Overcoming an Annotation Hurdle: Digitizing Pen Annotations from Whole Slide Images

Peter J. Schüffler*, Dig Vijay Kumar Yarlagadda*, Chad Vanderbilt, Thomas J. Fuchs

Department of Pathology, Memorial Sloan Kettering Cancer Center, New York City, NY, USA
*Both contributed equally

Abstract

Background: The development of artificial intelligence (AI) in pathology frequently relies on digitally annotated whole slide images (WSI). The creation of these annotations – manually drawn by pathologists in digital slide viewers – is time consuming and expensive. At the same time, pathologists routinely annotate glass slides with a pen to outline cancerous regions, for example, for molecular assessment of the tissue. These pen annotations are currently considered artifacts and excluded from computational modeling. **Methods:** We propose a novel method to segment and fill hand-drawn pen annotations and convert them into a digital format to make them accessible for computational models. Our method is implemented in Python as an open source, publicly available software tool. **Results:** Our method is able to extract pen annotations from WSI and save them as annotation masks. On a data set of 319 WSI with pen markers, we validate our algorithm segmenting the annotations with an overall Dice metric of 0.942, Precision of 0.955, and Recall of 0.943. Processing all images takes 15 min in contrast to 5 h manual digital annotation time. Further, the approach is robust against text annotations. **Conclusions:** We envision that our method can take advantage of already pen-annotated slides in scenarios in which the annotations would be helpful for training computational models. We conclude that, considering the large archives of many pathology departments that are currently being digitized, our method will help to collect large numbers of training samples from those data.

Keywords: Computational pathology, digital pathology, pen annotations, training data generation

INTRODUCTION

Algorithms in computational pathology can be trained with the help of annotated image data sets. In some scenarios, the knowledge of located tumor regions on an image is beneficial, as the models are designed to learn from the differences between cancerous tissue and surrounding normal tissue.[1-5] A large part of the corresponding pipelines for pathology AI development is therefore the creation of annotated data sets on scanned WSI such that cancerous regions are digitally accessible. Annotations are usually acquired with the help of pathologists, drawing on WSI with digital tools on a computer screen. Generating those annotated data sets can constitute a bottleneck since it is time consuming, cumbersome, and error-prone, depending on the level of granularity of the annotations.

At the same time, many glass slides are already physically annotated by pathologists with a pen to outline tumor regions or other regions of interest. As an example, glass slides are commonly annotated for molecular assessment to outline tumor regions to be sampled for genetic analysis and sequencing. Tissue from the original paraffin-embedded specimen can hence be sampled from the same region that the pathologist indicated on the glass slide after inspecting the slide. However, these pen annotations are hand-drawn on glass and not *ad hoc* utilizable by a digital algorithm. They have yet to be digitized.

With this work, we present a novel method to extract pen annotations from WSI to be able to utilize them for downstream digital processing. As illustrated in Figure 1 with a scanned pen annotation on a WSI (left), our method extracts binary digital

Address for correspondence: Dr. Peter J. Schüffler,
Memorial Sloan Kettering Cancer Center, 417E 68th St,
New York City, NY 10065, USA.
E-mail: schueffp@mskcc.org

How to cite this article: Schüffler PJ, Yarlagadda DV, Vanderbilt C, Fuchs TJ. Overcoming an annotation hurdle: Digitizing pen annotations from whole slide images. J Pathol Inform 2021;12:9.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2021/12/1/9/310041

Access this article online

Quick Response Code:

Website:
www.jpathinformatics.org

DOI:
10.4103/jpi.jpi_85_20

masks of the outlined regions (middle, blue mask). Hence, it allows us to take advantage of the annotations which have already been made from trained pathologists, reducing the need to collect new, manually drawn annotations, such as shown in Figure 1, right (red manually drawn digital annotation). Considering the plethora of archived image data in pathology departments, our method enables to access thousands of such hand-drawn annotations, making these annotations available for computational pathology for the first time.

Currently, pen annotations on digital WSI are usually considered artifacts, disturbing downstream computational analysis as they cover or stain the underlying tissue. Therefore, research exists aiming to automatically detect and exclude pen annotations on WSI from analysis along with tissue folds, out-of-focus areas, air bubbles, and other artifacts.[6-8] Instead, we propose to make use of the already annotated glass slides and digitize the inhibited information to make it accessible to computational algorithms.

