray *et al.*, 2018).

Through the mid-nineties, it was the primary cause of cancer-associated death worldwide, and, in the most recent years, it has been the 3rd most deadly, causing approximately 783,000 deaths per year (Rawla & Barsouk, 2019). Because *Helicobacter pylori* is an important agent in developing gastric cancer in humans (Rawla & Barsouk, 2019), there is an interest in the study of this bacterium. Infections from this bacterium induce proliferation of gastric epithelial cells with a risk of evolving into metaplasia, dysplasia, and invasive gastric cancer (Correa & Houghton, 2007). Pathologists use immunohistochemistry for diagnostics, which consists in the use of antibodies to identify antigens (also called markers) inside a tissue sample (Duraiyan *et al.*, 2012). One of these markers is called Ki-67, and it is highly associated with the proliferation of cells (Li *et al.*, 2014). Cells containing the antigen are stained, which allows them to quantify Ki-67 positive cells in the sample.

Quantifying the Ki-67 stained cells can be conducted using open-source programs and commercial software. These applications automate the process to some extent. For example, in the case of QuPath (QuPath, 2021), it is necessary to follow some steps: annotate the regions of interest (ROI); run the cell detection command indicating parameters such as type of tissue, pixel size, maximum and minimum nuclei area; train a classifier based on the annotations; finally, obtain the positive and negative cell count. On the other hand, ImageJ has a plugin called ImmunoRatio (Tuominen *et al.*, 2010), which calculates the Ki-67 index based on nuclei area ratio. In this case, it is also necessary for the user to indicate a series of parameters, for example, the segmentation thresholds for the nuclei (Yeo *et al.*, 2017). Other methods have been developed using techniques like k-means clustering for breast cancer histology images (Al-Lahham *et al.*, 2012) or for human nasopharyngeal carcinoma xenografts images (Shi *et al.*, 2016). In Barricelli *et al.* (2019), Bayesian classification trees are used, and in Xing et. al. (2014), the seed of the cells is located in order to classify them based on geometric descriptors, color intensity, cell morphology, and histogram intensities. However, these methods work for images whose staining is well marked, for example, with brown and light blue. There are also commercial programs such as the Aperio IHC Nuclear Image Analysis Tool (Aperio, 2007), which provides the option to annotate ROI and the complete automation of the whole quantification process. The problem with commercial programs is that they are neither free nor open source, reducing accessibility and customizability.

In the case of *H.pylori*-induced proliferation, immune cells infiltrate the gastric epithelium and start proliferating as part of the immune response to the pathogen. However, in many cases, pathologists are interested in the proliferation of epithelial cells and disregard these proliferating immune cells according to their location within the tissue. With the methods explained before, pathologists would be required to review and fix the automatic results by conducting the analysis, which negates the advantages of automation. Importantly, Ki-67 staining is used to measure proliferation in other tissues; therefore, a tool that can accurately

quantify proliferation and can be customized to different tissue structures is of great value. This work presents an automated Ki-67 quantification algorithm. It works for image sets with lower contrast in the staining and only requires the user to delimit the ROI (with an image analysis tool) where the quantification should take place (correctly marking the ROI by an expert can influence the accuracy of quantification).

Methodology

Modeling the Intensity of the Nuclei's Pixels

The ROI of the images in the dataset were manually annotated by an expert with QuPath and a red border line, as shown in Figure 2a, in order to delimit the processing to only the relevant part of the images. Afterward, a representative proportion of both Ki-67 positive and negative cells within the ROI of 10 images were manually annotated (labeled as images 1 to 10), for a total of 5702 *Ki-67 positive* and 6248 *Ki-67 negative* cells, see Table 1 (the ground truth data is available on request). For images 6 to 10, the manual counting of all the nuclei inside the ROI was performed, and the manual Ki-67 index was obtained. The first five images were only used to obtain samples of the color distributions of the nuclei but were not fully annotated. The annotations were done with an elliptical shape border, which allowed approximating the actual nuclei shape. Each ellipse is described by its center (*m,n*) and the lengths of its horizontal and vertical axes, *a* and *b*, respectively. These ellipses were used to draw binary masks with OpenCV that will be used to select pixels' samples to build a model for the classification. In order to remove any pixel not related to the specific type of nuclei, a *bitwise-AND* operation was carried out between the selected images in BGR and the masks generated before. The distribution of the pixel intensity of each nuclei type was analyzed via its histogram.

