WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Cervical Cancer Identification Using Machine Learning

Jayalakshmi P

Tirupur Kumaran College for Women

ABSTRACT

Because of its inadequate diagnostic system, prostate cancer (PCA) is a serious kind of cancer that kills a significant number of men. Images from cancer patients include important and intricate details that are difficult for conventional diagnostic methods to extract. It is adjusted and works without the need for hand-crafted features. J48 Algorithm, k-nearest neighbor-Cosine (KNN - Cosine), support vector machine (SVM), Gaussian Kernel, and other non-deep learning classifiers were used to compare the outcomes with manually created characteristics including texture, morphology, and grey level co-occurrence matrix. Recent research has concentrated on the suitability of Ensemble Learning (EL) and Transfer Learning (TL) techniques for prostate histopathology image evaluation. In any event, the phases of separation of prostrate CT images have not been seen in many exams. In order to organize well, moderately, and insufficiently separated prostrate CT scans, we thus suggest an Ensembled Transfer Learning (ETL) structure in this work.

Keywords: Cancer, Machine Learning, Cervical Cancer, Segmentation, image processing.

1. INTRODUCTION

Prostate cancer carries a significant risk of morbidity and mortality, making it the disease that kills men least frequently. However, because prostrate cancer grows slowly, there are options for prevention, early identification, and therapy as the disease progresses via precancerous alterations. The primary obstacle to eliminating prostate cancer is seen in low- and middle-income countries (LMICs), where over 88% of prostate cancer fatalities occur due to severe poverty and gender discrimination that significantly restricts a woman's ability to seek care. The process of converting an image into digital format and applying various adjustments to it to produce an improved image or extract some valuable information is known as image processing. This kind of signal distribution uses an image as the input, such as a picture or video frame, and outputs an image or features related to the image. In an image processing system, pictures are typically processed as two-dimensional signals using pre-established signal processing techniques. It is one of the modern technologies that is expanding quickly, having uses in many different facets of business. Within the fields of computer science and engineering, image processing is a fundamental research subject.

1.1 BIOMEDICAL IMAGE PREPROCESSING

processing is a general term for procedures involving the least abstract pictures; both the input and the output have the intensity of biological images. These recognizable biomedical pictures are identical to the original sensor data; typically, a matrix of image function values (brightness values) is used to represent the intensity of the biomedical image. Pre-processing aims to improve the image data by suppressing unwanted distortions and enhancing certain image features that are crucial for further processing. However, since similar techniques are used, geometric image transformations, such as rotation, scaling, and translation, are included in the pre-processing methods category. By enabling the extraction of quantitative information from confocal microscopy pictures of biological materials, image processing techniques have significantly expanded the set of biological topics that may be investigated through experiments. The goal of biology is to comprehend how cells function, what genes do to create a typical animal, and what goes wrong during illness or after damage. For this, they use confocal microscopy pictures of materials labelled with cell-specific markers to examine how changes in gene activity and medication administration influence tissue, organ, or whole body integrity. Although image-processing techniques are shockingly still underutilized, they offer a great deal of potential for extracting information from this type of data. Cell number is one helpful statistic to quantify. Cell number is the delicate balance between cell proliferation and cell death; it is strictly regulated during development and may be changed in diseases, most notably cancer and dementia. A homeostatic control of cell proliferation is accompanied with an increase in cell death following damage (e.g., spinal cord injury).

1.2 CELL SEGMENTATION

In medical diagnostics, cell segmentation has a very high practical value. However, the inconsistent cell border, accretive cells, and interior hollow of the cell picture pose challenges to image segmentation. A machine learning approach based on distance transform is proposed in this cell segmentation to solve adhesion pictures of cells. First, as part of the pre-processing step, the picture is enhanced. Next, it is segmented roughly using the OTSU threshold

segmentation method. Lastly, fine segmentation is achieved by using a machine learning algorithm that optimizes the seed points. According to the accretive cell pictures, machine learning segmentation based on distance transformation transform is therefore feasible. Although it remains difficult since many imaging settings are complicated, reliable cell segmentation is crucial for biological imaging research. This method employs an automatic image segmentation technique called immersion simulation based self-organizing (ISSO) transform. By using user-defined or default self-organizing functions, the approach enables users to tailor the immersion simulation process and incorporate previous knowledge into segmentation. Implemented and applied to a range of pictures and benchmark microscopy datasets from recent studies is a Size Filter based on the ISSO transform. The comparison with other algorithms clearly shows the advantages and versatility of the ISSO approach on different types of pictures, with benchmark error rates significantly lower than those published findings in the literature.