Our open-source code is available online at https://github.com/MSKCC-Computational-Pathology/PenAnnotationExtractor and can be used by other researchers to overcome the bottleneck of manually annotating digital slides.

METHODS

Pen annotation extraction

The annotation extractor is implemented as a command line script in Python 3. Its input is a folder containing thumbnail images of all WSI to be processed. We extracted the thumbnails stored in WSI prior to processing using the freely available library OpenSlide.[9] The output is a different folder with detected pen annotation masks for those images, each mask with the same dimensions as the corresponding thumbnail image. Seven processing steps compose the workflow for every thumbnail image in the input folder, as illustrated in Figure 2.

In step 1, a Gaussian blur filter with radius 3 is applied to the thumbnail image to reduce unspecific noise. In step 2, the blurred image is converted to the HSV (Hue, Saturation, Value) color space. We use the HSV color space as we found the RGB color space not to be robust enough to detect all variations introduced during staining and scanning. Further, HSV is better suitable to separate the markers by addressing the raw luminance values. The HSV image is used in step 3 to mask the tissue with H&E-related color thresholds. Pixel values between (135, 10, 30) and (170, 255, 255) are considered tissue without pen.

In step 4, pen-stroke masks are extracted from the HSV image based on pen color-related thresholds. Our data set comprises three pen colors, black, blue, and green. Pixel values between (0, 0, 0) and (180, 255, 125) are considered to originate from black pen. Pixel values between (100, 125, 30) and (130, 255, 255) are considered to originate from blue pen. And pixel values between (40, 125, 30) and (70, 255, 255) are considered to originate from green pen. These HSV values describe a spectrum of the corresponding colors and have worked well for us to capture the pen annotated pixels. As we do not differentiate between the pen colors, the three individual color masks are joined to the overall pen mask. Note that our method can be extended to other pen colors by including their specific thresholds.

To close gaps in the annotated pen contours, a morphologic dilation with a circular kernel is employed on the overall pen mask. The dilation thickens the contours of the pen by the given kernel size and thus closes holes in the lines. This step is needed to account for thin pen lines and for small gaps in the drawn lines, e.g., at *almost* closed ends of a circle. The larger the gaps are, the larger the kernel size has to be in order to close the shape. We run our algorithm in four rounds with an increasing kernel size of 5, 10, 15, and 20 pixels. In each round, pen annotations with too large gaps will result in empty masks (as the closed contour in the next step cannot be found), and we subject those images to the next run with larger kernel size.

In step 5, the dilated mask is subject to contour extraction and filling.[10] To reduce noise in the filled contours, components smaller than 3000 pixels are filtered. This threshold was chosen as it worked best on our data set by filtering small regions such as unrelated pixels, small contours, and text regions while letting tissue annotations pass. However, we propose to explore variable filter sizes based on thumbnail dimension and resolution in future work. The resulting mask is then subtracted in step 6 from the filled contour mask to preserve only the inner region.
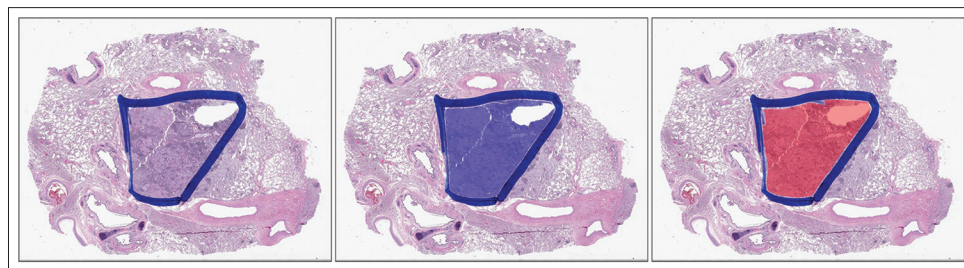