The resulting masked BGR image was converted to HSV and grayscale for the positive and negative nuclei, respectively. In the case of the positive nuclei, the V channel was used. Usually, the Ki-67 staining has good contrast between the color (brown) and the counter-color (light blue); so, at first, the H channel was the main candidate. However, in some of the images of the dataset, the counter-color, instead of being light blue, appeared to have a shade of purple and pink, while the color of the Ki-67 appeared to have a shade of auburn and purple. This was then reflected as low contrast in channel H images. On the other hand, the variation between color and counter-color was based to a greater extent on the darkness of the purple-brown tone, so the V channel was a better choice. Concerning the negative nuclei, the V channel was not so useful because it showed low contrast between the background and negative nuclei. Therefore, the grayscale was chosen.

Table 1

Manual Count of Ki-67 positive and Ki-67 negative nuclei.

Image/Ki-67	1	2	3	4	5	6	7	8	9	10	Total
Positive	231	261	261	297	166	629	434	1220	1001	1202	5702
Negative	133	172	192	156	185	731	866	1563	1119	1131	6248
Total	**364**	**433**	**453**	**453**	**351**	**1360**	**1300**	**2783**	**2120**	**2333**	**11950**

Note: derived from research.

Austin Blanco-Solano • Francisco Siles Canales • Warner Alpízar-Alpízar

Based on the histograms of the cells, for both the V channel and grayscale, Gaussian models were chosen to represent each distribution of the pixels $X \sim N(\bar{x}, s^2)$. The mean and standard deviation were calculated as in equation 1. In this equation, L is the amount of pixel intensities, h_i is the value for the given histogram bin, and i represents the bin number.

$$\bar{x} = \frac{1}{L}\sum_{i=0}^{L} i \cdot h_i \wedge s^2 = \frac{1}{L-1}\sum_{i=0}^{L-1} i \cdot (h_i - \bar{x})^2 \qquad (1)$$

The images with Ki-67 staining tend to have two types of quality problems: the uneven gathering of pigment on stained tissues and small pigment particles scattering around (Shi *et al.*, 2016). Because of this, the first step was to apply a Gaussian blur filter to the original BGR image in order to eliminate noise and smooth out the image. After filtering the image, it was converted from BGR to HSV, and the V channel was then segmented via thresholding. The segmentation was carried out as in equation 2, where T_1 and T_2 are the thresholds, α is a parameter that determines the number of standard deviations to consider, and f and g are the original and segmented images, respectively. This resulted in a mask for all the positive nuclei, and, finally, this mask was applied to the original BGR image to obtain the segmented Ki-67 positive nuclei.

$$Threshold = \begin{cases} T_1 = \bar{x} + \alpha \cdot s \\ T_2 = \bar{x} - \alpha \cdot s \end{cases} g = \begin{cases} 1, T_1 < f < T_2 \\ 0, otherwise \end{cases} \qquad (2)$$

Like *Ki-67 positive* nuclei, for the *Ki-67 negative* nuclei, the first step was to apply a Gaussian blur filter to eliminate noise and smooth the image; the difference is that, for the negative nuclei, the grayscale image was chosen. The segmentation was also carried out as in equation 2, which resulted in a mask that was applied to the BGR original image to obtain the segmented negative nuclei. However, this last image had residues of the Ki-67 stain pigment around the removed positive nuclei. This was removed by carrying out a *bitwise-AND* operation with the positive nuclei mask, but before this operation, the mask was dilated in order to cover more of the residues. Finally, an opening operation was applied to the image to remove any last particles.

Calculating the KI-67 Index

The nuclei in the training images have variations in morphology. Some have a shape like a circle, some to an ellipse, and some have a shape that cannot be so easily associated with a regular figure. Also, some nuclei are bigger than others and have different orientations. Because of this, we decided to calculate the Ki-67 index (K_i, see equation 3) as the ratio of the area of the segmented positive nuclei (A_{pos}) to the area of all the segmented nuclei ($A_{pos} + A_{neg}$), as studied in Barricelli *et al.* (2019), instead of counting each nucleus individually:

$$K_i = \frac{A_{pos}}{A_{pos} + A_{neg}} \qquad (3)$$

Analysis and Results

The parameters of the Gaussian model for the nuclei segmentation were calculated. For positive nuclei, the V channel mean was 121.341; the standard deviation, 23.898; and the α parameter, 1.6. For the negative nuclei, the grayscale mean was 145.506; the standard deviation, 15.600; and the α parameter, 1. The Q-Q plots in Figure 1 were used to determine whether the model selection introduced a substantial error. As can be observed, most of the points lie on the 45° line. The dotted curve seems to deviate

downwards, which indicates a slight skew to the right. The root-mean-square error (RMSE) for this fit is 0.440 for the positive nuclei and 0.430 for the negative nuclei. In a perfect Gaussian distribution, the mean and the median coincide; therefore, in order to quantify the amount of error the skew brings into the model, the difference between these metrics was calculated. The average difference for the positive nuclei was 1.45%, and for the negative nuclei, 0.45%.