1.3 CELL COUNTING

Cell counting refers to any of the several techniques used in the biological sciences, such as medical diagnosis and therapy, for the counting or comparable measurement of cells. It is a significant branch of cytometry having uses in both clinical and scientific settings. Cell scientists have employed the haemocytometer for cell counting for more than a century. Originally created for the quantization of blood cells, it quickly gained popularity as a useful instrument for counting different types of cells, particles, and even tiny species. Cell counting refers to any of the several techniques used in the biological sciences, such as medical diagnosis and therapy, for the counting or comparable measurement of cells. It is a significant branch of cytometry having uses in both clinical and scientific settings. For instance, a complete blood count can assist a doctor in figuring out why a patient is feeling under the weather and how best to assist them. The amount of cells per unit of volume, or concentration, is often stated as a cell count in liquid media (such as blood, plasma, lymph, or laboratory rinsate) (for example, 5,000 cells per milliliter). Determining the cell concentration is essential for many applications involving cell suspensions, including cell culture and microbiology. A counting chamber is the apparatus that counts the number of cells in a suspension per unit volume.

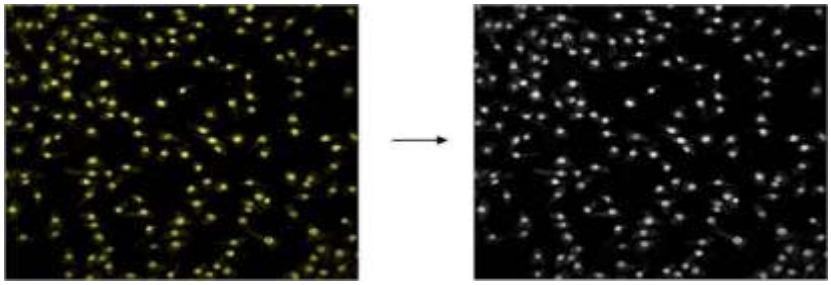


FIG 1. CELL COUNTING

2. LITERATURE REVIEW

2.1 REUSABLE CHITOSAN-PLATELET-RICH PLASMA IMPLANTS TO SUPPORT TISSUE REGENERATION: IN VITRO ASPECTS, IN VIVO DURATION, CELL RECRUITMENT, AND VASCULARIZATION: IN VITRO PROPERTIES

The system that A. CHEVRIER1 et al. suggested The aim of this work was to create formulations of freeze-dried chitosan that can be dissolved in platelet-rich plasma (PRP) to create injectable implants for tissue regeneration. A methodical approach was used to modify the formulation parameters, such as the chitosan number average molar mass (Mn), chitosan concentration, and lyoprotectant concentration, in order to find compositions that would completely and quickly dissolve in PRP (less than 1 minute), have paste-like handling qualities after doing so, and coagulate quickly (less than 5 minutes) to form stable and homogenous solid chitosan-PRP hybrid implants. Cakes that were freeze-dried and included calcium chloride, as well as different concentrations of lyoprotectant, chitosan Mn, and chitosan. PRP was utilized to solubilize the freeze-dried cakes and evaluate both in vivo and in vitro performance—the latter being administered to New Zealand White rabbits via dorsal subcutaneous injections. PRP quickly and thoroughly dissolved freeze-dried polymer compositions with low and medium chitosan Mn and concentrations [1].

2.2 A PROPOSED SIMPLE TOOL FOR TWO-DIMENSIONAL FILLER DISPERSION QUANTIFICATION

In this method, Costantino Del Gaudio and colleagues have suggested For a composite to behave as intended, an efficient filler dispersion within a polymeric matrix is acknowledged to be a vital prerequisite. It is not always easy to quantify this parameter, and several approaches have been put forth in the past; these approaches are typically based on intricate image processing methods. But as of yet, no standard has been created. Here, a unique, simple, two-dimensional alternative technique based on the examination of electron microscopy micrographs is provided to aid in the quantitative evaluation of the filler dispersion in a polymeric matrix. To evaluate the approach's dependability, a number of case studies were taken into consideration. The outcome from these case studies was either an index to gauge the filler dispersion or a qualitative polar plot that represented the studied image. The precise selection of appropriate elements, which are often described as matrix and filler, as well as their reciprocal interactions, determine a composite's real performance [2].