Figure 1: Example of a digitized pen annotation. Left: The original glass slide, manually annotated by a pathologist with a blue pen on the glass slide. Middle: Automatically segmented annotated region (blue) with our procedure based on the pen and tissue region. Right: For comparison: manually digitally annotated region (red) with a digital tool by a pathologist. This manual digital annotation is time consuming and redundant
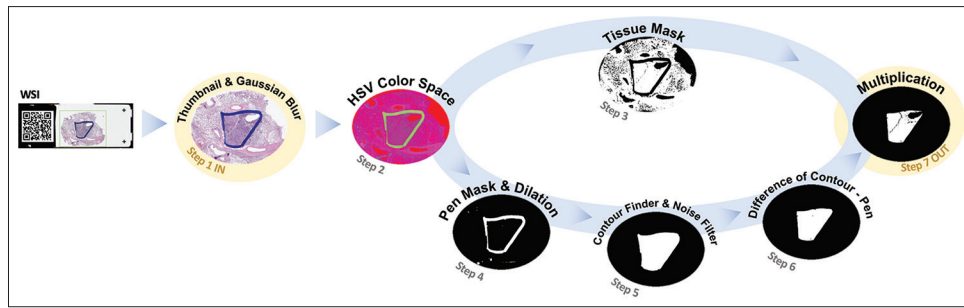
Figure 2: Annotation extraction pipeline. Step 1: From a WSI, a thumbnail is extracted as input on which a Gaussian filter is applied. Step 2: The blurred thumbnail is converted to HSV. Step 3: The tissue is separated from the background. Step 4: From the HSV image, the pixels of the pen colors are separated from the rest and dilated to close small gaps. Step 5: A contour finder fills closed contours identifying the "inner" regions. Then, noise such as small regions are filtered based on size. Step 6: The pen mask is subtracted from the contour mask to obtain the content of the annotated region only. Step 6: The final output is created by multiplying the tissue mask with the annotation mask

In step 6, the inner region mask is multiplied with the tissue mask to exclude background regions that are not tissue. The noise filter is applied again to remove small regions introduced at the annotation mask generation, resulting in the final mask of the pen annotated region.

Note that if there was no pen annotation on a slide in the first place, the final pen annotation mask will be empty.

Validation data set and manual annotations

To evaluate our method, we utilized 319 WSI with pen markers, scanned on an Aperio AT2 scanner (Leica Biosystems, Buffalo Grove, Illinois, USA). The WSI have been manually annotated by a pathologist using an in-house developed digital slide viewer[11] on a Microsoft Surface Studio with a digital pen as input device. The pathologist sketched the inner regions of the visible pen markers on the full WSI. Note that the pathologist could use any magnification level in the viewer to annotate the WSI. When the pen shape was coarse, the digital manual annotation was done on a low magnification level of the WSI. When the pen shape was fine or narrow, the pathologist zoomed in to higher magnification levels to annotate the WSI. In any case, the digital annotation mask was saved by the viewer internally at the original dimension of the WSI. The manual annotations were then down-scaled to the size of the thumbnail images.

To assess the performance of our method, we calculated the four similarity metrics Dice coefficient[12] (or F-Score), Jaccard index[13] (or Intersection over Union (IoU)), Precision, Recall and Cohen's Kappa[14] between an automatically generated annotation mask A and a manually drawn annotation mask M:

$$Dice = \frac{2*|A \cap M|}{|A| + |M|} = F\_score$$

$$Jaccard = \frac{|A \cap M|}{|A \cup M|} = IoU$$

$$Precision = \frac{|A \cap M|}{|A|}$$

$$Recall = \frac{|A \cap M|}{|M|}$$

$$Kappa = \frac{p_0 - p_e}{1 - p_e}$$

where p_0 is the probability of agreement on the label assigned to a pixel, and p_e is the expected agreement if both annotations are assigned randomly. All metrics were calculated using the Scikit-learn[15] package in Python. Although these metrics are similar, they highlight slightly different aspects. Dice and Jaccard express the relative amount of overlap between automatic and manually segmented regions. Precision expresses the ability to exclude areas which do not have pen annotations. Recall quantifies the ability to include regions with pen annotations. The Kappa value expresses the agreement between automatic and manually segmented regions as a probability. All values except Kappa range between 0 (poor automatic segmentation) and 1 (perfect automatic segmentation). Kappa values range between -1 and 1, with 0 meaning no agreement between manual and automatic segmentation better than chance level, and 1 and -1 meaning perfect agreement or disagreement, respectively.