Figure 1 shows the results of segmenting the positive and negative nuclei. As shown in subfigures 2a, 2b, and 2c, the automated quantifier segments only the nuclei inside the ROI.

Table 2 summarizes the Ki-67 indexes obtained for each validation image both by the algorithm and manual counting. The smallest error between indexes was 5.720%, while the highest was 22.330%, giving an average error of 13.159%. Figure 3 shows the correlation between the automated and manual Ki-67 index, which resulted in a Pearson correlation of 0.725, with an RMSE of 0.066.

Table 2
Comparison of the automated (A) and manual (M) Ki-67 index.

| Image	A	M	Abs(A-M)	(A-M)²
6	0.417	0.471	0.046	0.138
7	0.389	0.334	0.055	0.111
8	0.395	0.438	0.043	0.124
9	0.499	0.472	0.027	0.223
10	0.630	0.515	0.115	0.265
Average Differences			**0.057**	**0.172**
RMSE				**0.293**

Note: derived from research.

Discussion and Conclusions

This study developed an automated approach to estimate the Ki-67-index that does not require prior experience with cell counting and accepts images with annotated ROI. As shown in Figure 2, the proposed method can segment both the positive and negative nuclei that reside in the previously annotated ROI. Based on the area of the segmented nuclei, the automated method calculates the Ki-67 index with a strong Pearson correlation of 0.725 when compared to the manual method. This gives certainty that the automated method can be used to analyze an increment of stained cells.
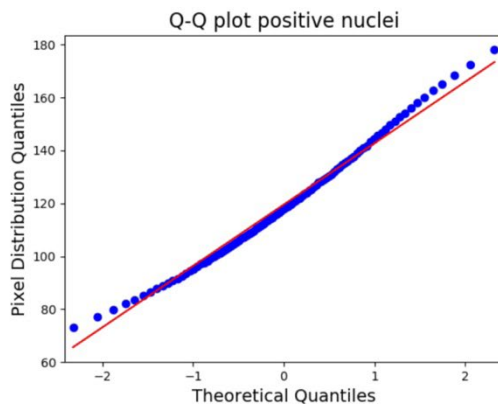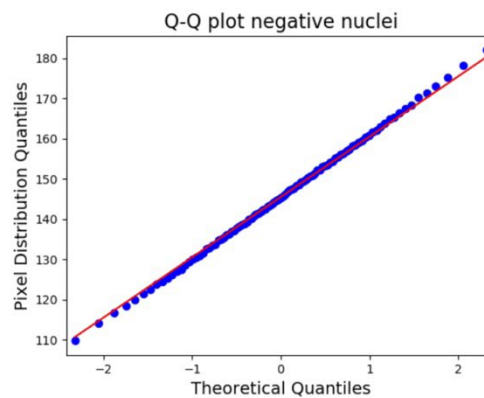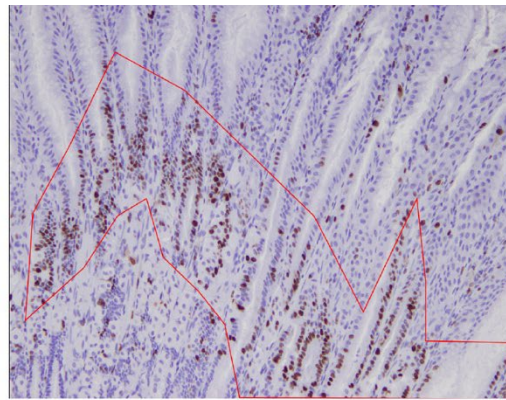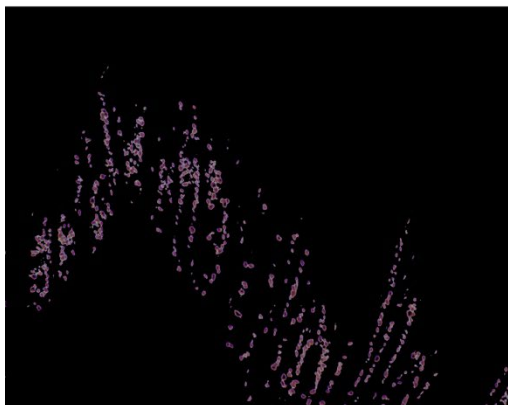