2.3 COMPOSITE MATERIALS: HOMOGENEITY QUANTIFICATION OF NANOPARTICLES DISPERSION

In this system, BINGCHENG LUO et al. have suggested We proposed two quantifiable tools for the evaluation of dispersion in the nanocomposites by using machine-learning algorithms: the information entropy derived from the probability density function and the coefficient of variation of K-nearest neighboring distances (CVKD value). This is because the quantitative assessment of nanoparticle dispersion is essential for the research of nanocomposites, including ceramic, metal, and polymer nanocomposites. Using the Gaussian mixture model, expectation maximization, K-nearest neighbors methods, and Kernel density estimation techniques, 230 distinct types of dispersion morphologies of nanocomposites were examined. The dispersion of nanoparticles was substantially more homogenous when CVKD or information entropy levels were less. As may be predicted, the data presented in this study could help with the logical design and production of excellent nanocomposites. 00:000–000, POLYM. COMPOS., 2018. VC 2018 Society of Plastics Engineers In conclusion, machine-learning methods were used to quantitatively analyse the dispersion of nanoparticles in the nanocomposites. For the computations, 230 distinct kinds of nanocomposites were used. K-nearest neighboring methods were used to compute the coefficient of variation of K-nearest neighboring distances. A more inhomogeneous dispersion of nanoparticles was indicated by a bigger CVKD value, whereas a smaller CVKD value indicated a more uniform dispersion of nanoparticles.

2.4 EVALUATION AND ASSESSMENT OF HOMOGENEITY IN IMAGES. PART 1: UNIQUE HOMOGENEITY PERCENTAGE FOR BINARY IMAGES

Leandro de Moura França et.al., has proposed in this system Texture features analysis is one of the most important approaches for the assessment of homogeneity on images. However, all of them are either relative to the comparison with a standardized set of images, or further multivariate models are strongly required to predict or classify the images according to their features. In this first work, we propose an alternative and novel methodology to calculate a percentage of homogeneity by only using the self-information contained on the image. This methodology is based on the macropixel analysis theory and the generation of what is called the “homogeneity curve”. The homogeneity curve is deeply explored and the knowledge to what it could be considered the most homogeneous and inhomogeneous distribution for every case is spanned. This first work postulates the theory and demonstrates its usefulness with several examples applied to binary images. This will provide a theoretical framework to fully understand the homogeneity curve, postulating a mathematical model to parametrize homogeneity and its plausible deviations One of the main keystones on image analysis is the characterization of the different spatial positions and colors of the elements in an image.

2.5 SINCE 2008, A THOROUGH INVESTIGATION OF MOSTLY TEXTUAL DOCUMENT SEGMENTATION ALGORITHMS

In this research, Sebastien Eskenazi et al. have proposed Segmentation is the process of identifying the different areas of a document in document image analysis. The growing number of uses for document analysis necessitates a solid understanding of the current technologies. The range of methods that have been put forth for document picture segmentation since 2008 is demonstrated by this survey. It offers an understandable taxonomy of both document types and document picture segmentation techniques. We also talk about the evaluation criteria, general community trends, and the technological constraints of these algorithms. Reliable document processing methods are becoming more and more necessary due to industrial document digitalization, document archiving that destroys the original copy, and security technologies that rely on document processing. To choose them appropriately, it would be quite helpful to have a comprehensive list of all the possible al 5 gorithms. Figure 1 illustrates a standard procedure for extracting content from paper documents. The goal of document segmentation is to separate the document picture into sections that make sense. These components can be words, glyphs, paragraphs, text lines, or areas (often containing a single kind of information, such text or graphics). Typically, these sections are utilized for further content extraction processes such word recognition, reading order determination, or document classification.