Results

We quantify the performance of our method on a data set of 319 WSI. The thumbnails of the WSI have a width of 485–1024 px (median 1024 px) and a height of 382–768 px (median 749 px). As shown in Figure 3, right, and Table 1, the median dice coefficient between the automatically segmented and manual pen masks is 0.942 (mean 0.865 ± 0.207), the median Jaccard index is 0.891 (mean 0.803 ± 0.227), the median Precision is 0.955 (mean 0.926 ± 0.148), the median Recall is 0.943 (mean 0.844 ± 0.237), and the median Kappa value is 0.932 (mean 0.852 ± 0.216). Figure 3, left, sketches a Precision/Recall curve describing our data set. Note that the precision is generally very high (>0.90), while the Recall distributes over a larger range with a median of 0.943, meaning that some manual annotations are missed. The extreme outliers with zero Precision and

Recall indicate disjoint annotations and are discussed in the next section.

Figure 4 illustrates two examples with high scores (Dice 0.983 and 0.981, top), two examples with medium scores (0.755 and 0.728, middle), and two examples with low scores (0.070 and 0, bottom). The easiest annotations are those with closed shapes such as circles or polygons. Still, even if the annotation is easy to process by our method, the score can be lowered if the tissue within the annotation is sparse while the manual digital annotation is coarse, as illustrated in the two medium examples. Difficult annotations for our method are shapes that are not closed and therefore cannot be filled, slides with artifacts such as broken cover slips [Figure 4 second from bottom], or complex annotations such as ring-shaped objects [Figure 4 bottom]. These difficult cases are outliers in our data set, as indicated by the statistics in Figure 3.
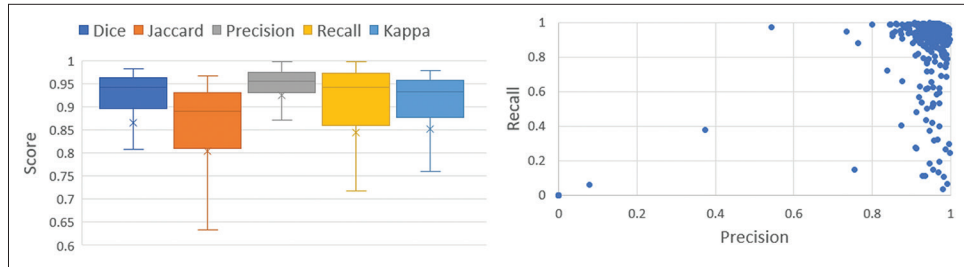

Figure 3: Performance metrics for the proposed annotation extraction method. Left: Dice coefficient (median 0.942), Jaccard index (median 0.891), Precision (median 0.955), Recall (median 0.943) and Kappa (median 0.932) of the automatically segmented annotated regions compared to the masks which were manually drawn by a pathologist. Right: Precision/Recall curve of automatically generated and manually drawn annotation masks. All measures are calculated pixel-wise. $n = 319$
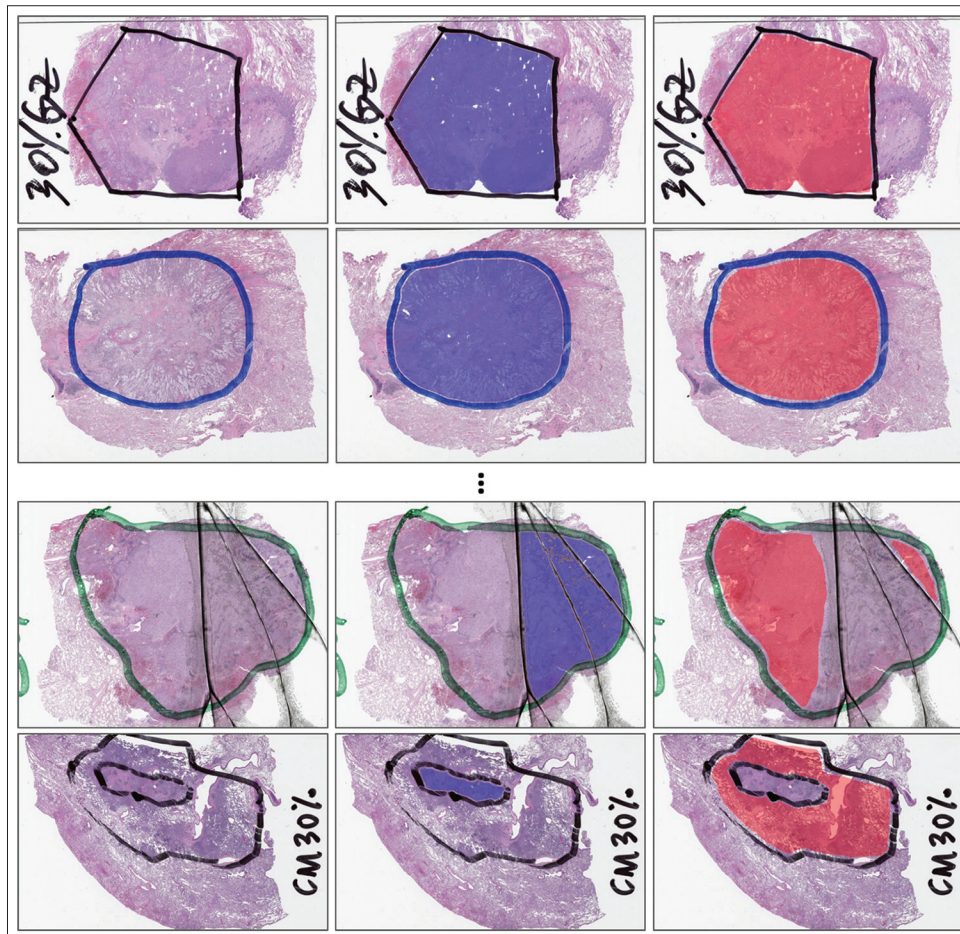