a: Positive nuclei   b: Negative nuclei

Figure 1. *Q-Q plot for the Gaussian model of the positive and negative nuclei.*
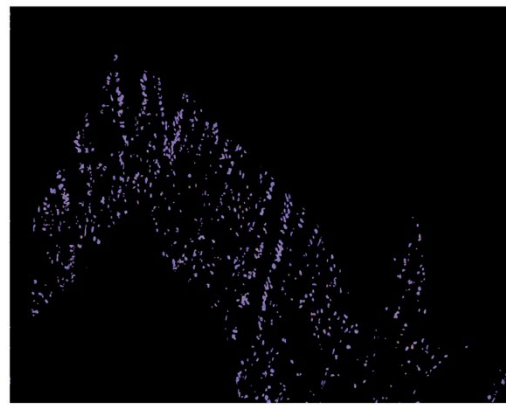*Note:* derived from research.

a. Annotated ROI

b. Positive nuclei

c. Negative nuclei

Figure 2. *Example of segmentation for the positive and negative nuclei.*
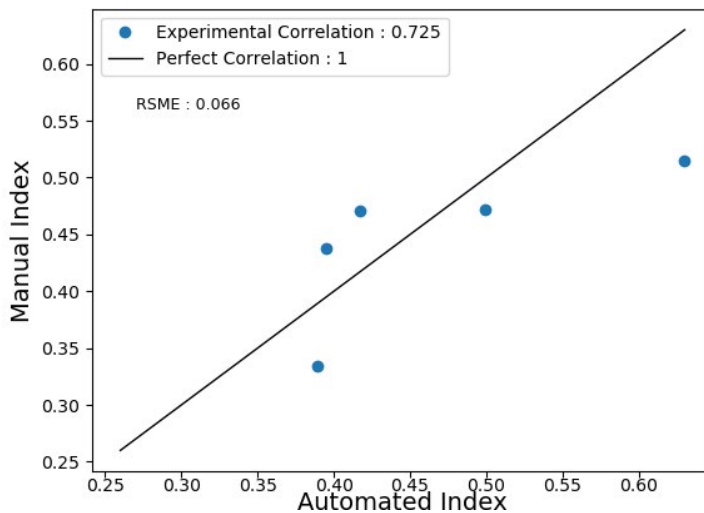*Note:* derived from research.

Figure 3. *Pearson correlation between the automated and manual Ki-67 index for images 6-10.*
Note: derived from research.

On the other hand, the Ki-67 indexes calculated by the automated method have an average difference of 0.057, an average square difference of 0.172, and a RMSE of 0.293, compared to the manual method, as shown in Table 2. It is equally important to evaluate both the error of the automated method and the accuracy of the insight these indexes provide. The St. Gallen International Consensus of Experts (Goldhirsch et al., 2011) recommends the categorization of the Ki-67 proliferation index into 3 groups: low (Ki-67 $\leq$ 15%), intermediate (15% < Ki-67 $\leq$ 30%), and high (Ki-67 > 30%). Based on these categories, the indexes calculated by our algorithm provide the same classification as manually determined by a trained expert. There are different error sources for this method. The Q-Q plots showed that the model selection introduced minor errors. In addition, an important parameter for these Gaussian models is α because of its direct relation with the index. The value corresponding to the α parameter for the positive nuclei was determined experimentally by means of a visual inspection of the segmentation of the V channel, similarly the α parameter for the negative nuclei was experimentally determined using the grayscale image. These parameters can be optimized by using a train/test/validation split. Despite the small sample size, we demonstrated the utility of our method. However, the low number of scored images did not allow us to fully sample the ranges of pixel values and intensities observed by pathologists. In future work, we will explore the use of kernel density estimation, optimize α, consider morphological features for segmentation, generate a new set of synthetic images, and subsequently access a larger dataset.

Conflict of Interest

The authors declare no competing interests.

Author Contribution Statement

All the authors declare that the final version of this paper was read and approved. The total contribution percentage for the conceptualization, preparation, and correction of this paper was as follows: A.B.S. 40%., F.S.C. 40% and W.A.A. 20%.