3. EXISTING SYSTEM

The current system introduces a quantification technique called MASQH in an attempt to resolve the uncertainty around the idea of homogeneity in different sectors. Although it is often used, homogeneity lacks a defined definition, which makes it difficult to compare data across research objectively and encourages subjective analysis. A single homogeneity index is provided by the multi-scale, statistical, segmentation-free MASQH method, which guarantees simplicity and robustness. In three case studies including synthetic pictures, histology images of biomaterials, and nanocomposite images, the article confirms the algorithm's performance and highlights its potential to help scientific community members make objective comparisons. The MASQH approach is widely used and has a significant influence because it is freely available online.

4. PROPOSED SYSTEM

One of the most popular machine learning techniques for continually and comprehensively examining data is the J48 algorithm. The C4.5 method (J48) is primarily utilized in several domains for data classification; examples include categorizing E-governance data and analysing clinical data for the diagnosis of coronary heart disease. The classification algorithm is learned during the learning phase of machine learning, and fresh data is labelled by the algorithm during the classification phase. Classification, or supervised learning, is a data mining task that organizes and classifications the data into predetermined categories. The proposed detection approach allows for a separation of foreground and background over the thresholding operation by quantizing grey levels using a clustering procedure. Pre-processing of the images could be necessary. The restoration of the picture only addresses localized distortions, which makes applying a global grey level clustering far more challenging. Therefore, the complete image with overlapping patches

is divided into groups in order to achieve the clustering. Following individual patch processing, the clusters that are left over are either background or foreground. Cervical cancer detection is improved by improved classification and increased accuracy, both of which are provided by the j48/C4.5 algorithms.

5. MODULES

5.1 IMAGE SEGMENTATION

Although it remains difficult since many imaging settings are complicated, reliable cell segmentation is crucial for biological imaging research. This method employs an automatic image segmentation technique called immersion simulation based self-organizing (ISSO) transform. Let h(x) represent the image I's normalized histogram. Given an image I of size r=m*n pixels, where each pixel may accept L potential grayscale level values in the range [0.L-1], let's analyze the picture.

5.2 J48/ C4.5 ALGORITHM

Based on information theory, the C4.5 algorithm generates decision trees for classification purposes. It is an expansion of the previous ID3 algorithm by Ross Quinlan, which is also referred to as J48 in Weka—the J standing for Java. Because C4.5 generates decision trees that are used for classification, the program is sometimes referred to as a statistical classifier. Many more capabilities are included in the J48 version of the C4.5 method, such as the ability to deduce rules, account for missing data, prune decision trees, and continuous attribute value ranges. J48 is an open-source Java implementation of the C4.5 algorithm used in the WEKA data mining application. J48 permits the use of decision trees or rules derived from them for categorization. Using the idea of information entropy, this method creates decision trees based on a set of training data, much as the ID3 algorithm. A collection S={s1, s2,...} of previously identified samples makes up the training data. Each sample si is made up of a p-dimensional vector (x1, i, x2, i,..., xp, i), where xj stands for the class in which the sample belongs and the attribute values or characteristics of the associated sample. The characteristic with the most information is the optimal one to divide on in order to get the maximum classification accuracy.

The C4.5 method selects the data attribute that divides its set of samples into subsets that are enriched in one or more classes at each node of the tree. The normalized information gain, which is derived from the entropy difference, serves as the splitting criteria. To determine the decision, the characteristic with the largest normalized information gain is selected. After applying a divide-and-conquer strategy to the partitioned sublists, the C4.5 algorithm builds a decision tree based on the greedy algorithm.

5.3 CLUSTERING MICROSCOPIC CELLS USING KMEANS METHODS

The initial stage in diagnosing some blood-related disorders, such as acquired immune deficiency syndrome and prostrate, is the identification of white blood cells (WBCs). Pathologists typically use an optical microscope to make these diagnoses. This procedure requires seasoned professionals in the area and is costly, time-consuming, and incredibly tiresome. This explains why a computer-aided diagnostics system that helps pathologists diagnose cases may be so successful.

Typically, the initial stage of creating a computer-aided diagnostic system involves segmenting WBCs. This work's primary goal is to separate WBCs from microscopic pictures. To achieve this, we provide a three-stage approach that combines thresholding, k-means clustering, and customized machine learning algorithms: (1) segmenting WBCs from a microscopic picture; (2) extracting nuclei from the cell image; and (3) separating overlapping cells and nuclei. The assessment findings of the suggested technique demonstrate that, for nucleus segmentation, the corresponding similarity measures, precision, and sensitivity were 92.07, 96.07, and 94.30%, and for cell segmentation, they were 92.93, 97.41, and 93.78%. Furthermore, statistical analysis shows that the outcomes of the suggested approach and manual segmentation are very comparable.