Figure 4: Examples of two high scored extractions (top, Dice 0.983 and 0.981) and two low scored extractions (bottom, 0.070 and 0.0). Left: Original image. The annotations are drawn with a pen on the glass slide. Middle: Automatically segmented annotations. Right: Manually segmented annotations. Note that our method can differentiate between text and tissue outlines. The two low scored examples are difficult due to a broken cover slip, or due to a ring-shaped annotation

Table 1: Statistical summary of the similarity metrics comparing the automatically segmented annotations with the manual annotations (*n*=319)

	Dice	Jaccard	Precision	Recall	Kappa
Mean±SD	0.865±0.207	0.803±0.227	0.926±0.148	0.844±0.237	0.852±0.216
Minimum	0	0	0	0	−0.143
25%	0.896	0.812	0.931	0.860	0.879
50%	0.942	0.891	0.955	0.943	0.932
75%	0.964	0.931	0.975	0.972	0.958
Maximum	0.983	0.967	0.999	0.998	0.979

SD: Standard deviation

An interesting observation is that text annotations are robustly ignored throughout all samples by our method, as illustrated in Figure 4 top. This is achieved by the size-based noise filter that removes small closed areas in roundish letters. We do not incorporate a specific text recognition program.

Annotation time

The time needed for manual digital coarse annotations on all WSI was approximately 5 h, with an average of 1 min per slide.

In contrast, our method runs in 15 min for all slides after finalizing all parameters. Note that images are being processed in sequence, and the script can further be optimized with parallel processing. Due to the time savings, we propose to use our method to extract coarse annotations whenever possible.

Note that this comparison has limitations. While the pathologist can annotate in the viewer at any magnification level, e.g., to account for fine-grained sections, our method runs solely on thumbnails without any option for fine-grained annotations. Further, we do not know the time needed to annotate the glass slides itself with a pen and cannot compare pen annotation time with manual digital annotation time.

CONCLUSION

WSI can contain analog, hand-drawn pen annotations from pathologists. These annotations are commonly used to coarsely outline cancerous areas subject to molecular follow-up or genetic sequencing. Therefore, these annotations can be very valuable for various cancer classification models in computational pathology. However, pen annotations are usually considered as unwanted image artifacts and are aimed to be excluded from the analysis. Instead, we consider the scenario in which these annotations would be beneficial for the classifier if they could be accessed by the algorithm. For this, we present a software that allows for the digital extraction of the inner part of hand-drawn pen annotations. Our method identifies and segments the pen regions, closes the contours and fills them, and finally exports the obtained mask. The tool is freely available at https://github.com/MSKCC-Computational-Pathology/PenAnnotationExtractor.

The performance of our algorithm has been assessed on a pen-annotated data set of 319 WSI, resulting in an overall Dice metric of 0.942 and overall Precision and Recall of 0.955 and 0.943, respectively. Most suitable pen shapes are closed areas as they are easily extractable by our method. However, problematic pen annotations include shapes that are improperly closed or complex by nature (e.g., with holes in them middle). Improperly closed shapes can be addressed with manual adjustments of the dilution radius. More complex shapes such as doughnut-shaped annotations would require further improvements of our method.