Data availability statement

The data supporting the results of this study will be made available by the corresponding author, F.S.C., upon reasonable request.

References

Al-Lahham, Heba Z., Alomari, Raja S., Hiary, Hazem. & Chaudhary, Vipin. (2012). Automating proliferation rate estimation from Ki-67 histology images. *SPIE Medical Imaging*, 8315. https://doi.org/10.1117/12.911009

Aperio Technologies, Inc. (2007). *IHC Nuclear Image Analysis User's Guide.* https://tmalab.jhmi.edu/aperiou/userguides/IHC_Nuclear.pdf

Barricelli, Barbara Rita., Casiraghi, Elena., Gliozzo, Jessica., Huber, Veronica., Leone, Biagio Eugenio., Rizzi, Alessandro. & Vergani,

Barbara. (2019). ki67 nuclei detection and ki67-indexestimation: a novel automatic approach based on human vision modeling. *BCM Bioinformaticss, 20*(733).

Bray, Freddie., Ferlay, Jacques., Soerjomataram, Isabelle., Siegel, Rebecca L., Torre, Lindsey A. & Jemal, Ahmedin. (2018). Global Cancer Statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *A Cancer Journal for Clinicians*, *68*(6). https://doi.org/10.3322/caac.21492

Correa, Pelayo & Houghton, Jeanmarie. (2007). Carcinogenesis of Helicobacter pylori. *Gastroentorology*, *133*(2), 659–672. https://doi.org/10.1053/j.gastro.2007.06.026

Duraiyan, Jeyapradha.; Govindarajan, Rajeshwar; Kaliyappan., Karunakaran. & Palanisamy, Murugesan. (2012). Applications of immunohistochemistry. *Journal of Pharmacy and Bioallied Sciences*, *4*(2), 307–309. https://doi.org/10.4103/0975-7406.100281

Goldhirsch, A., Wood, W. C., Coates, A. S., Gelber, R. D., Thürlimann, B. & Senn, H.J. (2011). Strategies for subtypes—dealing with the diversity of breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. *Annals of Oncology*, *12*(8), 1736–1747. https://doi.org/10.1093/annonc/mdr304

Li, Lian Tao., Jiang, Guan., Chen, Qian. & Zheng, Jun Nian. (2014). Ki67 is a promising molecular target in the diagnosis of cancer (Review). *Molecular Medicine Reports*, *11*(3), 1566-1572. https://doi.org/10.3892/mmr.2014.2914

Shi,Peng,., Jing, Zhong., Jinsheng, Hong., Rongfang, Huang., Kaijun, Wang. & Yunbin, Chen. (2016). Automated Ki-67 Quantification of Immunohistochemical Staining Image of Human Nasopharyngeal Carcinoma Xenografts. *Scientific Reports*, *6*(32127).

QuPath. (2021). *Cell classification - QuPath 0.3.0 documentation*. https://qupath.readthedocs.io/en/latest/docs/tutorials/cell_classification.html

Rawla, Prashanth. & Barsouk, Adam. (2019). Epidemiology of gastric cancer: global trends, risk factors and prevention. *Gastroenterology Review*, *14*(1), 26-38, https://doi.org/10.5114/pg.2018.80001

Tuominen, Vilppu J., Ruotoistenmäki, Sanna., Viitanen, Arttu., Jumppanen, Mervi. & Isola, Jorma. (2010). ImmunoRatio: a publicly available web application for quantitative image analysis of estrogen receptor (ER), progesterone receptor (PR), and Ki-67. *Breast Cancer Research*, *12*(R56).

Xing, Fuyong., Su, Hai., Neltner, Janna., & Yang, Lin. (2014). Automatic Ki-67 Counting Using Robust Cell Detection and Online Dictionary Learning. *IEEE Transactions on Biomedical Engineering*, *61*(3), 859-870. https://doi.org/10.1109/TBME.2013.2291703

Yeo, Min-Kyung., Kim, Hee Eun., Kim, Sung Hun., Chae, Byung Joo., Song, Byung Joo. & Lee, Ahwon. (2017). Clinical usefulness of the free web-based image analysis application ImmunoRatio for assessment of Ki-67 labelling index in breast cancer. *Journal of Clinical Pathology*, *70*(8), 715-719. http://dx.doi.org/10.1136/jclinpath-2016-204162