6. EXPERIMENTAL SETUP

We provide the experimental findings from the segmentation of three types of fluorescent cellular pictures: ground truthed nucleus images, synthetic cell images, and microscopic images of brain cells. The quantitative performance of the four segmentation algorithms is assessed using the first two categories of picture data, and the outcomes are compared to the ground truth. Because there is no ground truth, the brain cell pictures are segregated using qualitative performance analysis.

6.1 QUANTITATIVE MEASURE

At the pixel level, we employ the conventional accuracy, recall, and F-score as quantitative metrics. These metrics are common ways to assess how well segmentation results compare to the original data.

a) b) c)

Fig. (2). Segmentation result for synthetic cell images of low quality in Fig. (1a). a) knn, b) j48 c) svm

measures quantify discrepancy between segmentation results and binary ground truth mask as follows

$$\text{precision} = \frac{\#SR \cap GT}{\#SR} \text{ -------} \rightarrow (26)$$

$$\text{recall} = \frac{\#(SR \cap GT)}{\#SR} \text{ ----} \rightarrow (27)$$

$$\text{f-score} = \frac{2.(precesion.recall)}{precesion+recall} \text{---} \rightarrow (28)$$

where SR is the segmentation result and GT is the ground truth of images. The symbol '#' refers to the pixel numbers in the sets.

6.2 SEGMENTATION OF SYNTHETIC DATA

Since we lack appropriate actual cell pictures with ground truth for evaluation, we choose the second benchmark set, which comprises of multichannel cell images. This collection includes simulations of nuclei, cytoplasm, and subcellular components that are adjusted based on size, position, shape randomization, and other background or fluorescence characteristics. The picture sets are separated into two groups, each with 20 cell images: high quality and low quality (examples displayed in Fig. 1). The second set features a chaotic backdrop with overlapping cells. Every picture has fifty cells. Because every simulated picture has an associated binary mask that serves as the ground truth, binary operations may quickly compute the previously stated quantitative measure.

a) b) c)

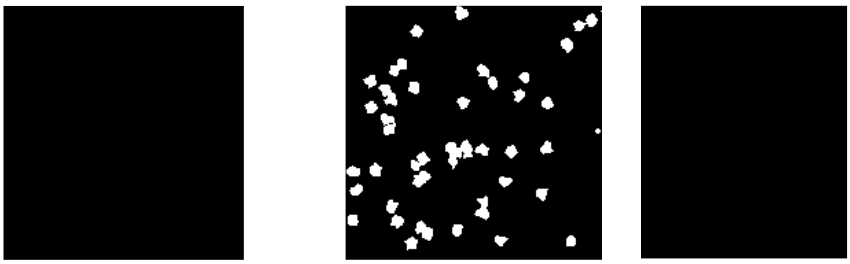


Fig. a) knn, b) j48 c) svm

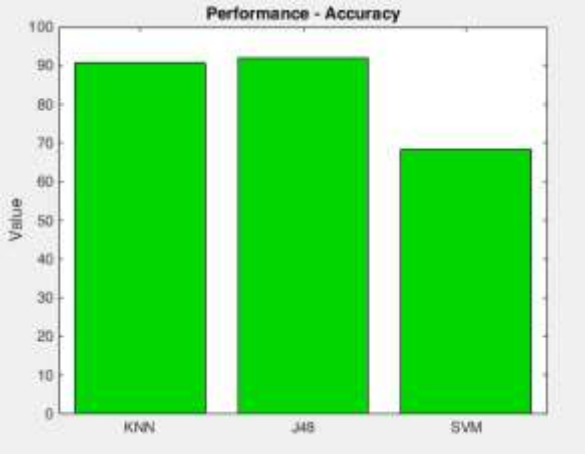
values for the segmentation outcomes utilizing high-quality subcellular pictures. We find that segmentation outcomes from photos of lower quality—with louder backgrounds and overlapping cells—perform worse than those from images of higher quality. As determined by Fscore, accuracy, and recall, Kmeans, Otsu's threshold, and GMAC achieve comparable segmentation quality in both sets of photos. Compared to EM, their performance is more resilient to noise. Furthermore, especially for cell pictures with noisy backgrounds, the EM method maintains substantially greater recall values at the expense of reduced precision.