In general, the approach that we present can be extended to other data sets, for example to process WSI with a different staining than H&E, or to account for more pen colors. It is not a *fully* automatic pen-annotation extraction method, since it needs potential adjustments of the used parameters. Still, we showed that it is able to capture a bulk part of common annotations which would need much more time to draw manually. Further, we provide guidance to fine tune potential parameters.

Pen annotations can be very diverse and might have various meanings. Our method appeared to be robust against text, possibly since the text does not contain large closed shapes and is typically on the white background and not on the tissue area.

However, pen annotations can be very imprecise since they are drawn on the glass directly, which can be a limitation. It is almost impossible to outline the exact boarder of cancerous regions without any magnification. It has to be kept in mind that using our tool will lead to digital regions at the same precision as the original annotation.

We conclude that a primary use case for our method can be the gathering of enriched tumor samples for training or fine tuning of pathology AI in scenarios in which pen-annotated tumor regions are available.

Conflicts of interest

PJS and TJF are both co-founders of Paige.

REFERENCES

1. Ho DJ, Yarlagadda DV, D'Alfonso TM, Hanna MG, Grabenstetter A, Ntiamoah P, *et al*. Deep Multi-Magnification Networks for Multi-Class Breast Cancer Image Segmentation. Computerized Medical Imaging and Graphics 2021, 88:101866.
2. Ho DJ, Agaram NP, Schüffler PJ, Vanderbilt CM, Jean MH, Hameed MR, *et al*. Deep Interactive Learning: An Efficient Labeling Approach for Deep Learning-Based Osteosarcoma Treatment Response Assessment. In: Martel AL, Abolmaesumi P, Stoyanov D, Mateus D, Zuluaga MA, Zhou SK, *et al*., editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Lecture Notes in Computer Science, vol 12265. Springer, Cham.
3. Bándi P, Balkenhol M, van Ginneken B, van der Laak J, Litjens G. Resolution-agnostic tissue segmentation in whole-slide histopathology images with convolutional neural networks. PeerJ 2019;7:e8242.
4. Li Z, Zhang J, Tan T, Teng X, Sun X, Zhao H, *et al*. Deep Learning Methods for Lung Cancer Segmentation in Whole-Slide Histopathology Images – The ACDC@LungHP Challenge 2019. IEEE J Biomed Health Inform 2020.
5. Sornapudi S, Hagerty J, Stanley RJ, Stoecker WV, Long R, Antani S, *et al*. EpithNet: Deep Regression for Epithelium Segmentation in Cervical Histology Images. J Pathol Inform 2020;11:10.
6. Kothari S, Phan J, Wang M. Eliminating tissue-fold artifacts in histopathological whole-slide images for improved image-based prediction of cancer grade. J Pathol Inform 2013;4:22.
7. Mousavi H, Monga V, Rao G, Rao AU. Automated discrimination of lower and higher grade gliomas based on histopathological image analysis. J Pathol Inform 2015;6:15.
8. Janowczyk A, Zuo R, Gilmore H, Feldman M, Madabhushi A. HistoQC: An open-source quality control tool for digital pathology slides. JCO Clin Cancer Inform 2019;3:1-7.
9. Goode A, Gilbert B, Harkes J, Jukic D, Mahadev S. OpenSlide: A vendor-neutral software foundation for digital pathology. J Pathol Inform 2013;4:27.
10. Suzuki S, Abe K. Topological structural analysis of digitized binary images by border following. Comput Vis Graph Image Process 1985;30:32-46.
11. Hanna M, Reuter VE, Ardon O, Kim D, Sirintrapun SJ, Schüffler PJ, *et al*. Validation of a digital pathology system including remote review during the COVID-19 pandemic. Mod Path 2020;33:2115-27..
12. Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26:297-302.
13. Jaccard P. Lois de Distribution Florale Dans La Zone Alpine. Bulletin de la Société Vaudoise des Sciences Naturelles; 1902. Available from: https://books.google.com/books?id=uMJ6ygAACAAJ.
14. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37-46.
15. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al*. Scikit-learn: Machine Learning in Python. J Mach Learn Res 2011;12:2825-30.