Table 1. Average Measures of the Segmentation Methods Applied on Low Quality Synthetic Cell Images.

	F score	**Precision**	**Recall**
K-NN	0.9745	0.9726	0.9765
J48	0.9840	0.98267	0.9986
SVM	0.9738	0.9798	0.9679

Table 2. Average Measures of the Segmentation Methods Applied on High Quality Synthetic Cell Images.

	F score	**precision**	**Recall**
K-NN	0.9350	0.9530	0.9180
J48	0.9840	0.98267	0.9986
SVM	0.9738	0.9798	0.9679



7. RESULT AND DISCUSSION

Future research will refine the current deep learning method even further by optimizing the loss function and making minor changes to the network design. Secondly, in order to improve predictions and lower the possibility of post-processing mistakes, we will investigate novel network designs. Third, we will research and experiment with other approaches to data augmentation, such as creating artificial data. Fourth, we'll focus on enhancing the post-processing algorithm's precision and speed. Additionally, employing the j48 method to anticipate cell boundary positions might enhance our approach by removing or at the very least lowering the amount of post-processing stages.

Using the cell-to-cell boundaries from the ground truth segmentation, we attempted to construct a cell boundary class in one of our early studies. Sadly, this led to a poor estimation of the cell borders. Our imprecise ground truth segmentation findings may be the cause of the subpar performance. In further research, we may look into giving pixels near cell borders in the loss function larger weights.

8. CONCLUSION

For the purpose of segmenting cells in fluorescence microscopy pictures, a new K-Means using EM approach was created. This method produced outcomes that were satisfactory. This technique works well for cell separation, enabling proper cell-by-cell characterisation for intricate investigations including the examination of viral infections. Prior to extracting the cells from the background, a machine learning approach was employed. The machine learning method's two-stage algorithm used this first segmented image as input. To properly divide the cells, it uses the Split and Merge procedures based on the Machine Learning Transform. Using fitted cell parameters like area and solidity, the split procedure finds the clustered cells. Machine learning is then used by calculating the distance transform. The merge technique involves morphological operations to remove the divisions and uses the area and eccentricity to detect the over-segmented regions. Furthermore, because this approach does not rely on a geometric correction, it adheres to the uneven form of the cells. They fared better than our approach when compared to the most advanced deep learning architectures, although K-Means (KNN), SVM, and J48 with EM Machine learning works well with databases that have less data since it doesn't require a training procedure. While cell identification in a fluorescent picture was successfully accomplished using the j48 algorithm segmentation approach, further work is required to refine the indications for missing cells and noise.

REFERENCES

1. A. Chevrier et al., “Injectable chitosan-platelet-rich plasma implants to promote tissue regeneration: In vitro properties, in vivo residence, degradation, cell recruitment and vascularization,” J. Tissue Eng. Regenerative Med., vol. 12, no. 1, pp. 217–228, Jan. 2018.
2. C. Del Gaudio and G. A. Licciardi, “A simple tool for two-dimensional quantification of filler dispersion: A proposal,” Fullerenes, Nanotubes Carbon Nanostruct., vol. 27, no. 5, pp. 446–452, May 2019.
3. “Homogeneity quantification of nanoparticles dispersion in composite materials,” Polym. Composites, vol. 40, no. 3, pp. 1000–1005, Mar. 2019.
4. L. de Moura França, J. M. Amigo, C. Cairós, M. Bautista, and M. F. Pimentel, “Evaluation and assessment of homogeneity in images. Part 1: Unique homogeneity percentage for binary images,” Chemometric Intell. Lab. Syst., vol. 171, pp. 26–39, Dec. 2017, doi: 10.1016/j.chemolab.2017.10.002.
5. S. Eskenazi, P. Gomez-Krämer, and J.-M. Ogier, “A comprehensive survey of mostly textual document segmentation algorithms since 2008,” Pattern Recognit., vol. 64, pp. 1–14, Apr. 2017.